Grape minds think alike

Talyah Greyling[a]

[a]*Stellenbosch University, Cape Town, South Africa*

Abstract

UPDATE ABSTRACT * Copy table output, put in ChatGPT & ask to rewrite (**Latex?**) format, paste in Paper_23761067 * ggsave & include_graphics("{location}tables/table1.png") for plots Main research question: NLP: Which words in reviews are associated with higher wine ratings? * interesting sentiment/wine words/descriptives literature * Method: 2 step: random forrest (or lasso) - variable selector (words as columns & ratings as rows), top N words (20/50) filter only on those columns & draw word clouds, variable importance plots & PDP (Partial Dependence Plots) * Then run regressions of best words (OLS)

Abstract to be written here. The abstract should not be too long and should provide the reader with a good understanding what you are writing about. Academic papers are not like novels where you keep the reader in suspense. To be effective in getting others to read your paper, be as open and concise about your findings here as possible. Ideally, upon reading your abstract, the reader should feel he / she must read your paper in entirety.

Corresponding author: Talyah Greyling
Email address: 23761067@sun.ac.za (Talyah Greyling)

Contributions:
The author would like to thank her mother, Melina, for the ample amounts of coffee supplied to her desk.

1. Introduction

Grape minds think alike... or do they? The objective of this study is to investigate which words (if any) used in wine reviews are associated with higher wine ratings on an online marketplace platform. The data utilised in this study is a sample drawn from Vivino [1] reviews for Calitzdorp, South Africa for the years 2014 to 2016. Section 2.3 contains a review of the existing literature surrounding sentiment and, more specifically, wine descriptives. Thereafter section 2 explains how the data was cleaned and transformed and embarks on an exploratory description and analysis of the data by making use of histograms and word clouds. In section 3 statistical modelling is employed in a two-step process consisting of a random forest variable selection method and an OLS regression to determine which words used in wine reviews are the best predictors of good ratings. Thereafter a conclusion is drawn in section 4.

[1] *follow this link to peruse the website*

2. Literature review

2.1. Economic importance

The South African Wine Harvest Report (2023) lists South Africa among the top 10 largest wine producers globally, noting that the country produces approximately 4% of the global wine supply. Additionally, over R55 billion of the country's GDP can be attributed to the wine industry which employs roughly 269 000 workers (Wines of South Africa (WoSA), 2023). The Economic importance of this research question therefore lies in the ability to understanding how to boost one's ratings and thereby ultimately sales by using the correct words in reviews. This research can therefore have an impact on both the country's GDP as well as employment if the boost in ratings and income generated is significant enough.

2.2. Sentiment analysis

Ali, Farooq, Imran & El Hindi (2025) describes sentiment analysis as the undertaking of determining the opinions and emotions conveyed via text and other media. This has a multitude of applications in our increasingly technologically driven society in the domains of e-commerce and marketing, social media, politics and health. With the explosion of text based content during the fourth industrial revolution, Dang, Moreno-García & Prieta (in press) explain that a vast amount of opposing representations of anything ranging from facts to opinions can be found online. This gives sentiment analysis an important role to play in deciphering this new technological landscape we find ourselves in. Alalwan (2018) note that utilising data mining tools and social media analytics have indeed given businesses invaluable insight into the preferences, attitudes and behaviour of their customers.

However, these online platforms where individuals have the freedom to write whatever and however they deem fit are equally hazardous as they can facilitate the spread of incorrect information. Vosoughi, Roy & Aral (2018) found incorrect information to actually spread wider and faster on social media than its true counterpart. Nevertheless, these platforms form an integral part of sentiment analysis since they provide an ever changing source of data that is becoming more complex and connected by the day (Dang *et al.*, in press).

Vivino is predominantly an online marketplace, but can also be considered as a social media platform of sorts. The mechanics of the site's interface and features, to a large extent, mirror those of any standard social media platform. Vivino users create a profile with an accompanying profile picture and are able to follow and therefore be followed by other users. Furthermore users are able to post images alongside their wine reviews, which other users can then like, comment on or share at their own discretion. The reviews shared on Vivino are therefore subject to the same emotional and behavioural biases as social media posts which means that value can be added to understanding the inner workings of this platform's data by applying sentiment analysis techniques.

2.3. Wine descriptives

Sentiment analysis has proven to be a useful tool to analyse the wine industry before. Previous accounts include the study done by Barbierato, Bernetti & Capecchi (2022), who used reviews on packaged wine tour experiences extracted from TripAdvisor, an online travel review website, to identify the themes present in wine tourism experiences and seperate them into positive or negative bins based on their sentiment. They also established the elements defining users' perceptions of quality to understand how wine tourists value the different features of the winescape.

Bangwayo-Skeete & Skeete (2025) also draw on data from TripAdvisor to analyse sentiment by focusing on user's online discussions regarding 2 US wine festivals (yet another element of wine tourism) to identify the usefulness of topics & emotions in guiding the purchasing behaviour of future users.

Another wine related sentiment study, done by Rui, Sparacino, Merlino, Brun, Massaglia & Blanc (in press), aims to bridge the gap in comprehending post-purchase satisfaction within wine e-commerce by analysing consumers' feedback in prominent markets located in the US, UK and China. Using text mining they find differences, by market in consumers' online wine shopping preferences and aversions and provide a guide that producers and platforms alike can utilise to improve their sales and services.

3. Exploratory data description & analysis

3.1. Data, transformation and cleaning

The dataset utilised in this analysis comprises 2298 observations of 107 variables, each encapsulating aspects and features related to the user, vintage of wine and winery concerned in each review. An excerpt from the original dataset can be found in Table 1. This is supplemented by Table 2 which displays a description for each of the variables extracted for use in this analysis.

The original data set was modified slightly by filtering out all reviews posted in languages other than English since using APIs to translate them to English was too costly an endeavor to undertake. 1667 of the original observations remained.

The reviews were tokenised and 30 800 tokens were recorded where-after I followed the standard pre-processing steps of removing punctuation, numbers, symbols, characters and taking everything to lower case. I also used the default tidyverse stop_words dataset to remove stop words using 3 respective lexicons: snowball (from the tm package), onix (from the ONIX text retrieval system) and smart (from the SMART information retrieval system). After these intermediate steps there were 16 869 tokens remaining. All 'clean' tokes were ultimately used in the random forest model to avoid unnecessary sampling bias.

3.2. Data Exploration

First, I was interested in uncovering the distribution of ratings associated with review tokens. This was done by calculating the average rating associated with each unique token and visualising the frequency observed for each rating through the histogram presented in Figure 1. The high frequency of observations that can be observed at a rating of 3.5 and 4 suggest that there may be certain words driving better ratings.

To gain a bit of insight into which words these might be I created a popularity index that used formula (1) to determine the popularity of each token.

$$\text{popularity\_index} = 0.8 \times \text{word\_count} + 0.2 \times \text{ave\_rating} \tag{3.1}$$

The most popular words (i.e. words that get high ratings and are used frequently) in reviews based on my index can be seen in the word cloud depicted in Figure 2. It comes as no surprise that wine is the most 'popular' word in a collection of wine review data. The words to take notice of however are port, dark, nose, sweet and red. It is interesting to note that 4/5 of these most popular words seem best suited to describing red, instead of white wines.

The wide difference in importance of tokens as portrayed by both Figures 1 & 2 are of a large enough extent to merit the statistical modelling in the next section.

4. Statistical modelling

To uncover the underlying relationship between words used in wine reviews and high ratings a two-step statistical modelling process is employed.

4.1. Step 1: Random Forest

The first step entails utilising a random forest model as a method to select the 20 words (or 'variables') that best predict a high rating. This model was chosen for its ability to handle complex datasets effectively as well as its scope for completing classification and regression tasks in a manner that is easy to interpret.

These variables are then fed into a simple OLS regression in 2 step of the modelling process to calculate coefficients and standard errors and inspect the significance of these words in predicting high ratings.

To employ step 1 the tokenised data was first converted into a document term matrix of words and each re-linked to their respective ratings. Thereafter I partitioned the data set into two subsets for training and testing. I decided to adopt the most generally used splitting practice and divide the data by means of a 70/30 split, whereby the training set makes up 70% of the data and the other 30% of the data is saved for testing. By employing this split a healthy balance is maintained between availing a sufficient amount of data to effectively train the model and keeping a significant portion of the data unseen to enable assessing how the model performs without the risk of overfitting that would have been present if the data where trained and tested on the same set of data. Figure 3 plots the distribution of ratings found in respectively the training (red) and the testing (green) sets. These followed a satisfyingly similar trajectory that enabled me to confidently continue with my analysis.

The recipe (a set prepossessing pipeline created for the data) was then defined and set to facilitate 3 transformations: it includes a hot-encoding applied to categorical variables, removes numeric predictors with zero variance and normalizes numeric predictors.

Next, a tree-based model called a Random Forest is specified. This model utilises numerous trees fit on bootstrapped samples of data reserved for training and using an arbitrary subset of predictors at every split (see the mtry parameter) which then averages the predictions made by all the trees to reduce variance. This method is therefore robust to overfitting, can efficiently process interactions and nonlinearities and is insentitive to scaling (by which variables do not ned to be normalized). My regression forest model employs 100 trees with a minimum of 10 observations in a node pre split.

A workflow containing the recipe and model specification is built to improve reproducibility and reduce errors by ensuring that all transformations are recorded and performed in the correct order. A 5-fold cross-validation approach is employed for model evaluation and hyperparameter tuning whereby the

training data is divided into 5 parts that are each employed once as the validation set and act as the training set otherwise. This ensures a robust model performance estimate that doesn't rely on a single split between training and testing.

Random Forests require minimal hyperparameter tuning. The most important parameter to optimise is mtry since a value that is too low till alleviate bias and a number that is too large will increase variance. I defined a regular grid over values of mtry that were possible between 2 and 18 and tune the forest with the same workflow and CV folds as earlier. Figure 4 plots the root mean squared error performance of my Random Forest by mtry. The model is tuned to the lowest mean metric value, found at mtry = 16, for the best performance.

Finally the ultimate model was fit on the full training set, the optimal hyperparameters were plugged into the workflow and the final models' performance evaluated on the test set. The trained model was inspected and an assesment was carried out on the models' performance on unseen data. These results are displayed in Table 3. The RMSE (Root Mean Squared Error) indicates that on average, predictions deviate from actual ratings by 0.64. The R-squared tells us that 12% of the variation in ratings is explained by the model and the MAE (Mean Absolute Error) tells us that the typical absolute error in prediction is around 0.46. The low RMSE and MAE are both extremely satisfying whilst the low R-squared conveys disappointing explanatory power.

The top 20 words identified by the Random Forest model are then visualised in the variable importance plot displayed in Figure 5. As can be expected descriptive words conveying a positive sentiment like 'excellent', 'absolutely', 'fantastic', 'amazing' and 'delicious' make up our top ranks. Lesser expected words are 'smooth' and 'port' confirming our suspicions that red wines being better associated with high ratings than white wines from Figure 2. Furthermore Figure 6 depicts a word cloud of the top 50 words in reviews identified by the Random Forest model where further positive descriptives like 'wonderful', 'beautiful', 'bliss', 'brilliant', 'elegant' and 'stunning' can be found.

4.2. Step 2: OLS Regression

During the second step ratings are regressed on the top 20 words, as identified by the Random Forest model during step 1, to get coefficients for their probability of determining good ratings. These regression results are summarised in Table 4 and we can see that to our satisfaction most of the coefficients display with great significance at even up to the 0% level. We see that the probability of receiving a good rating increases by 50% when the word 'excellent' is used in a review, but reduces by 168% when the word 'bland' is included. Interestingly Port is the only vintage that makes it into the top 20 tokens. This once again hints to an overall preference for red above white wine, but could also be due to the concentration of wines made in Calitzdrop.

Figure 5 portrays a plot of the coefficients such that the playoff between those words with positive

(INSERT COLOUR) and negative (INSERT COLOUR) effects on ratings can be seen. We note that there is a higher collection of positive descriptives that increase the probability of a good rating than those predicting a bad rating and note that negative descriptives have an effect of a greater magnitude in reducing the probability of a goot rating than the positive descriptives have of increasing a good rating. This can be driven by humankinds tendancy to loss aversion where a loss (buying a bad wine) is valued at a greater magnitude than a gain (buying a good wine) of a similar size.

5. Conclusion

I therefore find satisfactory evidence that there is indeed a correlation between the words used to describe wine in reviews and their subsequent ratings. These findings are significant and were created to be robust to errors by implementing a Random Forest model to identify which words to include as explanatory variables in my OLS regression. As expected positive descriptives has a positive effect on ratings and vice versa for negative descriptives.

References

Alalwan, A.A. 2018. Investigating the impact of social media advertising features on customer purchase intention. *International journal of information management*. 42:65–77.

Ali, H.M.U., Farooq, Q., Imran, A. & El Hindi, K. 2025. A systematic literature review on sentiment analysis techniques, challenges, and future trends: A systematic literature review on sentiment analysis techniques. *Knowledge and information systems*. 67(5):3967–4034.

Bangwayo-Skeete, P. & Skeete, R.W. 2025. Sipping and sharing: Impact of emotions and topics on the usefulness of travelers' wine festival forum posts. *International journal of wine business research*. 37(1):134–158.

Barbierato, E., Bernetti, I. & Capecchi, I. 2022. Analyzing TripAdvisor reviews of wine tours: An approach based on text mining and sentiment analysis. *International journal of wine business research*. 34(2):212–236.

Dang, C.N., Moreno-García, M.N. & Prieta, F.D. la. (in press). An approach to integrating sentiment analysis into recommender systems. *Sensors (Basel, Switzerland)*. 21(16):5666–.

Rui, M., Sparacino, A., Merlino, V.M., Brun, F., Massaglia, S. & Blanc, S. (in press). Exploring consumer sentiments and opinions in wine e-commerce: A cross-country comparative study. *Journal of retailing and consumer services*. 82:104097–.

Vosoughi, S., Roy, D. & Aral, S. 2018. The spread of true and false news online. *Science (American Association for the Advancement of Science)*. 359(6380):1146–1151.

Wines of South Africa (WoSA). 2023. *South African Wine Harvest Report 2023*. Wines of South Africa. [Online], Available: https://www.wosa.co.za/The-Industry/Vintage-Reports/South-African-Wine-Harvest-Report-2023/ [2025, June 13].