

# Of Wines and Reviews: Measuring and Modeling the Vivino Wine Social Network\*

Neema Kotonya<sup>1</sup>, Paolo De Cristofaro<sup>2</sup>, and Emiliano De Cristofaro<sup>1</sup>

<sup>1</sup>University College London   <sup>2</sup>Tipicamente Wine Blog

neema.kotonya.13@ucl.ac.uk, e.decrisofaro@ucl.ac.uk, paolodecri@alice.it

## Abstract

This paper presents an analysis of social experiences around wine consumption through the lens of Vivino, a social network for wine enthusiasts with over 26 million users worldwide. We compare users' perceptions of various wine types and regional styles across both New and Old World wines, examining them across price ranges, vintages, regions, varietals, and blends. Among other things, we find that ratings provided by Vivino users are not biased by cost. We then study how wine characteristics, language in wine reviews, and the distribution of wine ratings can be combined to develop prediction models. More specifically, we model user behavior to develop a regression model for predicting wine ratings, and a classifier for determining user review preferences.

## 1 Introduction

Over the years, research has shown that food and alcohol consumptions are strongly shaped by social influences [8, 22], often mirroring those of people with shared social connections [19]. In fact, eating and drinking habits are increasingly part of our social media footprints. Moreover, regional factors, such as economic, cultural, and lifestyle variations, inevitably affect perceptions and choices [2]. Therefore, the rise of dedicated social apps, as well as the growing use of social media to discuss and share such habits, offer new opportunities to elicit valuable insight at a large scale.

Previous work has studied the cultural and socio-economic factors determining what we eat and drink [4]. More recently, researchers have analyzed food consumption on mainstream social networks such as Twitter [1] and Instagram [9], as well as beer habits on social networks like Untappd [3]. However, to the best of our knowledge, wine consumption has long been overlooked, even though it has played an important societal role in many cultures for thousands of years [23].

Aiming to bridge this gap, this paper studies Vivino, a social network application for wine enthusiasts with over 26 million active users worldwide. Vivino provides a platform for reviewing and rating wines, forums for sharing experience and knowledge, as well as a marketplace for buying and selling wines. We focus on Vivino as it allows us to perform a large-scale

measurement of wine consumption.

Specifically, we focus on two main research objectives: (1) Characterizing the Vivino social network in terms of how regional factors, user biases, and wine characteristics affect wine ratings; and (2) Integrating wine characteristics, reviews, and wineries on Vivino to develop practical models for predicting wine ratings and user reviews.

We crawl Vivino to collect data about wines, including prices and ratings, as well as attributes, such as their vintage, type, regional style, and the winery from which they originate. We also collect user data, including ratings, reviews, taste profile, ranking, country, and followers, as well as the users whom they follow. Next, we study the relationship between ratings and a number of wine characteristics, and examine users' biases as to which wines they drink and how they score them, as well as the language used in both the user biographies and the wine reviews. Lastly, we integrate the findings of our analysis to develop systems for predicting users' ratings and reviews. In other words, we show how to use the patterns uncovered in the data analysis of Vivino wines, users and user-generated content, to generate: (i) a predictive model for wine ratings, and (ii) a model for categorizing user preferences.

**Main findings.** Overall, we find that:

1. The ratings and the reviews supplied by Vivino users display the same rich knowledge of wines as professional wine reviews. However, unlike the latter, Vivino users' ratings do not seem to be affected by wine prices;
2. Vivino users have an affinity for rating local wines. There are also strong geographical similarities in how wines from adjacent countries are rated;
3. Vivino's user-generated data produces accurate practical models for the prediction of wines. The regression models for predicting wine ratings achieved  $R^2$  scores  $>65\%$ , and the user preference classifier has a mean accuracy score of almost 80%.

## 2 Related Work

Previous work has studied online diaries of food and drink consumption habits, e.g., examining how dietary choices are linked to food-related tweets by Twitter users in the United States. Abbar et al. [1] find a strong correlation between food mentions on Twitter with obesity and diabetes statistics, while

\*A preliminary version of this paper appears in the Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018). This is the full version.

Mejova et al. [10] present a large-scale analysis of pictures taken at US restaurants. Sharma and De Choudhury [18] investigate how nutritional information about food is communicated through Instagram.

Researchers have also analyzed recreational eating on social networks. Guidry et al. [7] study how hashtags in social media are used to express (dis)approval of fast food chains, while Mejova et al. [9] collect a dataset of Instagram conversations including certain hashtags, and characterize three kinds of users: singletons, residents, and travelers, showing that travel drives emotions associated by users to hashtags.

Another line of work introduces models for predicting the cuisine and the original geographical region a dish from its recipe [6, 16], collecting data from wiki-like pages, rather than social networks. These mostly aim to develop supervised learning models to predict the cuisine, via multi-class learning. This is similar to one of the models we propose, although we also use a regression based one to determine wine ratings in real-value terms. Moreover, the datasets they use are significantly smaller than ours, i.e., thousands vs millions of entries. We also find that confusion matrices presented in previous work show that the most commonly misclassified cuisines are from the same regions of the world, thus, the presumable source of their ingredients is from the same locations. In our work, we aim to discover if there was any detectable confusion between Old and New World wines; besides looking at prediction models, we also consider models for recommendation. These allow us to learn more about users, as they introduce users to new genres, in our case, wine types, which they might have not previously considered.

Overall, while numerous studies focus on eating habits and food consumption, fewer analyze the behavior driving consumption of drinks, in particular alcoholic beverages. A notable exception is the work by Chorley et al. [3], who analyze the beer-oriented app Untappd. Specifically, they study beer ratings and highlight that Untappd users express a generally favorable opinion of lagers and ales. They also find correlation between scores assigned to beers by American and European drinkers, showing that Untappd displays a power-law distribution like most real networks [5].

## 3 Datasets

### 3.1 Background: Vivino

Vivino is a wine marketplace and online community for wine enthusiasts, which is available both as a web and a mobile application. It was founded in 2009 by Heini Zachariassen, with his colleague Theis Sondergaard joining the venture in 2010. Vivino has grown rapidly since then, boasting 29 million users as of March 2018. In a nutshell, the application allows users to review and purchase wines through third-party vendors. The mobile application also provides a *wine scanner* functionality, i.e., users can upload pictures of wine labels and access reviews and details about the wine/winery.

Vivino is really a social network, as it allows wine enthusiasts to communicate with and follow each other, as well as share reviews and recommendations. As of March 2018,

Vivino reports featuring 9.2M wines (including dessert and port wines), covering a multitude of wine styles, grapes, and geographical regions, as well as 89.4M ratings and almost 29.9M reviews.<sup>1</sup> Users can also earn a variety of rewards for their activity, e.g., receiving *likes* on their posts, and get prompted to unlock achievements, e.g., if they “scan a wine from Argentina.” High-performing reviewers also become *ambassadors* or receive labels like “Top Ten in Country Californian Meritage.”

In terms of reviews, Vivino claims that their 5-star rating system (with 0.5 granularity) has a good correlation with Robert Parker’s 100 point scale [14], and argues that its users are able to produce a greater number of ratings than the seven most prolific wine experts (in fact, Vivino users have produced ratings for 1.4M wines in the period between 2011 and 2015, while only 370k wines received expert ratings).<sup>2</sup>

### 3.2 Crawler

Vivino does not offer an API to collect data from their site, therefore, we gathered data about wines, wineries, and users from the Vivino website using a custom web crawler in Python, relying on the *Selenium* and *requests* packages. To avoid generating an extensive amount of traffic, causing possible issues for the site operators, we throttled the crawler to 0.2 requests per second, and ran it over five months (November 2016 to March 2017).

### 3.3 Wines, Users, Reviews, and Wineries

Our crawl yields four datasets, respectively, containing wines, users, reviews, and wineries, as discussed next.

**Wines dataset.** In total, we collected data for 1.06M wines. We gathered the following attributes: *Name* (of the wine), *Type* (i.e., white, red, rosé, port, dessert, or sparkling), *Vintage* (i.e., year of production), *Average Price*, *Average Rating*, *Number of Ratings*, *Ratings Breakdown* (i.e., the number of 1.0, 2.0, 3.0, 4.0, and 5.0 stars received), *Country*, *Region* (e.g., Bourgogne), *Regional Style*, (e.g., Spanish Rioja) *Winery*, *Food Pairings* (e.g., lamb), and *Grapes* (e.g., merlot).

Note that some fields are sparsely populated, e.g., *Regional Style*. Also, the *Average Price* really depends on the availability at suppliers working with Vivino, and 474.3k wines are not listed at any supplier, thus the corresponding field is empty.

**Users dataset.** Next, we selected the 10k most active users from each of the 15 countries with the highest level of wine consumption according to the International Organization of Vine and Wine (OIV) [13], i.e., Argentina, Australia, Brazil, Canada, China, France, Germany, Italy, Portugal, Romania, Russia, South Africa, Spain, UK, and USA. We did so since the majority of Vivino users do not submit any review, while we wanted to make sure to capture a non-negligible number of reviews per user. In order to get the top users, we used Vivino’s ranking system, which ranks users by country according to their activity (i.e., number of reviews) and contributions to the platform (i.e., posts receiving likes and comments). Out

<sup>1</sup><https://www.vivino.com/about>

<sup>2</sup><https://www.vivino.com/wine-news/vivino-5-star-rating-system>

Country	Wines		Users		Wine Reviews		Wineries	
Argentina	43.5k	4%	8.8k	6%	35.5k	5%	0.19k	2%
Australia	56.0k	5%	9.5k	6%	55.0k	7%	0.41k	5%
Brazil	3.9k	0%	8.6k	6%	0.8k	0%	0.01k	0%
Canada	7.9k	1%	9.4k	6%	9.5k	1%	0.05k	1%
China	0.0k	0%	9.9k	7%	0.0k	0%	0.00k	0%
France	252.0k	24%	9.9k	7%	139.6k	18%	1.89k	23%
Germany	23.6k	2%	9.7k	7%	8.4k	1%	0.08k	1%
Italy	222.7k	21%	10.7k	7%	132.1k	17%	1.76k	21%
Portugal	52.1k	5%	9.4k	6%	20.4k	3%	0.44k	5%
Romania	2.1k	0%	9.5k	6%	0.5k	0%	0.02k	0%
Russia	0.4k	0%	10.8k	7%	0.0k	0%	0.00k	0%
South Africa	41.2k	4%	9.8k	7%	50.3k	7%	0.31k	4%
Spain	73.6k	7%	10.5k	7%	48.1k	6%	0.62k	8%
UK	1.3k	0%	10.9k	7%	1.2k	0%	0.00k	0%
USA	173.0k	16%	9.8k	7%	209.8k	27%	1.70k	21%
Others	109.4k	10%	6.7k	5%	60.7k	8%	0.76k	9%
<i>Total</i>	1.06M		147.1k		771.9k		8.3k	

**Table 1:** Summary of the wine, user, wine reviews (initial posts, not replies) and winery datasets collected by crawling the Vivino website, broken down by country.

of the 150k top users (10k for 15 countries), we acquired data for 147.1k (98%); the rest was missed due to crawler failure, errors, and/or time limitations. We were unable to collect all the top users because of web crawling restrictions and time limitations.

Overall, for each user, we collected: *Username*, *Biography*, *Country*, *Ranking* (i.e., how the user ranks in terms of rating/review contributions compared with other users in their country), *Number of Followers* the user has acquired, *Number of Users Followed*, *Taste Profile* (the regional styles reviewed by the user, with counts and average ratings), the *Total Number of Ratings* supplied by the user to Vivino, and whether the user has been *Featured* by Vivino – a special status where the users profile is promoted on the application to other users.

**Reviews dataset.** We also collected the reviews posted by each of the top users in our user dataset, gathering 771.9k review posts for 86.6k unique wines. (In other words, we created a database with wine reviews collected using the IDs of a sample of the wines scraped.) For each review, we gathered: *Wine*, *Vintage*, *Content* (which includes the author’s username and the number of ratings provided by them), *Date*, and *Replies*. Overall, we gathered 771.9k reviews, which were posted by 370k unique authors. Additionally, these initial review posts garnered 617.7k replies collectively.

**Wineries dataset.** Similar to the wines dataset, we also got records for 8.3k wineries, which represent the 1.06M wines scraped from Vivino. Specifically, we retrieved: *Basic Details* (i.e., name, number of wines produced, URL of its Vivino profile page), *Ratings* (number and average score), *Wine Maker* (if any), *Location* (GPS, region, and country), and *Websites* (Twitter, Facebook, and official websites).

### 3.4 General Characterization

In Table 1, we present a summary of our datasets. As discussed above, we have data for 1.06M wines as well as 147.1k

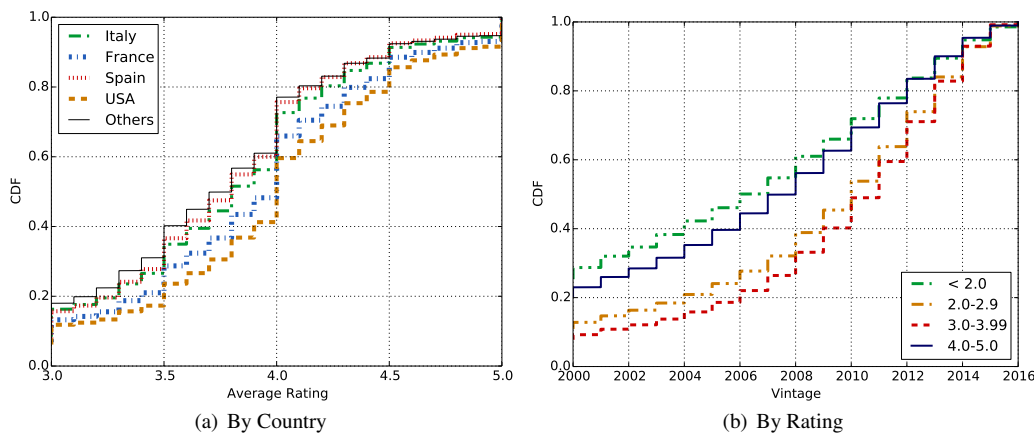
user profiles, along with 771.9k reviews and 8.3k profiles of wineries.

**Countries.** The entire wine dataset includes wines from 49 distinct countries. Interestingly, 98% of them (1.04M) are produced in 21 different countries<sup>3</sup> The top four countries, i.e., France, Italy, Spain, and the United States, account for more than two third of all the wines on Vivino (69.1%). This is not surprising since these are also the four largest wine producing countries, accounting for 62% of the worldwide wine production in 2014 [12].

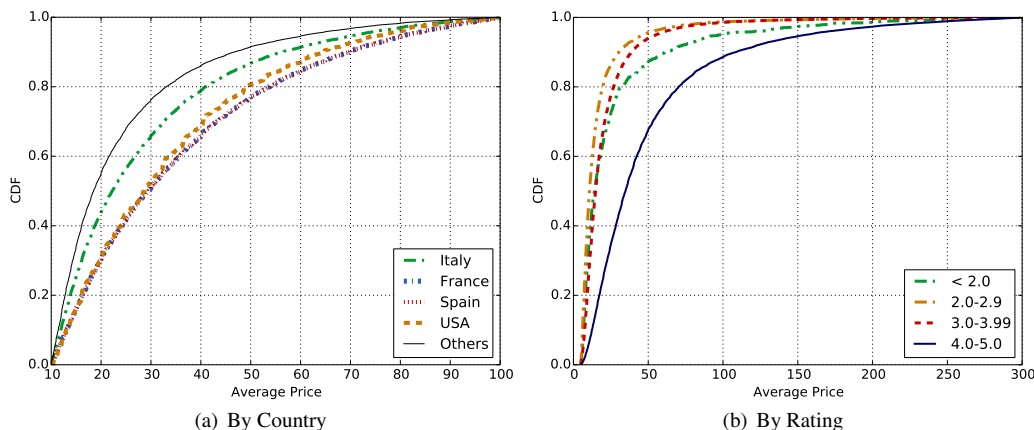
**Ratings.** The ratings awarded to wines are generally favorable. The most common ratings are either 3.0 or 4.0 stars. All of the 21 wine-producing nations have a mean rating higher than 3.2, and no country has wines whose average rating exceeds 3.9. Brazilian (3.36), Chilean (3.46) and Romanian (3.28) wines have the lowest average ratings. The highest average ratings belong to the US (3.84), Germany (3.79), and France (3.78). The countries with the highest number of wine ratings per wine are Argentina (145.35), Chile (112.57), and Brazil (99.51). However, the wines with largest number of total ratings are Italy (1.4M), France (1.1M) and the United States (1.0M). In Figure 1, we report the CDF of wine ratings divided by country (a) and of vintages by rating (b).

**Prices.** As mentioned, we also collected the (average) price for each wine, when available. We find interesting differences across countries. The country with the most expensive wines is France (£127.12 average price); this is inline with the fact that a disproportionate number (41 out of 50) of the world’s most expensive wines are produced in France [17]. After France, we find the United States (£67.54), followed by Portugal (£64.69). Conversely, the three countries with the lowest mean prices are Croatia (£15.94), South Africa (£15.54), and Romania

<sup>3</sup>As per most wines: France, Italy, USA, Spain, Australia, Portugal, Chile, Argentina, South Africa, Germany, New Zealand, Austria, Canada, Hungary, Israel, Brazil, Greece, Romania, Georgia, Uruguay, Mexico, Switzerland.



**Figure 1:** CDFs of average ratings by country and vintages by rating.



**Figure 2:** CDFs of the average price for the wines in our dataset.

(£11.62). In Figure 2, we report the CDFs of wine prices divided by country (a) and ratings (b).

**User Reviews.** Finally, we find that wines from US, France, and Italy receive the bulk of the written reviews – 209.8k, 139.6k, and 132.1k, respectively (see Table 1), in addition to the highest number of replies to these review posts, with 127.7k, 148.7k, and 117.7k. Whereas, the highest levels of engagement, i.e. replies per comment, are for wines from Croatia (1.8), United Kingdom (1.689) and Israel (1.321). The levels of engagement for wines from US, France and Italy are 0.609, 1.065, and 0.891. The mean engagement overall is 1.2519 replies per post. Also, reviews for Italian and French wines have the largest number of unique authors – 83.8k, 60.3k, and 65.5k respectively.

## 4 Rating Analysis

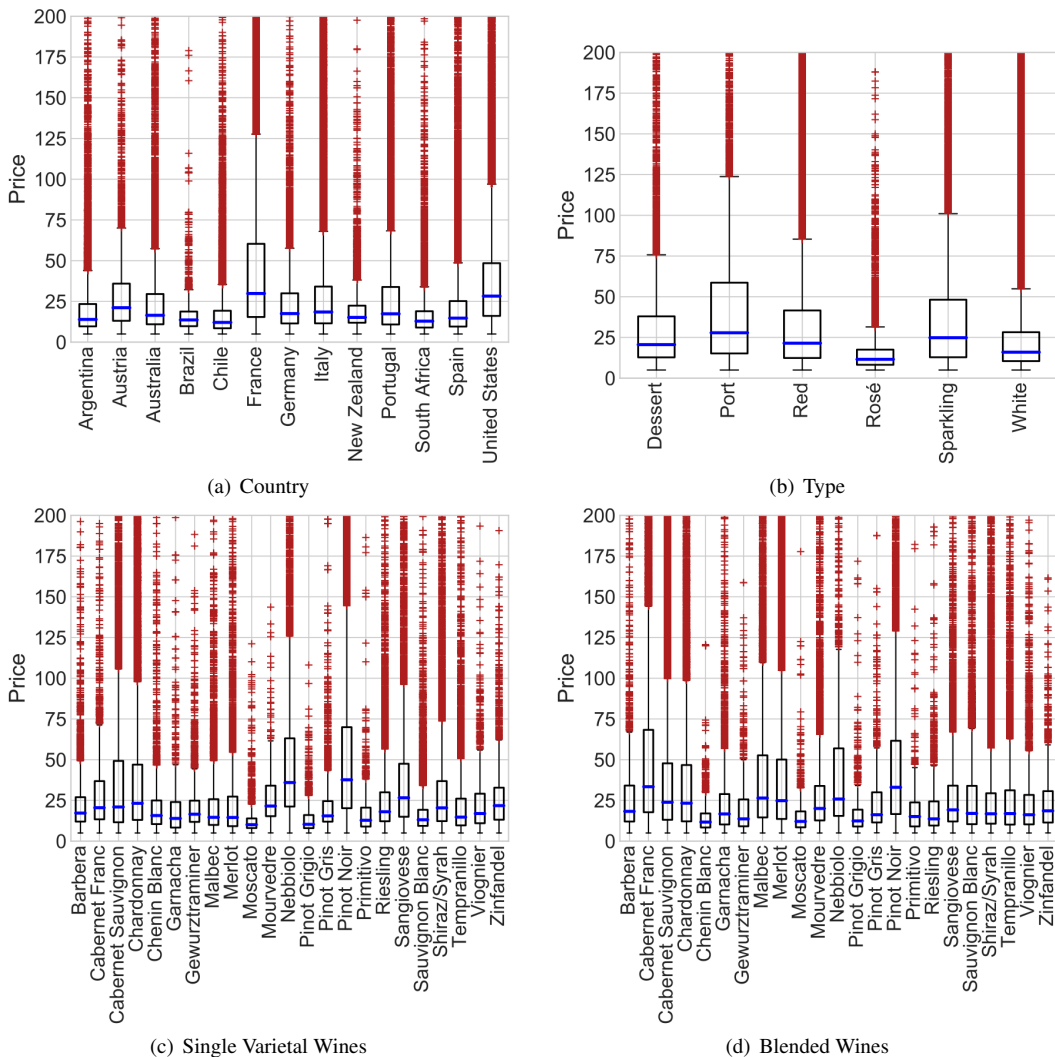
As mentioned, Vivino users assign ratings to wines on a scale between 1.0 and 5.0 stars, with a 0.5 granularity. We now analyze the Vivino wine ratings, investigating the relationship between them and prices, origins, types, style, and grapes.

### 4.1 Rating Trends

**Evolution over time.** Expert wine reviews have at times been criticized for a perceived inflation in wine reviews, i.e., that the

preferences of wine suppliers for highly rated wines influences the reviews produced by experts [21]. (Naturally, the higher the rating of a wine, the greater the demand, which, according to market forces, also drives up the price.) This trend has also somewhat manifested itself in the fast growing number of wines receiving a perfect score on the Robert Parker scale, which went up from 17 in 2004 to 38 in 2009, 103 in 2013, and 69 in 2014. Calls for new evaluation methods for wine quality have been increased [20]. By contrast, we find that, regardless of type, country of origin, varietal, or blend, “older” wines, i.e. those produced between 1960 and 2000, are preferred to “newer” wines, i.e., those produced after 2000. Across all wine variables: type, country, varietal, blend, there is a decline in average ratings between 2000 and 2010, but begin to increase again for 2016 vintages.

Grapes (both varietals and blends) and countries also show a similar ratings decline over the years. However, of the four most popular wine-producing countries: France, Italy, Spain and the United States, wines from the latter show the least decline (from 3.87 to 3.83 a decrease of 0.04 in the period between 2000 and 2010), whereas, Italy wines show a decline of 0.21, French wines experience a decline of 0.26, and Spanish wines fall in popularity by 0.17 stars. As we discuss later, this is partially due to biases of Vivino users, and in particular American users, for local wines. One reason for the general



**Figure 3:** Average prices distribution for Vivino wines with average prices in the interval  $[\pounds 5, \pounds 1M)$ . The plots have been truncated at  $\pounds 200$ .

trend in falling ratings could be that users have a more positive perception of older wines. Another reason is that as the number of users on the social network has grown, the distribution of wine ratings may have changed to reflect a more diverse (and perhaps discerning) user base.

## 4.2 Wine Ratings and Prices

We first examined the trend in prices across (a) country, (b) type, (c) varietals, and (d) blends as shown in Figure 3. Price listings on Vivino are based on averages generated from a limited set of vendors and the price data for wines has a very low frequency (see Section 3). Therefore, we cannot guarantee the reproducibility of these results, however, our analysis can provide an indication of the correlation between price and ratings. We find that prices vary substantially across wine variables, in particular vintage, as expected older wines are more expensive than newer ones. Similarly, wines from specific countries: France in particular, but also the United States are more expensive than others.

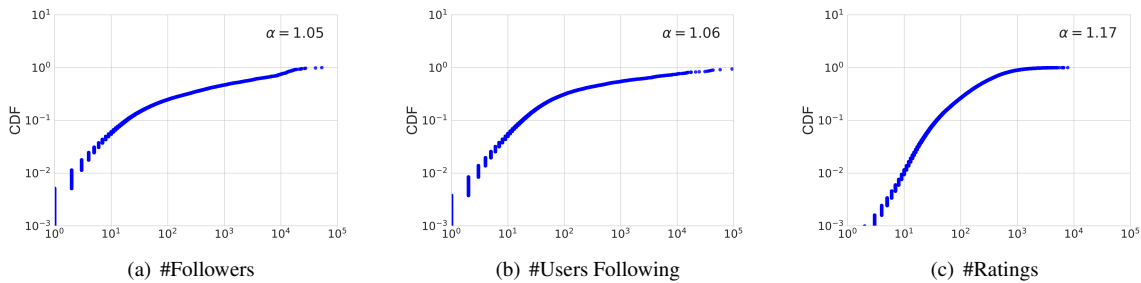
Next, we set out to test whether the average price (in GBP)

of the wine correlates to high average ratings. We found no evidence to support a relationship between the price of a wine and its average rating i.e. that more expensive wines receive higher scores than less expensive wines (see Figure 3).

We split the wine prices into quartiles, and looked for evidence of discrepancies in ratings between the quartiles with a minimum of 10 and a maximum of 914.9k. Q1, Q2 and Q3 having the values 15.73, 24.99 and 46.40. When grouped by both country of origin and wine type, only a slight difference in the average ratings is observed. The median rating for all wine types is within 0.1 star for all four price ranges. Similarly, for all wine types there is no difference in the interquartile range, apart from rosé wines, which have a smaller interquartile range and maximum for the upper most quartile. Also, for countries price has no significant effect on ratings.

## 4.3 Use of Language on Vivino

For our analysis of language employed by Vivino users, we examine the biographies and reviews of users. User biographies are tag-lines and short snippets of text, which appear on



**Figure 4:** Distributions for the top 10k Vivino users from countries with the highest levels of wine consumption.

#	Comment n-gram	Freq.	LR	Reply n-gram	Freq.	LR
1	('full', 'bodied')	1916	2455.13	('long', 'finish')	93	141.959
2	('long', 'finish')	1574	1976.26	('full', 'bodied')	88	136.86
3	('easy', 'drinking')	1244	1773.29	('pinot', 'noir')	51	97.2578
4	('well', 'balanced')	1574	1602.5	('thanks', 'sharing')	77	82.3297
5	('pinot', 'noir')	854	1584.19	('easy', 'drinking')	49	74.0608
6	('easy', 'drink')	941	1198.65	('looking', 'forward')	37	72.807
7	('good', 'value')	1499	1139.07	('black', 'cherry')	52	69.0606
8	('green', 'apple')	577	1126.39	('medium', 'body')	48	68.653
9	('medium', 'bodied')	788	922.062	('well', 'balanced')	68	68.1572
10	('fruit', 'forward')	807	831.71	('dark', 'fruits')	44	58.9347

**Table 2:** Top 10 n-grams from wine reviews comments and replies ranked according to Likelihood Ratio (LR).

a profile page, they often state a user’s motivation for joining the social network and occasionally provide a link to the individual’s personal or professional websites. User wine reviews, which are typically accompanied by a quantitative score, vary from single word comments, stating either approval or disapproval, to lengthier descriptive texts, which outline a user’s experiences of a wine. These usually detail the taste and appearance of the beverage, and suggest dishes which complement the wine and from time to time describe occasions where the wine could be served.

#### 4.3.1 User Biographies

Only 16,045 (11.6%) of the 137.5k high-ranking users have biographies. This finding, coupled with the power-law distributions plotted in Figure 4, which show  $\alpha = 1.05$ ,  $\alpha = 1.06$ , and  $\alpha = 1.17$  for followers, users following, and ratings respectively, indicates that wine ratings, and not social connections, are the main motivation behind users’ interactions on the platform.

We find that 683 biographies contain the word *sommelier*; 434 biographies abbreviations *WSET*, *N2*, *N3*, *N4* or *N5* (names of professional wine tasting qualifications); 120 contain the word *expert*; 33 the phrase *wine expert*; 135 simply the word *professional*; 15 the bigram *wine professional*, and 94 mention *ABS*, the wine agencies group. Conversely, 300 users describe themselves as *amateur*, 159 as *learning*, and 9 think of themselves as a *learner*. Interestingly, significantly more Vivino biographies contain keywords implying that users are wine experts, than those implying they are novices. However, the sample size is relatively small, and therefore few findings can be confidently drawn from these data. Furthermore, 274 biographies contain a web address, of these, 114 are email addresses. Thus, it would be fair to assume that these users are looking to

forge professional connections, rather than social ones.

We also examine how user biographies vary across region, ranking and number of ratings contributed. Biography unigram and bigram frequency distribution does not vary significantly by user ranking. However, it does when users are grouped by the number of ratings contributed. For users with fewer ratings, their bios typically have a high frequency of the words *learning*, *student*, *amateur* and *enthusiast*. On the other hand, for users who have contributed more than 1,000 wine ratings on Vivino, the frequency for *enthusiast* is significantly lower. These users are much more likely to describe themselves as *certified*, having a WSET qualification and provide their email address in their biography.

#### 4.3.2 Wine Reviews

We analyze reviews published between the dates 8 September 2012 and 27 March 2017. These comments, which we call reviews in this section, are also associated with replies or feedback posts. There is a one-to-many relationship between reviews and replies. And although many reviews do not receive a reply, on average a review has 1.2519 replies. The contents of replies are not of as much interest to us as the content of reviews: typically, replies express agreement or gratitude as shown in Table 2. In fact, the two most common bigram collocations for replies are (*dear*, *thank*), with a likelihood ratio of 77.2083, and (*Thanks*, *sharing*), with a likelihood ratio of 69.8736. Recall that a collocation is a sequence of tokens which of appears with a high probability in a text, and can be determined, e.g., using the Natural Language Toolkit (NLTK) Python package. We use the likelihood ratio association metric to determine bigram collocations. Mean word count is 64.05, the standard deviation is 37.33. Although the written

Regression Model	Train $R^2$	Test $R^2$	Train MSE	Test MSE
MLP	0.645	0.640	0.0459	0.0469
DT	0.701	0.608	0.0383	0.0520

**Table 3:**  $R^2$  Scores and mean squared error (MSE) for wine ratings prediction models for a 75% train set and 25% test set.

reviews provided by Vivino users are short in length, they contain a wide range of vocabulary. In particular, the words employed by users to describe wines are taken from a lexicon commonly employed by professional wine critics [15]. The most frequently occurring descriptors are: *good, nice, great, smooth, dry, fruity, light, red, sweet, well*. The most common unigrams include *tannins* and *acidity* - which are among the most common words used in professional reviews [15].

## 5 Wine Prediction Models

Next, we set out to build two predictive models: (1) a regression model for predicting the average rating of a wine; and (2) a classification model for predicting the taste profile or reviews history of a Vivino user. For the former, we evaluate the performance of both a Decision Tree (DT) regressor and a Multilayer Perceptron Neural Network (MLP), while, for the latter, we use a Support Vector Machine (SVM).

### 5.1 Predicting Wine Ratings

First, we develop feature engine for the model as part of the pre-processing stage. The task of constructing a good feature representation is essential in order to train supervised learning models effectively. As we aim to implement a generalized model, with capabilities of predicting the average rating of any given wine, we discount the following wine and winery specific columns of the wine dataset: wine name and winery name. We find that, even though there are recurring noun phrases in the wine names, these noun phrases either referred to the grapes, which constituted to the wine or the vintage or the wine’s type – data which is already present in other columns of the of the wine dataset. We also exclude, from the feature representation, wines with less than 75 ratings. One-hot encoding is used to represent the categorical data: region, country, wine type (e.g. red, white, sparkling, port), regional style, food pairings and grapes. The outcome of the preprocessing stage is a feature representation with 4.4k dimensions. Univariate feature selection is then employed to reduce the dimensions to 1k. The combined train and test sets account for 124,397 wines. Finally, we use a train-test ratio of 75% : 25% for both the MLP and DT regressor models.

Although DT regression outperforms the MLP approach on the training set, the MLP shows the best generalization on the test set. The accuracy metric chosen to compare the two models is the coefficient of determination. This regression model error metric is commonly denoted as  $R^2$ [11], where:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The maximum value for  $R^2$  is 1.0. Negative values can also be produced for poorly performing models.

#	Decision Tree (DT) Feature	Gini
1	average price	0.82400
2	number ratings	0.02160
3	food pairing: poultry	0.00954
4	grape: Zinfandel	0.00678
5	regional style: Italian Vino Nobile Di Montepulciano	0.00588
6	regional style: Austrian Pinot Gris	0.00502
7	regional style: New Zealand Chardonnay	0.00420
8	regional style: Italian Ripasso	0.00392
9	regional style: Israeli Syrah	0.00320
10	regional style: Spanish Montsant Red	0.00314
11	regional style: South African Malbec	0.00314
12	food pairing: Cured meat	0.00286
13	regional style: South African Cabernet Franc	0.00249
14	grape: Cabernet Franc	0.00232
15	regional style: Northern Italy Muller Thurgau	0.00226
16	regional style: Australian Viognier	0.00222
17	regional style: Greek Amyndeon Red	0.00214
18	country: South Africa	0.00213
19	grape: Mourvedre	0.00207
20	regional style: South African Merlot	0.00207

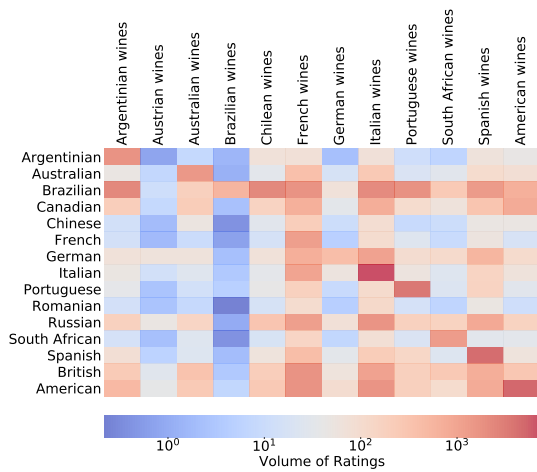
**Table 4:** Gini importance for Decision Tree (DT) features used to train wine ratings predictors, in descending order of importance.

$R^2$  for both methods is shown in Table 3. The DT model achieves an accuracy of 70% on the train set, whereas the MLP achieves a lower accuracy score of 64%. When the models are performed on the test set, the accuracy of the DT model decreases by 15.4%, whereas the MLP model only decreases by 0.744%. This is most likely because the one-hot coded representation of the wine features allows the MLP to generalize to unseen data more effectively.

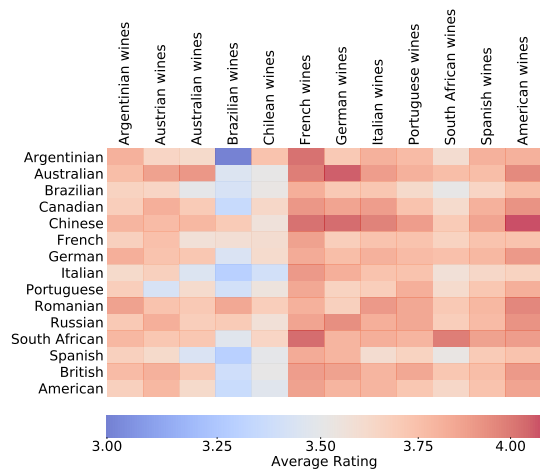
Table 4 shows the Gini index values of the ten most important features used to train the Decision Tree (DT) model. The feature with the highest Gini index is *average price*, which is far greater than that of *number ratings*, i.e., the second most important feature. The other eight features represent one hot-encoded features – one grape, one food pairing, and the remaining five are related to wine regional styles. The least important features are *region: Dao* (998th), *year: 1858* (999th), and *region: Knights Valley* (1000th).

### 5.2 Predicting Wine Preferences

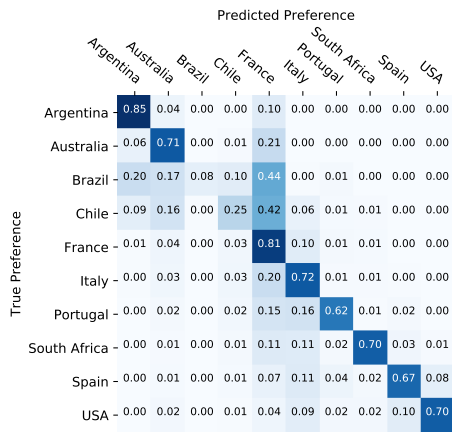
Next, we attempt to use the data analysis in order to create a wine style preferences classifier. Before we begin to design this model, we first extend our wine ratings analysis, with a particular focus on users. From this analysis, we learn that users show a preference for rating wines which originate in their home countries. Figure 5 shows that for of 60% the countries, Vivino users are mostly likely to rate a local wine. For the 40% for whom this is not the case, users either have a strong preference for rating wines that are produced in neighboring countries or European wines. Most commonly rated wines are not necessarily the most highly rated wines, as shown in Figure 6. Brazilian wines receive the lowest volume of ratings from all Vivino users, apart from Brazilian users themselves.



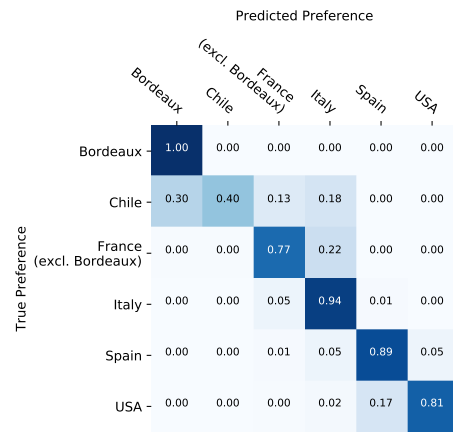
**Figure 5:** The total volume of ratings of wines from users from 15 countries with the highest levels of wine consumption in 2015 [13].



**Figure 6:** The average rating of wines from users from 15 countries with the highest levels of wine consumption in 2015 [13].



**Figure 7:** Confusion matrix for user preferences classifier based on the 10 countries with the most ratings in the wine dataset.



**Figure 8:** Confusion matrix for user preferences classifier based on the countries/regions with the most ratings in the user taste profiles.

As well as receiving the lowest volume of ratings, Brazilian wines also receive the greatest proportion of ratings below 3.5. French wines receive the greatest proportion of highest ratings: users from all countries with the exception of British and Russian users. Aside from the US, New World wines (from Brazil, Chile, Argentina, Australia and South Africa) receive lower ratings than Old World wines. Argentina receives higher ratings than other New World countries.

Preferences are acquired from the user taste profile, a history of average ratings and number of wines rated of region styles a Vivino user has reviewed. Given the tendency of users from all countries to show more favorable ratings to wines from France, and less favorable ratings to wines from Brazil. The mean and standard deviation of both French and Brazilian wine ratings are chosen as features for the classifier. Each example in the user set is labelled as either a user with a preference for wines from Chile, France (excluding the Bordeaux region), Bordeaux, Italy, Spain or US. These countries are chosen as their wines accounted for the majority of the taste profile sam-

ples. A decision tree ensemble method is used to classify user preferences. The mean accuracy of the model was 79.8%.

We develop two versions of the classifier:

1. A model differentiating between the preferences for wines from the 10 most commonly rated countries (from the wine dataset): Argentina, Australia, Brazil, Chile, France, Italy, Portugal, South Africa, Spain, and United States. The accuracy for this model is 71.8%.
2. A second model based on the country of origin of wines most frequently reported on in the taste profiles in the user dataset. These countries and regions are Bordeaux, Chile, France (excluding Bordeaux), Italy, Spain, and United States. This model achieves a higher accuracy, namely, 79.8%. The confusion matrices for these two models are shown in Figure 7 and Figure 8.

We also report the confusion matrix for the classifier in Figure 8. The confusion matrix shows that with the exception of wines from France (excluding Bordeaux), which are most often confused for Bordeaux, all other wine preferences are most



often confused for wines from France (excluding Bordeaux).

## 6 Conclusion

In this paper, we examined how user biases, regional factors and wine characteristics affect wine ratings on Vivino, a social network application for wine enthusiasts. Our analysis found that the biggest indicator of how a user will rate a wine is its vintage and region, i.e., there is a strong preference for French wine styles, and users also have a more favorable opinion of home-grown wines. Furthermore, our work shows that on Vivino users wine preferences are not based on price. This is in contrast to the bias researchers have found exists among professional critics [21]. Language analysis also shows that Vivino users generate relatively high quality user reviews, showing that wine ratings and reviews produced by amateur wine enthusiasts can be quite useful.

We also aimed to explore how wine characteristics, natural language analysis of wine reviews and regional analysis of wineries on Vivino could be integrated to develop a practical model for predicting wine reviews. We developed two predictions based on average wine ratings and users' ratings histories (which wines users rate and how they rate them). The models were evaluated using unseen examples from the dataset to gauge their efficacy. The results demonstrate that there is consistency across the ratings given by Vivino users, thus, spam and/or troll content does not affect the credibility of ratings on the social network. Our model also shows that wine ratings are not random, but the ratings assigned to wines by users based on informed and considered decisions. Overall, we believe that the analysis of factors that influence wine ratings and the development of models for predicting wine reviews are useful contributions to the understanding of how specialist social media platforms influence and shape our eating and drinking habits, and how we can minimize the subjectivity of online food and drinking ratings.

On the other hand, we acknowledge that our models do have some limitations. For instance, the user data collected was not of the same granularity as the wine data, and this may be one reason for poorer performance of our ratings profile classification compared to the wine ratings regression. The models produced were only tested on Vivino datasets, thus, as part of future work, we plan to evaluate them on user and wine reviews data from other sources.

## References

- [1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. In *CHI*, 2015.
- [2] P. Atkins, I. Bowler, et al. *Food in society: economy, culture, geography*. Routledge, 2016.
- [3] M. Chorley, L. Rossi, G. Tyson, and M. Williams. Pub crawling at scale: tapping untappd to explore social drinking. In *ICWSM*, 2016.
- [4] European Food Information Council. Why we eat what we eat: Social and economic determinants of food choice. <http://bit.ly/2Dx8RpA>, 2018.
- [5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *ACM CCR*, 1999.
- [6] R. Ghewari and S. Raiyani. Predicting cuisine from ingredients. *University of California San Diego*, 2015.
- [7] J. D. Guidry, M. Messner, Y. Jin, and V. Medina-Messner. From #mcdonaldsfail to #dominossucks: An analysis of instagram images about the 10 largest fast food companies. *Corporate Communications: An International Journal*, 20(3), 2015.
- [8] S. Higgs and J. Thomas. Social influences on eating. *Current Opinion in Behavioral Sciences*, 9, 2016.
- [9] Y. Mejova, S. Abbar, and H. Haddadi. Fetishizing Food in Digital Age: #Foodporn Around the World. In *ICWSM*, 2016.
- [10] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. #FoodPorn: Obesity patterns in culinary interactions. In *ICDH*, 2015.
- [11] N. J. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.
- [12] OIV. Global economic vitiviculture data. <http://www.oiv.int/public/medias/2256/en-communique-de-presse-octobre-2015.pdf>, 2015.
- [13] OIV. State of the vitiviculture world market. <http://www.oiv.int/public/medias/4710/oiv-noteconjmars2016-en.pdf>, 2016.
- [14] R. M. Parker and P.-A. Rovani. *Parker's wine buyer's guide*. Simon and Schuster, 2002.
- [15] M. Puckette. 40 wine descriptions and what they really mean. <http://winefolly.com/tutorial/40-wine-descriptions>, April 2012.
- [16] S. Sajadmanesh, S. Jafarzadeh, S. A. Ossia, H. R. Rabiee, H. Haddadi, Y. Mejova, M. Musolesi, E. De Cristofaro, and G. Stringhini. Kissing Cuisines: Exploring Worldwide Culinary Habits on the Web. In *WWW Web Science Track*, 2017.
- [17] W. Searcher. The World's Top 50 Most Expensive Wines. <http://www.wine-searcher.com/most-expensive-wines>, 2017.
- [18] S. S. Sharma and M. De Choudhury. Measuring and characterizing nutritional information of food and ingestion content in instagram. In *WWW*, 2015.
- [19] T. H. Silva, P. O. de Melo, J. Almeida, M. Musolesi, and A. Loureiro. You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. In *ICWSM*, 2014.
- [20] M. Snipes and D. C. Taylor. Model selection and akaike information criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3(1), 2014.
- [21] A. Teeter. How a ratings system designed to help consumers is tearing the \$40 billion dollar wine industry apart., Nov. 2014.
- [22] L. Tiu Wright, C. Nancarrow, and P. M. Kwok. Food taste preferences and cultural influences on consumption. *British Food Journal*, 103(5):348–357, 2001.
- [23] M. Trevisan. Wine and society. *Wine Studies*, 1(1), 2011.