In Vi(vi)no Veritas? Expertise, review accuracy and reputation inflation

Rebecca Janssen*, Matthew Ribar†

August 25, 2023

Abstract

Review systems including quantitative measures as well as text-based expression of experiences are omnipresent in today's digital platform economy. This paper studies the existence of reputation inflation, i.e. unjustified increases in ratings, with a special focus of heterogeneity between experienced and non-experienced users. Using data on more than 5 million reviews from an online wine platform we compare consistency between numerical feedback and textual reviews as well as sentiment measures. We show that overall the wine platform displays strongly increasing numerical feedback over our time period from 2014 to 2020 while this is not the case for our control measures. This gap appears to be even stronger for users with less experience or expertise in wine reviewing. We conclude, that online platforms as well as potential customers should be aware of the phenomenon of reputation inflation and simplifying feedback to one number might do a disservice to review platforms' goal of providing a representative quality assessment.

JEL Classification:

Keywords:

*ZEW Mannheim; P.O. Box 103443, D–68034 Mannheim, Germany. rebecca.janssen@zew.de.
†Department of Political Science, Stanford University. mkribar@stanford.edu.

1 Introduction

Almost every digital online platform today uses score mechanisms. E-commerce, online streaming, travel agents, reviewing services, and car sharing services all include options to rate transactions. Some platforms use ratings for likes and community aspects, but many platforms which offer products or services use these score mechanisms to reduce information asymmetries for consumers. Providing past customer experiences to potential future customers delivers additional data points and is especially helpful when the quality of a product or service can be evaluated only after purchase. Consumers depend on these ratings and reputation systems (Li et al., 2013; Luca, 2016; Luca and Zervas, 2016) but these systems can also drive product sales (Archak et al., 2011; Jabr and Zheng, 2014; Moe and Trusov, 2011).

However, over the past few years, ratings distributions from many platforms have begun to skew more positive—to 'inflate.' This highly skewed distribution of online ratings is often referred to as a J-shape distribution (Zervas et al., 2021). Filippas et al. (2022) show that the share of workers with a perfect 5-star rating on an online marketplace for labor grew from 33 percent to 85 percent in only six years and describe similar patterns for other platforms. Similarly, the majority of Uber drivers have a perfect five star-rating (Athey et al., 2019). Do these changes reflect a real increase in user satisfaction or merely "reputation inflation", i.e. better ratings without higher satisfaction? Filippas et al. (2022) decompose much of the upward trend in ratings to the latter cause.

Are all users on these platforms equally susceptible to unsubstantiated reputation inflation? The existing literature treats users as largely homogeneous, but on many platforms, some users are more experienced or knowledgeable than others. Such experts may be less prone to social pressure to give better ratings and more focused on their objective quality evaluation. Furthermore, experts — defined by an educational criteria or on a status as an influencer with a large network — may face a higher risk of prestige loss when leaving ratings which fail to match real quality assessment. Both platform users and platform providers could benefit from insights about heterogeneity in reputation inflation. The former could rely more on expert reviews if there is evidence for more consistent rating behavior in this group; the latter could adjust their internal metrics of performance or quality based on the consequences of reputation inflation and deliberate

expert definitions.

Score mechanisms often rely on quantitative measures like star ratings or school grades, supplemented by written text reviews. Numeric feedback is easier to compare: it is generally a one-dimensional measure. Text reviews, on the other hand, are more complex to study. Length, tone and syntax are highly individual. At the same time, text reviews can contain detailed and granular information on a variety of dimensions considered by consumers: price, quality, alignment with description, service, and so on. Previous studies investigated the correlation of text-based feedback and its numeric counterparts, as well as the added value of reviews for quality and rating score prediction (Katumullage et al., 2022; Klimmek, 2013; McCannon, 2020; Yang et al., 2022). Their overarching results are threefold: product quality can be related to several review characteristics; there is a positive correlation between text length and product price; and text reviews are a better source of information for quality prediction than pure numerical input.

We leverage written reviews and numerical ratings from experts and non-experts on the world's largest online wine marketplace. Specifically, we use differences in quantitative and written feedback to identify reputation inflation. We use a variety of machine learning techniques to predict quantitative ratings based on the written feedback. In doing so, we compare numerical ratings with predicted quality based on text reviews, similar to Filippas et al. (2022). Furthermore, we check for robustness with developments in reviews' sentiment.

We show that numerical ratings increase over time. The average rating of 3.67 in 2014 increased to 3.86 in 2020, reflecting an increase by 4.3 percent. In contrast, when we predict the ratings of wines based purely on their written reviews, the average wine scores 3.77 points in 2021 and 3.76 points in 2014.[1] Comparing those values to predicted quality ratings as well as sentiment measurements we find strong evidence for reputation inflation on the platform. Our predictions are more accurate for experts: the root mean squared error (RMSE) of the model is 0.50 for experts and 0.60 for non-experts, which suggests that expert reviews are indeed more informative. Importantly, however, we find that reputation inflation is even more severe for experts than for non-experts. Among experts, the average rating increased 0.20 points from 2014 to 2020; among non-experts,

[1]These figures are from the 20 percent of the data we held out as a test set. Expert is defined here as having an average number of comments per review above the 80th percentile.

this increase was only 0.15 points.

The remainder of the paper proceeds as follows. Section two describes background information about reputation inflation and expert status on platforms. Section three introduces our raw data and details on the chosen expert classification. Section four introduces our empirical strategy. Section five presents our main results and section six concludes.

2 Background

2.1 Reputation inflation in online marketplaces

The most common form of product reviews online are numerical ratings. Users assign a rating to the product or service. These ratings are often expressed in a star format ranging from one to five. A one star rating represents the lower bound, i.e. a negative review, and five stars the most positive review possible (Yin et al., 2016). While numerical feedback measures reflect a user's overall evaluation of the product or service and serves as a signal of quality and value to the customer (Li and Hitt, 2010), textual reviews bring additional information valuable to potential buyers. Textual reviews can include additional thoughts, detailed reports about different aspects of the product experience, and often some kind of emotional tone (Li et al., 2019).

Highly skewed ratings distributions with a large share of good or perfect scores in online marketplaces' rating systems are common. Filippas et al. (2022) show that 80 percent of the evaluations on their online labor market have a rating of at least 4.75 (out of 5.00) and list other marketplaces with comparable results. On eBay, the percent positive measure for sellers is 99.3 percent on average (Dellarocas and Wood, 2008; Nosko and Tadelis, 2015; Tadelis, 2016). Almost 95 percent of Airbnb properties are rated at 4.5 or 5 stars (Zervas et al., 2021). Zhu and Liu (2018) show similarly high seller ratings on Amazon (Zhu and Liu, 2018). Almost 90 percent of UberX trips are rated at the maximum of five stars; to leave any lower review signals a problem (Athey et al., 2019).

A large and increasing share of fake reviews online may account for these overwhelmingly positive reviews (Glazer et al., 2021; He et al., 2022; Luca and Zervas, 2016). Beyond an increasing number of fake or inauthentic reviews, bias within the pool of reviewers or within the review content could decrease the ability of these reviews to reduce infor-

mation asymmetries. This disservice to the initial benefit of review systems leaves users without helpful experience reports from other customers.

Another explanation for the positive skew of online ratings may be an actually increasing level of satisfaction with the services provided. However, it seems unlikely that, for example, UberX drivers have unambiguously become better drivers over time. Using objective telemetry data Athey et al. (2019) show that UberX customers prefer rides with fewer sharp accelerations or brakes—and that they rate drivers accordingly. A more likely explanation is review inflation: an increase in customer ratings over time which is not justified by increased satisfaction (Filippas et al., 2022). Potential explanations for review inflation include manipulated positive ratings, self selection within the group of reviewers, social pressure where users feel pressured to leave positive ratings, or incentive alignments on the platform which nudge people for positive ratings.

How can we decompose the this increasingly positive skews into its constituent parts? One option would be to compare these subjective ratings to a feedback mechanism that is less susceptible to inflation. Text reviews, for example, do not suffer many of the same pressures for higher reviews that numerical reviews do. Ironically, it is the ambiguity of text reviews that make them less susceptible. While written reviews generally express either positive or negative opinions, this positivity is a latent dimension. People may feel less guilty leaving a review that says "this wine lacks acid" than leaving a one-star review of the wine.

To summarise, we expect numerical ratings to increase over time. However, much of this inflation will be unrelated to actual product quality. As a result, the predictions of these ratings will remain steady over time. More formally, we can distill the following hypothesis:

H1: Numerical ratings of wines will increase over time.

H2: The difference between numerical ratings and the predicted values of ratings from text reviews will increase over time.

2.2 Expert status in online marketplaces

One might also expect expert users to be less susceptible to the review inflation phenomenon. Professionalism and expert status mean different things in different settings.

This not only holds for the reviewer side itself but also the products, services or persons rated on a platform. For example, Airbnb assigns "Superhost" status to specific hosts of accomodations based on guests' ratings; the physician rating platform Jameda not only offers physicians—who present themselves on the platform and are rated by patients—to buy special premium memberships which come along with profile badges but also assigns badges like on Airbnb for "Top10" or "Top5" within a specific region and area.

Because the platform we study rates products, we focus on the expert status of reviewers. Herein, professionalism can take on characteristics of quality measures or influence. Several platforms, such as Amazon, provide the opportunity to users to evaluate reviews of others as helpful or leave likes as some kind of agreement. The Q&A website Stackoverflow uses helpfulness measures or evaluations of the best answer to assign reviewers special badges and reputation measures. The area of influence can often be quantified by the number of followers of users where applicable which might also speak in the direction of expertise on rating platforms.

Experts may be less susceptible to social pressure to give positive reviews because they have an incentive to provide useful reviews. First, Vivino is a sufficiently prominent platform that many aspiring wine critics use it to attempt to grow their reach. Many expert reviewers within our sample link their Vivino profiles to their external websites—suggesting that the users are attempting to drive traffic to their own page. Experts may also feel an intrinsic reward for increasing their follower count.

Experts are likely to provide more informative reviews. In our context, this increase in information is likely to reduce the error in predicting ratings from text reviews. Moreover, if experts are indeed less susceptible to review inflation, then growth in prediction error over time—as text reviews remain unaffected but numerical ratings increase—should be lesser among Vivino users we label as experts. These conclusions generate two additional hypotheses:

H3: The prediction error of our model will be lower for reviews written by experts. .

H4: The growth in prediction over time will be lesser for reviews written by experts.

3 Data and Descriptives

3.1 Dataset

For our analysis we use large-scale, web-scraped data from Vivino. Vivino is the world's largest marketplace for wine. It is offered as an app as well as a web version. The platform serves as a community forum where users can interact and leave reviews and ratings for individual wines. They can also purchase wine on the platform. Importantly, many Vivino members use the platform simply to catalog their tasting notes on wines. Currently, over 62 million users are registered on the platform.

Vivino is a particularly interesting setting to study reputation inflation and experts' additional value for several reasons. First, its reviews are comparable: at the wine and vintage level, the products are more or less identical. Second, the platform is used by reviewers of different expertise and knowledge levels about wine, which allows differentiation and investigation of heterogeneity effects. Third, Vivino's market size permits a large number of observations for our analysis. Finally, the platform design includes information about followers, comments and likes for every reviewer.

One important caveat for wine is that products will evolve over time. Wine is a natural product; it ages. Specifically, a wine tasted in one year may not taste the same in subsequent years. Many wines—such as vintage port or classically made Bordeaux—'open up' and improve over many years in bottle. In contrast, wines like Beaujoulais Nouveau are thought to be best consumed immediately upon release. The overwhelming majority of wines listed on Vivino will not improve with age.[2] In our data set, the median wine is consumed three years after vintage date; 90 percent of wines are reviewed before they are eight years old.

Our random sample covers more than 600,000 different wines across almost 3 million vintage-wine combinations, where vintage means the year the grapes were harvested.[3] Our sample covers 11.8 million unique reviews. The data were collected during the summer of 2020. They include up to the 100 most recent reviews left for each wine. As our method relies on text analysis of reviews, we decide to use only a subsample of our

[2]In addition, the characteristics of aged wine are not universally appealing, meaning that review scores are not strictly increasing in wine age. For consumers who do prefer aged wine, more specialized platforms are available, such as cellartracker, a wine review software that allows users to manage their wine cellar.

[3]The oldest wine in the sample is an 1840s bottle of Madeira.

dataset. First, we drop 100,817 reviews which do not contain a written review. Second, for simplicity we restrict our analysis to English-language texts. We drop 6,081,342 tasting notes written in other languages, mostly in French, Spanish, German, Russian, and Portugese. These deletions leave us with around 5.45 million English-language reviews for which we have both the score (our outcome variable) and an informative tasting note. These reviews still cover 490,000 unique wines.[4]

3.2 Variables of interest

The following selection provides further details on the data used and variables of major interest for our analysis.

Information on reviews: To investigate how user heterogeneity affects reputation inflation, we leverage detailed information within Vivino's review system. Users can decide between leaving pure star ratings or adding a textual review. The star reviews are either integers or midpoints between them (i.e. 4.5) between 1 and 5. Figure 1 shows the distribution of scores in our data for both expert and non-expert reviews. The average rating for reviews included in our dataset is 3.77 (3.83 for non-experts, 3.72 for experts). The distribution of ratings is similar to other platforms: slightly skewed with a higher share of good reviews (Engler et al., 2015; Filippas et al., 2022; Hu et al., 2006; Luca and Zervas, 2016).

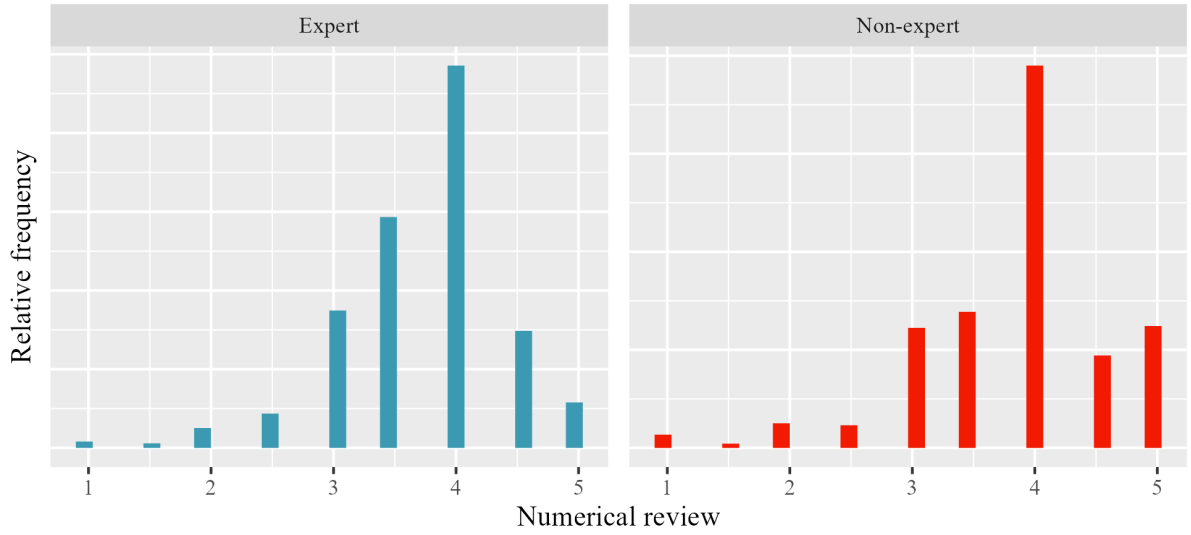
This numerical feedback for wine is our outcome variable—the rating whose inflation we measure. We compare these scores to written feedback, a tasting note or wine review. As we argue above, we consider the written feedback to be more time-invariant and less susceptible to review inflation. The average review is 106 characters long.

Ultimately the content of these reviews is quite varied. Some provide detailed tasting notes of the wine; other reviews are less informative. One example of a highly detailed review is "Violet and dark fruit on the nose. Initially boysenberry and cherry with strong tobacco. After 30 mins did it open up with flavors exploding. More fruit forward. Licorice, currants coming through. Long finish. Medium body. Doesnt feel 100% cab. Some tannic finish." Another example of a shorter but nevertheless informative review is "Fruity and off-dry with exotic fruits, citrus and minerals. 87."[5] On the other hand,

[4]When we refer to unique wines, we mean unique at the "label" level. 2008 and 2009 Dom Pérignon, for instance, would count as the same wine.

[5]The 87 here likely references a 100 point scale for wines, as popularized by the influential wine writer

Figure 1. Distribution of star ratings by expert level



Note: This figure distinguishes experts by whether they have a number of followers above the 80th percentile. This figure includes only the scores we use in later analysis—those with a written comment.

an example of an uninformative wine review is "Great with slow cooked lamb shanks." Many reviews also contain only details about when, where, and with whom the wine was consumed.

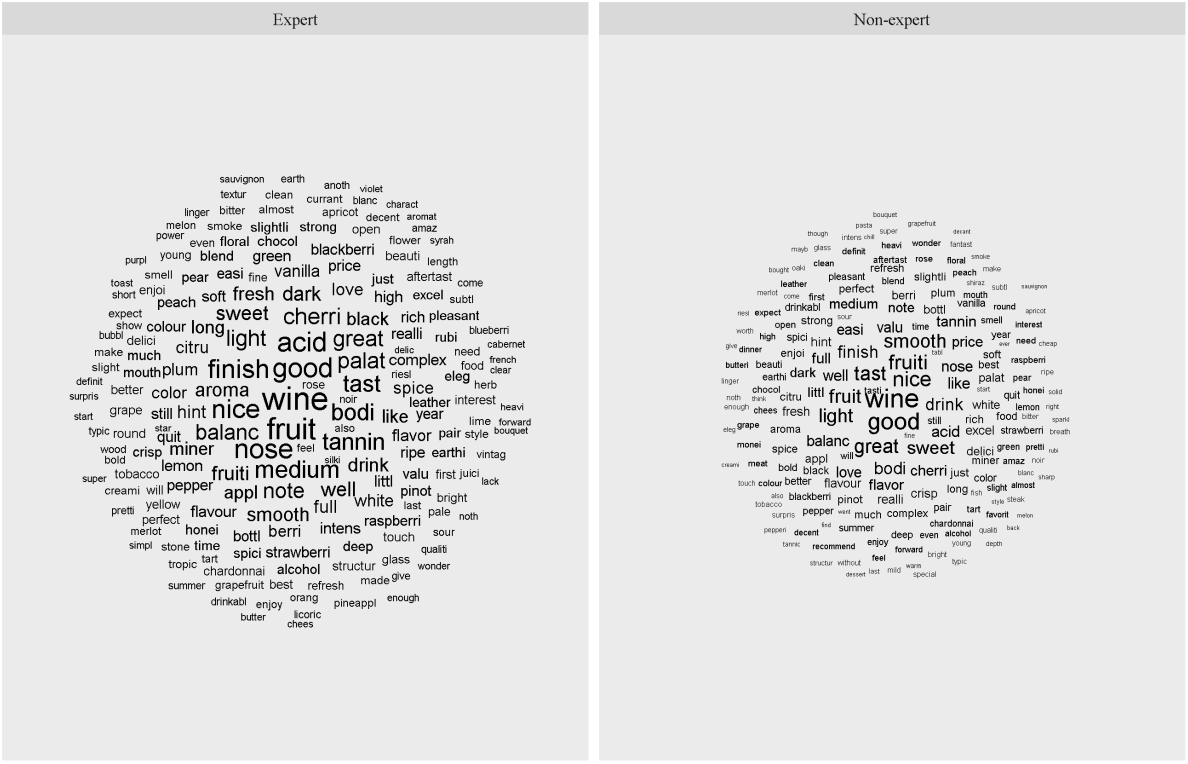
Examining word choice can begin to deliver valuable insights about differences and salience. Figure 2 shows basic wordclouds for all experts' and non-experts' reviews. These wordclouds are based on the frequency of stemmed words.

Non-experts appear to use words which are more vague or subjective in their reviews. For example, the word 'smooth' is the eight most common word in non-expert reviews and the 32nd most common word in expert reviews. Smooth is a word for which the meaning is often unclear; it could refer to mouthfeel, tannin, acid, fruitiness, or a number of other characteristics. In contrast, acid is the third most common word for experts (13th for non-experts). Tannin is the eight most common word in expert reviews, and the 16th most common word for non-experts. These terms are often paired with words like 'medium' (seventh for experts, 24th for non-experts) or 'balanced' (15th for experts, 20th for non-experts).

On the other hand, non-experts were more likely to used price-based terminology when describing wines. For non-experts, value was the 32nd most common term; for experts it was the 77th. For non-experts, price was the 22nd most common term; for experts,

Robert Parker.

Figure 2. Wordcloud of text reviews



Note: This figure distinguishes experts by whether they have above the 80th percentile of followers. This figure includes only the scores we use in later analysis—those with a written comment.

the 54th. While these observations are purely descriptive, they do start to suggest a pattern of wine experts using more precise language. This difference could explain why our models are better able to predict the scores for expert reviews: like many readers, the models do not understand what 'smooth' means.

Information on users and expert status: A strength of Vivino's review system is that it allows for interaction between different users. Users can like or comment on existing reviews. While this feature is rarely used, it nevertheless permits us to distinguish between expert and non-expert reviews.[6] We use two decision criteria to separate expert and non-expert Vivino users.

First, we define an expert as a user with a number of followers larger than the 80th percentile.[7] If we assume that users are more likely to follow an account that produces informative reviews, then experts will end up with the highest follower counts. Users may also follow accounts that review the wines that a particular user is most interested

[6]The median review has zero likes and zero comments.

[7]At the user-level, the 80th percentile of followers is five. The overall user-level median is one follower.

in. If a user is interested in Rieslings from upstate New York, they can likely find an account which reviews many of them. We believe that such specialized accounts are also likely to be experts. Second, we define experts as accounts with an average number of comments per review above the 80th percentile. However, comments can be either positive or negative, so we use the follower count as the key metric of expertise through the rest of the paper.

Our sample contains 5,448,634 reviews, of which 2,340,568 are written by non-experts and 3,108,066 are written by experts according to our first definition. This discrepancy implies that experts write more reviews, which is unsurprising if experts are the most active users of Vivino. Overall, there are 952,368 reviewers in our sample, of whom 167,740 are experts and 784,628 are not experts. In other words, this strategy identified approximately 17.6 percent of users as experts. It also means that the average expert user in our sample has reviewed 18.5 wines and the average non-expert user has only reviewed three wines.

An alternative measurement for expertise is whether the user linked a website on their Vivino profile. While listing a website suggests a strong commitment to wine tasting and the platform, it may also be aspirational. These reviewers may only plan on becoming experts; or list websites not related to wine tasting at all. We include analyses with this measurement for expertise in the appendix; the results are very similar.

Next to expert status, reviewers can be described by several other characteristics. Most of the users in our sample come from the US (40.5 percent), the UK (12.2 percent), Australia (5.3 percent), or Canada (5.2 percent). Like other social media platforms, the distribution of followers and people followed is rather uneven. The median user has one follower, but the mean user has 9.9. These numbers reflect that the distribution has a long right-hand tail: the maximum follower count in our sample is 71,926.

Information on wine characteristics: The products within our sample come from a variety of countries and regions. Two-thirds of the reviews refer to wines from either France (23.7 percent), the US (19.5 percent), Italy (15.9 percent) or Australia (7.1 percent).[8] 2.8 percent of reviews describe "natural" wines. For a subset of products we

[8]When we include the non-English reviews, most of the wine products come from France (27.4 percent), followed by Italy (20.0 percent), the US (11.2 percent), Spain (8.46 percent) and Australia (4.6 percent).

Table 1. Summary Statistics

	Non-Experts			Experts		
	Mean	**Median**	**N**	**Mean**	**Median**	**N**
Review level						
Rating	3.83	4.0	2340568	3.72	4.00	3108066
# Likes	0.24	0.0	2340568	11.34	1.00	3108066
# Comments	0.02	0.0	2340568	1.44	0.00	3108066
Text Length	71.91	50.0	2340568	139.12	104.00	3108066
Sentiment	0.30	0.3	2340568	0.26	0.25	3108066
$D_{Natural}$	0.02	0.0	2340568	0.03	0.00	3108066
$D_{Homecountry}$	0.33	0.0	2340568	0.27	0.00	3108066
D_{Expert}	0.00	0.0	2340568	1.00	1.00	3108066
User level						
# Reviews	12.34	5.0	784628	100.29	18.00	167740
# Ratings	23.70	9.0	784628	172.38	45.00	167740
$D_{Website}$	0.01	0.0	784628	0.07	0.00	167740
# Followers	0.89	0.0	784628	51.99	13.00	167740
# Following	1.40	0.0	784628	44.21	14.00	167740

Note: This table summarizes descriptive statistics of our main variables. While the upper panel has been calculated on our restricted review dataset, the panel at the bottom is based on user-level. The values are taken from the users' main profile page and do not necessarily coincide with statistics calculated from our data sample. For example, while an expert user in our data sample has written 18.53 reviews on average from the reviews used in our analysis, she has written 100.29 reviews on average overall.

also have information about the price.[9] The mean price listed for a product reviewed is \$45.00; the median price is \$40.00.

3.3 Descriptive Statistics

Table 1 summarizes main descriptive statistics for experts and non-experts. Experts leave lower ratings on average (3.72 vs. 3.83); their reviews also contain less positive sentiments (0.25 vs. 0.30). However, reviews by experts are longer on average, have more comments, and have more likes. At the user level, the results show differences in the quantitative measurement of user activity. While experts have left 172.4 ratings and 100.3 reviews on average, non-experts have only left an average of 23.7 ratings and 12.3 reviews. By definition, experts have more followers than non-experts, but they also follow more users.

[9]Prices displayed on the platform can either be market prices products are sold for on Vivino itself or prices reported by users who have bought the wine somewhere else.

4 Empirical Strategy

Our empirical strategy comprises three stages. Our goal is to compare the numeric ratings that users give wines to more 'objective' text based reviews. To do so, we use an ensemble machine learning strategy to obtain predicted values for the numerical ratings. These ratings are time-invariant in the sense that they do not take into account the year in which a review was written. We then compare these predicted values to the actual values to show how the latter drift over time. The next stage is to compare the predictions over time and within groups to illustrate the extent to which review inflation pervades both groups. Finally, we use sentiment analysis to further show that individuals rate wines better over time without expressing more positive sentiments in their written reviews.

4.1 Predicted Quality based on Text Reviews

Reputation inflation does not simply mean that ratings increase over time; rather, reputation inflation means that numerical ratings increase without a commensurate increase in the quality of products. To identify whether this discrepancy increases over time, we use the text reviews to predict a numerical rating. Reputation inflation would be associated with less consistency between the actual numerical feedback and the predicted value from the texts. Our specific quantity of interest is the root mean squared error (RMSE):

$$RMSE_{ijt} = \sqrt{\text{mean}\left(Rating_{ijt} - Rating_{ijt}^{pred}\right)^2} \tag{1}$$

for product i by user j at time t. The higher the discrepancy between the numerical rating and the predicted numerical rating based on text reviews, the more evidence for review inflation. Importantly, we regularly examine RMSE within subgroups, for within expertise categories or years.

Building the predicted text measure is a three step process. First, we use the Doc2Vec algorithm to extract features from the text reviews. Second, we use the resulting feature matrix in a variety of machine learning strategies to obtain predicted values. Third, we construct weighted and unweighted ensembles of these predictions.[10] We train the model using 80 percent of the data, leaving 20 percent to evaluate the model.

[10]The weighted ensemble is a convex combination of the predicted values; weights are the coefficients obtained by regressing the true value on the algorithm-specific predicted values while constraining the coefficients to sum to one. The unweighted ensemble is the mean of the predicted values.

For the feature extraction we considered three strategies. The simplest strategy is a term frequency (TF) matrix. After stemming the individual words in each review, this strategy creates a matrix which counts how many times the 1,000 most common words were used in each review. However, this strategy cannot distinguish between the relevant importance of different words. A term frequency–inverse document frequency (TF-IDF) strategy overcomes this problem. This approach takes the TF matrix and weights it by the word's 'importance,' measured by the logarithm of the total number of documents divided by the number of documents in which the term appears. However, like the simpler TF strategy, this approach has some flaws. It is a 'bag of words' approach in that it does not consider the associations between words or their orders.

After consideration, we adpot a feature extraction strategy called Doc2Vec (Le and Mikolov (2014)). Doc2Vec is an extension of a framework called Word2Vec, introduced by Mikolov et al. (2013). Like the other strategies, Word2Vec produces a vector space of features. Using a two-layer neural network, it transforms the inputs (i.e. the corpus of documents) into a feature space. However, Doc2Vec also includes some document specific information, making it more suitable for document-level prediction.[11] Unlike the bag of words strategies, this approach preserves relations between neighboring words. For the remainder of this paper, we report results obtained using features from Doc2Vec, as they produce the clearest results. However, the results are largely similar when replicated with features spaces extracted from the raw text using TF or TF-IDF strategies.

After extracting a feature space from the raw data, we then attempt to predict the numerical score by these features. Often it is unclear which machine learning strategy will perform best in any given space. As such, we try a variety of approaches. We first deploy three flavors of penalized linear regression: Ridge, LASSO, and an elastic net with a mixing parameter of $\alpha = 0.5$. All three models use cross validation to select the penalty which minimizes mean square error (MSE). We also use a random forest, a neural net, and an XGBoost (eXtreme Gradient Boosting) model. For ease of reading, we compare these models on their root mean square error (RMSE): the average distance of each prediction from the actual answer.

Table 2 shows that between these eight prediction methods, we can achieve an RMSE between 0.581 and 0.616 within the test set. In other words, the predictions made by our

[11] In this context, we treat each review as a separate document.

Table 2. RMSE by prediction method

Model	Training set	Test set
LASSO	0.616	0.616
Ridge	0.616	0.616
Elastic Net	0.616	0.616
Random Forest	0.617	0.620
XGBoost	0.561	0.581
Neural Net	0.724	0.725
Unweighted Ensemble	0.604	0.608
Weighted Ensemble	0.563	0.581

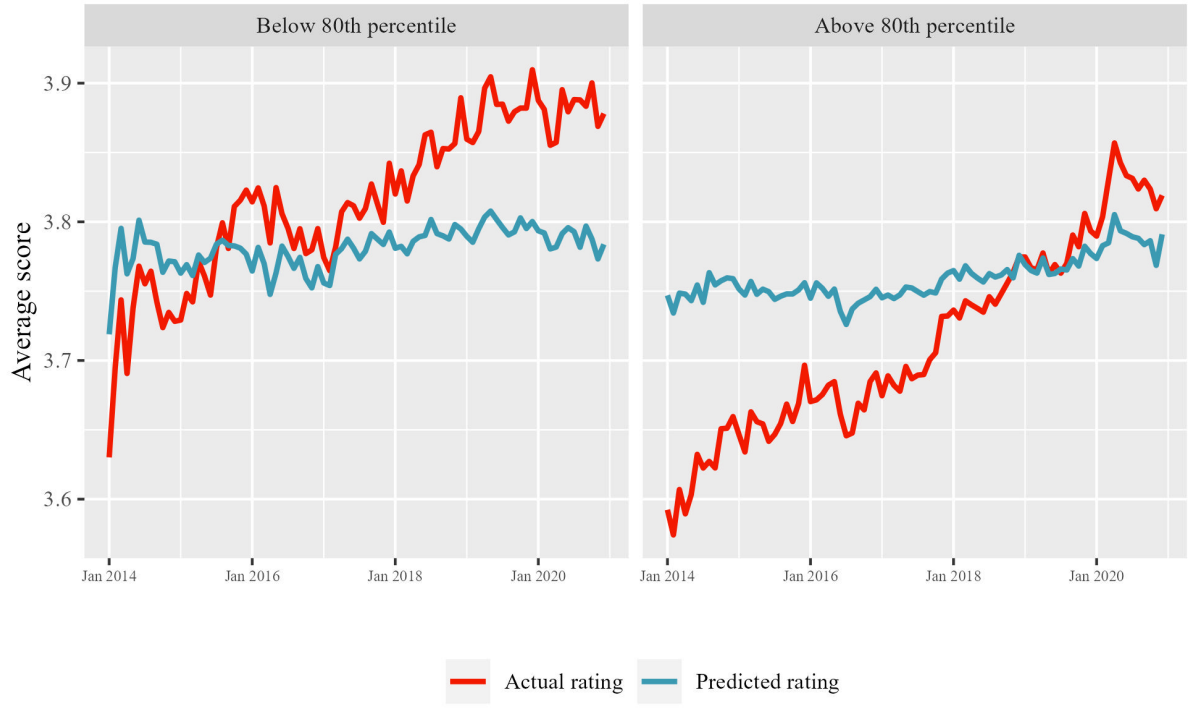
Note: We extract features using Doc2Vec. Expert is defined here as having above the 80th percentile of followers. We hold out 20 percent of data as the test set.

models are on average roughly 0.6 away from the true ratings. We can slightly reduce this error by using the weighted ensemble. We derive weights for the weighted ensemble using a linear regression of each prediction on the true value. The coefficients of this regression are constrained to lie between zero and one, and to sum to one. This gives weights of 0.09 for the LASS0, 0 for the Ridge regression and elastic net, 0.05 for the random forest, 0.86 for the XGBoost, and 0 for the neural net. Table 2 displays the RMSE for each prediction method, broken out by the training and test sets. We use the weighted ensemble method for the remainder of our analyses.

4.1.1 Sentiment Analysis

Beyond predicting quantitative ratings from the text of wine reviews, it is also important for our analysis whether the users actually liked the product in question. To this end, we turn to sentiment analysis of text reviews. As in the previous approach, review inflation would appear as a growing discrepancy between the numerical ratings of wines and the sentiment score we derive from the text reviews. Li et al. (2019) highlight the effect of numerical ratings on the one hand and textual sentiment on the other hand on product sales. Previous literature in that context had already shown that textual reviews impact product sales (Floh et al., 2013; Jabr and Zheng, 2014; Ludwig et al., 2013) as well as star ratings (Villarroel Ordenes et al., 2017). In addition to existing literature showing that in the case of online reviews the specific attributes food, service, ambience, and price are of major importance, Gan et al. (2017) find that consumers' sentiments in the attributes mentioned especially explained variation in star ratings. Kim (2021)

Figure 3. Actual and predicted wine ratings by user follower count



This figure breaks out actual ratings and the predicted ratings by whether or not the user has a follower count above the 80th percentile (5 followers). This figure includes only the scores we use in later analysis—those with a written comment. Predictions are obtained using the weighted ensemble method.

investigates the usability of sentiment scores for online reviews and concludes that these are less likely skewed to extreme values than numerical ratings.

Sentiment analysis also uses an ordered bag-of-words approach in which it calculates a sentiment for each word and uses a series of weights to aggregate up to a review-level sentiment. These weights account for so-called valence shifters, which can change the sentiment of a word.[12] We create a sentiment value for each written review in our dataset using the R package *sentimentr*. This package generates a positive value for reviews with positive sentiment and negative values for a negative sentiment.

5 Results

5.1 Prediction based evidence

Figure 3 shows how the average numerical rating users assign to wines increased between 2014 and 2020. The red line shows the average prediction, binned within months.

[12]For example, "this wine is not good" and "this wine is good" are two opposite reviews. In this case, the word "good" carries a positive sentiment, but the word "not" which precedes it shifts the valence.

The figure also distinguishes between expert and non-expert users, the cut-off being whether the user's follower count is above the 80th percentile. Across both subgroups, the story is consistent.[13] Among non-experts, the average score increases from 3.63 in January of 2014 to 3.88 in December of 2020. Among experts, the average score increases from 3.60 in January of 2014 to 3.82 in December of 2020. These figures represent a 6.8 percent percent and a 6.3 percent increase, respectively. These results are strongly consistent with our first hypothesis: the average rating increases over time.

Importantly, Figure 3 is also strongly consistent with our second hypothesis. At the beginning of our study period in January 2014, predicted ratings were on average 0.15 points higher than actual ratings for experts and 0.09 points higher than actual ratings for non experts. However, as time passes, the actual ratings increase while the predicted ratings stay constant. At the end of our study period in December 2020, the predicted rating was on average 0.03 points *higher* than the actual rating for experts and 0.09 points higher for non-experts. Contrary to our third hypothesis, review inflation is present in both. Experts do not appear immune to review inflation—they are equally subject to it. These results suggest that both experts and non-experts are consistently give higher scores for wines over time despite providing similar qualitative descriptions.

Figure 3 shows clear evidence of review inflation over time. Our prediction model does not take the date of the review or the score given by the user into account. As a result, even while the actual ratings are increasing over time, the predicted ratings are relatively constant. Because the model is predicting scores consistent with the last few years of numerical ratings, the difference between actual and predicted values shrinks over time, but this is nevertheless consistent with our second hypothesis. Numerical feedback and pure qualitative evaluations diverge over time.

5.2 Greater precision in expert reviews

Furthermore, we find much more consistent ratings and reviews for experts than non-experts as expected. RMSE for those reviews are lower and expert reviews are better suited to predict associated scores, therefore. A potential complication, however, is length of a review. Expertise is not always perceived to be associated with brevity. It may be the

[13]The figure starts in February of 2014. We delete previous months where the number of reviews was less than 2,000.

Table 3. Average prediction error is lower for expert reviewers

	Short Reviews		Long reviews	
	Expert	Non-expert	Expert	Non-expert
Has website	0.595	0.630	0.501	0.535
Above 80th %tile comments	0.577	0.652	0.495	0.601
Above 80th %tile followers	0.591	0.653	0.500	0.597

Note: We obtain predictions using a weighted ensemble on features extracted via Doc2Vec. Experts have 5 or more followers, the user-level 80th percentile. Long comments have above the median word length, independent of user expertise. These statistics come from only the test set.

case that expert reviewers are simply writing longer reviews, which therefore contain more information that we can use to predict the numeric scores. For example, reviewers with a linked website on their Vivino profiles have an average review length of 162 characters; those without average only 97 characters.

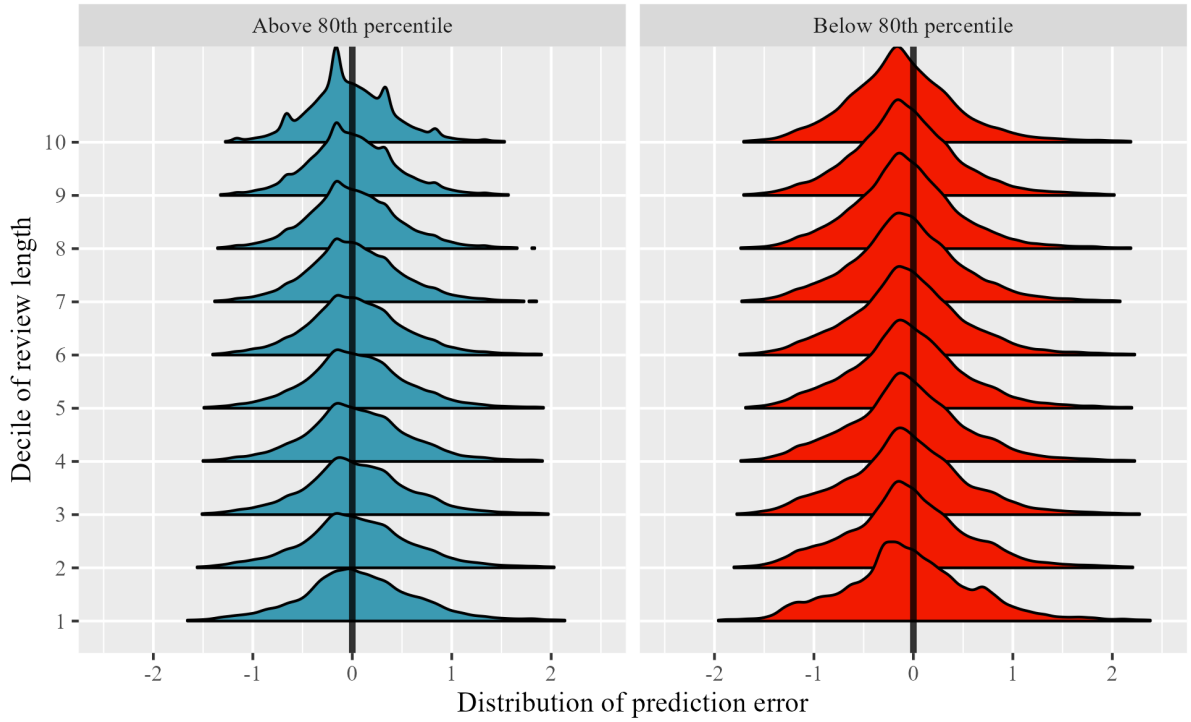
Table 2 shows the RMSE broken down by subgroup. Each entry shows the average distance between the true value and the predicted value. Expert reviews have consistently lower RMSEs than non-expert reviews—they contain more information. Moreover, long reviews are not necessarily more informative than short reviews. If anything, they produce noisier estimates.

Figure 2 shows the distribution of error in these estimates for each decile of review length, broken down by whether or not the reviewer is an expert. This figure confirms the intuitions provided by table 2: the length of the review does not have a systematic effect on the accuracy of our predictions.

5.3 Sentiment based evidence

Our sentiment analysis further supports that review inflation afflicts both expert and non-expert reviews. Figure 5 shows how overall sentiment remains relatively stationary in both groups. While we do not display the quantitative ratings in this figure due to scaling issues, the ratings rise over time (cf. Figure 3). This figure further confirms that growth in ratings among experts and non-experts is independent of the actual quality of the wine. The sentiment analysis approach captures the extent to which individuals speak positively of the wines. In this case, it illustrates that the upward trend in ratings is not matched by more positive sentiments. People are giving higher scores to wines

Figure 4. Density of prediction error by decile of review length and user follower counts



This figure distinguishes experts by whether they have an above median number of followers. This figure includes only the scores we use in later analysis—those with a written comment. Errors are obtained using the weighted ensemble method.

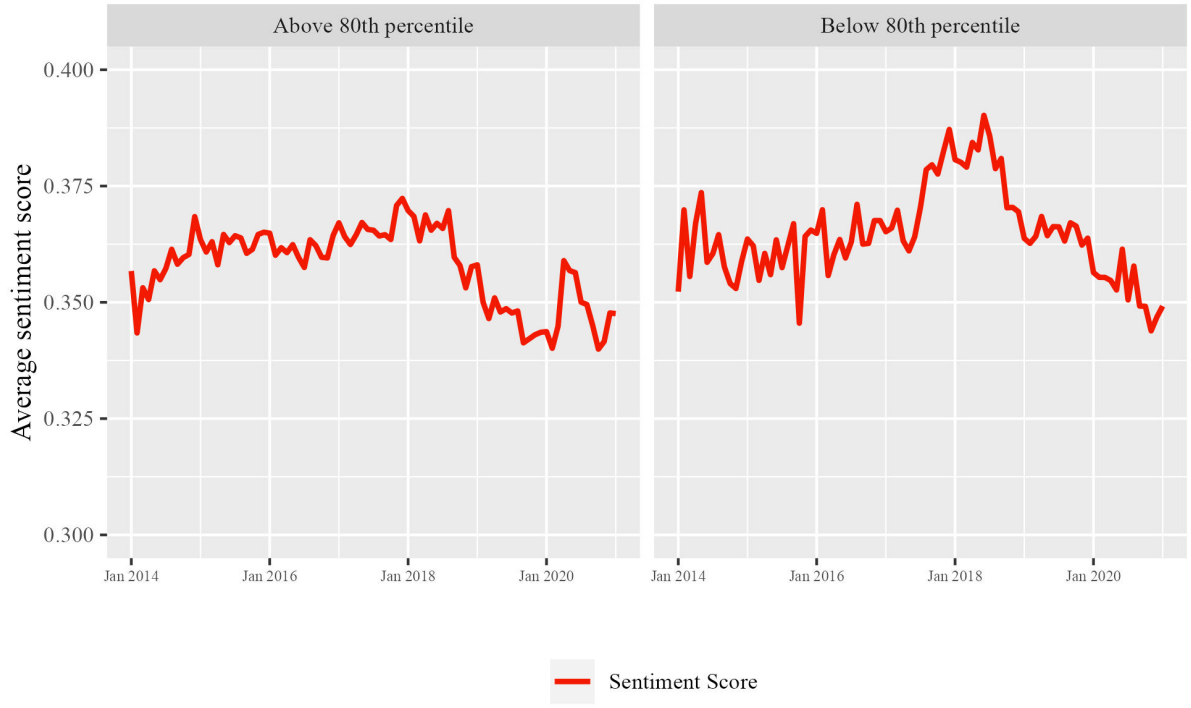
regardless of whether they like the wines more or less.

6 Conclusion

In this paper, we study the existence of reputation inflation and whether certain groups of users are more susceptible than others. We define reputation inflation as an increase in quantitative ratings without an equivalent increase in actual ratings. Specifically, we examine review inflation using numeric and text ratings from Vivino, a large online wine review and market platform. Such ratings are an important part of the digital economy, but they increasingly skew towards the positive end of the spectrum, leaving little information for both producers and consumers.

We first show that ratings do increase over time, from 3.67 out of five in 2014 to 3.86 in 2020. We then use an ensemble machine learning strategy to predict numerical ratings purely from the text-based qualitative reviews. Unlike the numeric ratings, the ratings predicted from the text are steady over time. These results are consistent with review inflation: numeric ratings increased without an increase in underlying quality. These

Figure 5. Average sentiment score by reviewers' expertise



This figure shows the average sentiment score of textual reviews for a given point in time. We differentiate between reviews from users with a below median level of followers, an above median level of followers, and those at the 80th percentile of followers and above.

differences are independent of the length of the text review. We also use a sentiment analysis of these written reviews to rule out a positive change in reviewer sentiment over time.

We also distinguish between so-called expert and non-expert reviewers, using a cut-off of the 80th percentile of follower count. In other words, we consider users with more than 8 followers to be experts. We find that reviews left by experts are more informative, in that our ensemble prediction has a lower RMSE when predicting numerical ratings left by experts. A qualitative review of expert and non-expert reviews suggest that experts use fewer vague words like "smooth" or price-based terminology, which may not hold as much predictive weight. However, we find that expert reviewers are equally susceptible to review inflation.

The prominence of review systems in the online economy makes review inflation an important subject. Truthful reviews benefit consumers by helping them to distinguish between different products and services on an online market. Truthful reviews also benefit service providers or sellers; Uber, for example, rates both riders and drivers. This articles

contributes to a growing literature which quantifies and explains this review inflation. While previous literature has largely treated reviewers as homogeneous, we distinguish between standard users and so-called experts, who we would expect to be less susceptible to review inflation. While both standard users and expert users alike have contributed to review inflation on "the world's largest wine marketplace," the fact that expert reviews more accurately predict quantitative ratings shows the need for a greater understanding of online review systems and review inflation specifically.

References

Archak, Nikolay, Anindya Ghose, and Panagiotis G Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews," *Management science*, 2011, *57* (8), 1485–1509.

Athey, Susan, Juan Camilo Castillo, and Bharat Chandar, "Service quality in the gig Economy: Empirical evidence about driving quality at Uber," *Available at SSRN*, 2019.

Dellarocas, Chrysanthos and Charles A Wood, "The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias," *Management science*, 2008, *54* (3), 460–476.

Engler, Tobias H., Patrick Winter, and Michael Schulz, "Understanding online product ratings: A customer satisfaction model," *Journal of Retailing and Consumer Services*, 2015, *27*, 113–120.

Filippas, Apostolos, John Joseph Horton, and Joseph Golden, "Reputation inflation," *Marketing Science*, 2022, *41* (4), 733–745.

Floh, Arne, Monika Koller, and Alexander Zauner, "Taking a deeper look at online reviews: The asymmetric effect of valence intensity on shopping behaviour," *Journal of Marketing Management*, 2013, *29* (5-6), 646–670.

Gan, Qiwei, Bo H Ferns, Yang Yu, and Lei Jin, "A text mining and multidimensional sentiment analysis of online restaurant reviews," *Journal of Quality Assurance in Hospitality & Tourism*, 2017, *18* (4), 465–492.

Glazer, Jacob, Helios Herrera, and Motty Perry, "Fake reviews," *The Economic Journal*, 2021, *131* (636), 1772–1787.

He, Sherry, Brett Hollenbeck, and Davide Proserpio, "The market for fake reviews," *Marketing Science*, 2022.

Hu, Nan, Paul A. Pavlou, and Jennifer Zhang, "Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication," in "Proceedings of the 7th ACM Conference on Electronic Commerce" EC '06 Association for Computing Machinery New York, NY, USA 2006, p. 324–330.

Jabr, Wael and Zhiqiang Zheng, "Know yourself and know your enemy," *Mis Quarterly*, 2014, *38* (3), 635–A10.

Katumullage, Duwani, Chenyu Yang, Jackson Barth, and Jing Cao, "Using Neural Network Models for Wine Review Classification," *Journal of Wine Economics*, 2022, *17* (1), 27–41.

Kim, Rae Yule, "Using Online Reviews for Customer Sentiment Analysis," *IEEE Engineering Management Review*, 2021, *49* (4), 162–168.

Klimmek, Martin, "On the Information Content of Wine Notes: Some New Algorithms?," *Journal of Wine Economics*, 2013, *8* (3), 318–334.

Le, Quoc V. and Tomas Mikolov, "Distributed Representations of Sentences and Documents," 2014.

Li, Mengxiang, Liqiang Huang, Chuan-Hoo Tan, and Kwok-Kee Wei, "Helpfulness of online product reviews as seen by consumers: Source and content features," *International Journal of Electronic Commerce*, 2013, *17* (4), 101–136.

Li, Xiaolin, Chaojiang Wu, and Feng Mai, "The effect of online reviews on product sales: A joint sentiment-topic analysis," *Information & Management*, 2019, *56* (2), 172–184.

Li, Xinxin and Lorin M Hitt, "Price effects in online product reviews: An analytical model and empirical analysis," *MIS quarterly*, 2010, pp. 809–831.

Luca, Michael, "Reviews, reputation, and revenue: The case of Yelp. com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, 2016, (12-016).

_ **and Georgios Zervas**, "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management Science*, 2016, *62* (12), 3412–3427.

Ludwig, Stephan, Ko De Ruyter, Mike Friedman, Elisabeth C Brüggen, Martin Wetzels, and Gerard Pfann, "More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates," *Journal of Marketing*, 2013, *77* (1), 87–103.

McCannon, Bryan C., "Wine Descriptions Provide Information: A Text Analysis," *Journal of Wine Economics*, 2020, *15* (1), 71–94.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estima-

tion of Word Representations in Vector Space," 2013.

Moe, Wendy W and Michael Trusov, "The value of social dynamics in online product ratings forums," *Journal of Marketing Research*, 2011, *48* (3), 444–456.

Nosko, Chris and Steven Tadelis, "The limits of reputation in platform markets: An empirical analysis and field experiment," Technical Report, National Bureau of Economic Research 2015.

Ordenes, Francisco Villarroel, Stephan Ludwig, Ko De Ruyter, Dhruv Grewal, and Martin Wetzels, "Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media," *Journal of Consumer Research*, 2017, *43* (6), 875–894.

Tadelis, Steven, "Reputation and feedback systems in online platform markets," *Annual Review of Economics*, 2016, *8*, 321–340.

Yang, Chenyu, Jackson Barth, Duwani Katumullage, and Jing Cao, "Wine Review Descriptors as Quality Predictors: Evidence from Language Processing Techniques," *Journal of Wine Economics*, 2022, *17* (1), 64–80.

Yin, Dezhi, Sabyasachi Mitra, and Han Zhang, "Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth," *Information Systems Research*, 2016, *27* (1), 131–144.

Zervas, Georgios, Davide Proserpio, and John W Byers, "A first look at online reputation on Airbnb, where every stay is above average," *Marketing Letters*, 2021, *32* (1), 1–16.

Zhu, Feng and Qihong Liu, "Competing with complementors: An empirical look at Amazon. com," *Strategic management journal*, 2018, *39* (10), 2618–2642.

Table A1. Average prediction error is lower for expert reviewers

	Short Reviews		Long reviews	
	Non-expert	Expert	Non-expert	Expert
Has website	0.587	0.610	0.499	0.532
Above 80th %tile comments	0.602	0.609	0.482	0.545
Above 80th %tile followers	0.573	0.612	0.477	0.546

Note: We obtain predictions using a weighted ensemble on features extracted via Doc2Vec. Expert is defined as above the 80th percentile of comments/followers.

A Online Appendix

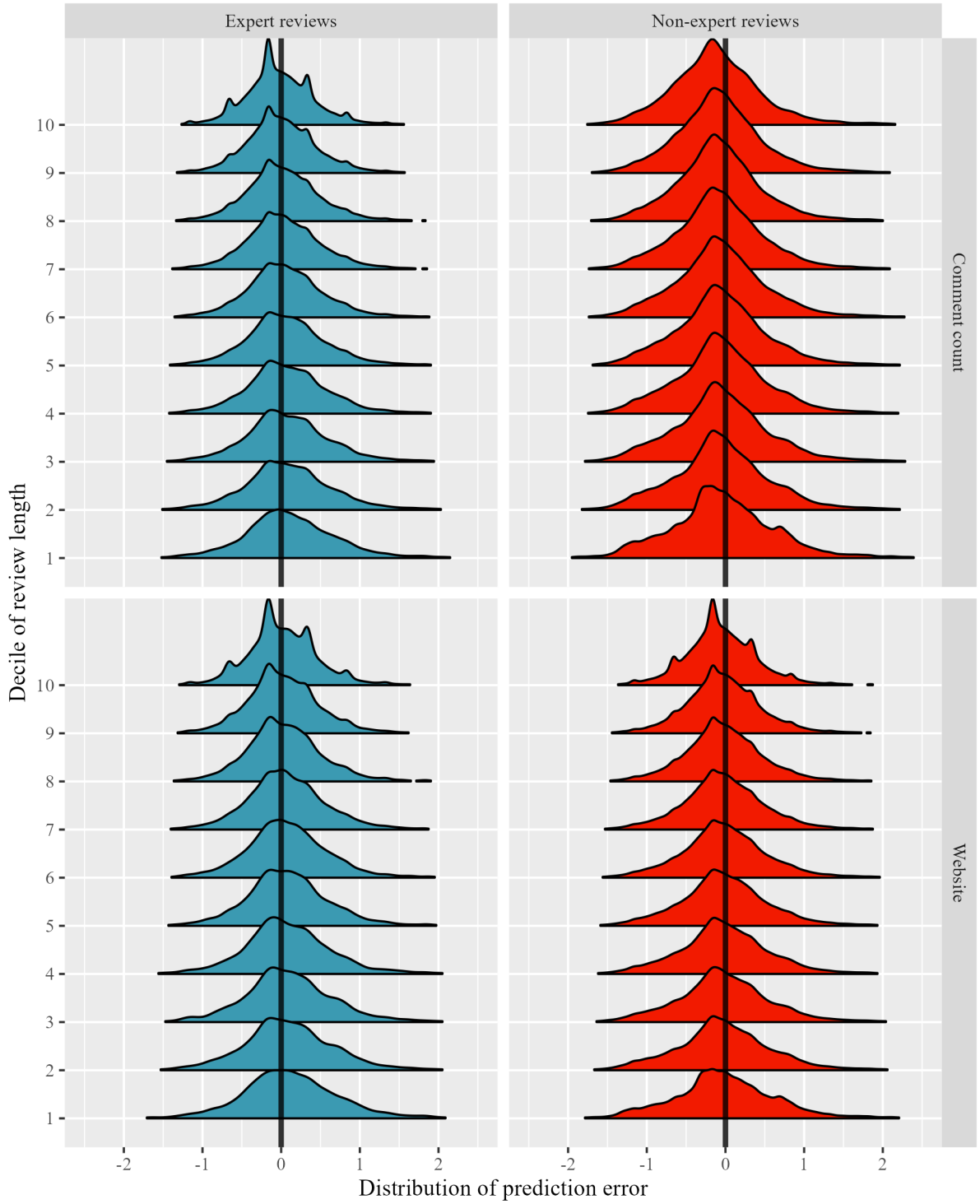
B Alternative metric for expertise

In this article, we decline 'experts' as having an above-median number of number of followers on Vivino. However, we also considered two alternative definitions of reviewer expertise: whether the average number of comments per post is above the 80th percentile, and whether or not a user links a website to their Vivino profile.

Figure A1. Actual and predicted wine ratings by user expertise

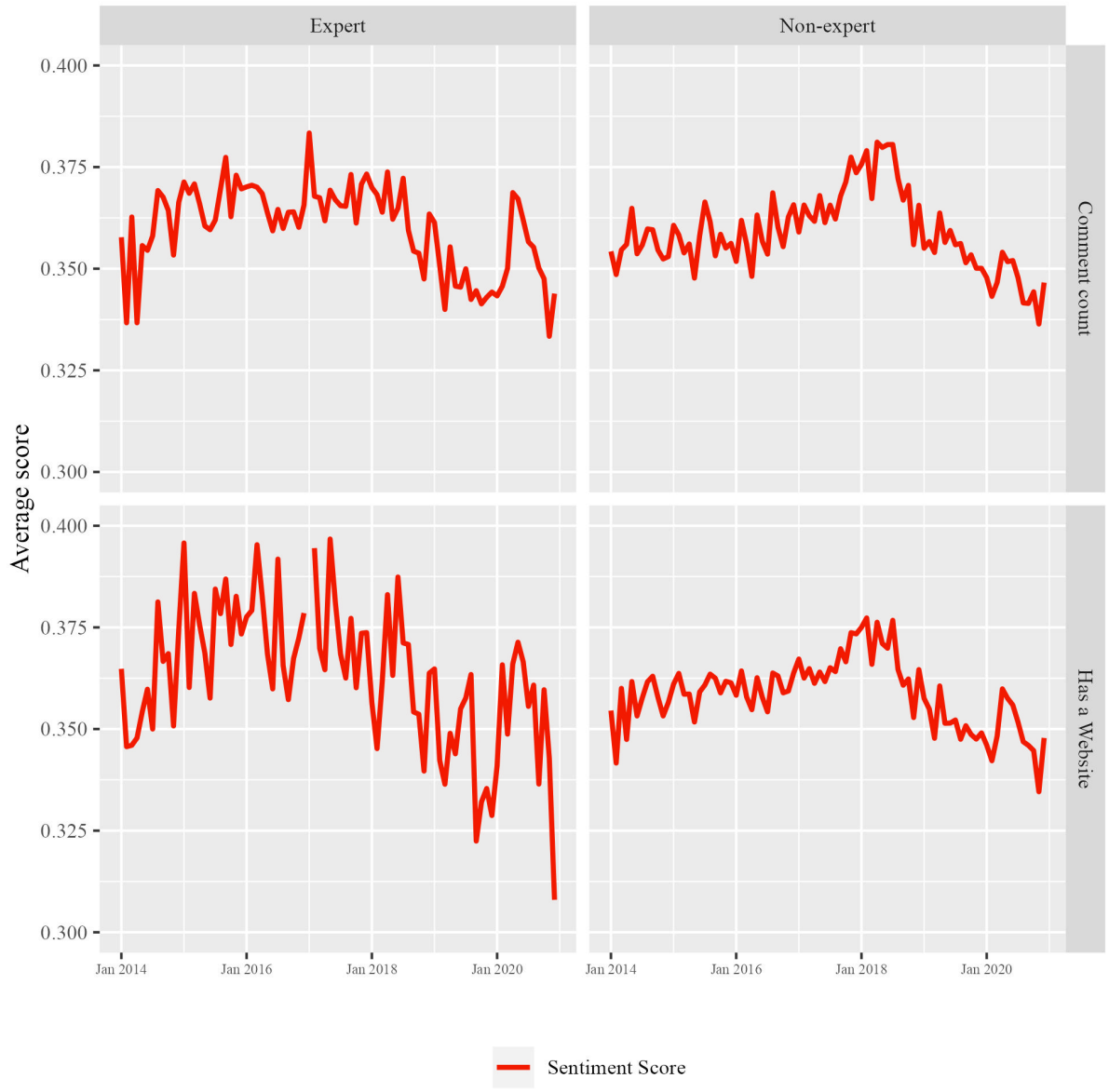
This figure distinguishes experts by whether they have above the 80th percentile of average per-review comments and whether they link a website to their Vivino profile. This figure includes only the scores we use in later analysis—those with a written comment. Errors are obtained using the weighted ensemble method.

Figure A2. Density of prediction error by decile of review length



This figure distinguishes experts by whether they have above the 80th percentile of average per-review comments and whether they link a website to their Vivino profile. This figure includes only the scores we use in later analysis—those with a written comment. Errors are obtained using the weighted ensemble method.

Figure A3. Average sentiment score by user expertise



This figure distinguishes experts by whether they have above the 80th percentile of average per-review comments and whether they link a website to their Vivino profile.This figure includes only the scores we use in later analysis—those with a written comment. Errors are obtained using the weighted ensemble method.