Rich or *Pour* Taste? Predicting the Price of Wines

Atharva Mehendale
atharva@berkeley.edu

Jim Zhu
jyz6@berkeley.edu

Will Dudek
wdudek@berkeley.edu

August 6, 2023

ABSTRACT: Our team explored ways to use natural language processing techniques on reviews of various wines by sommeliers in order to predict the price of a bottle of those wines. We implemented a baseline model using a support vector machine (SVM) architecture as it was the main model in previous works, but was used for easier objectives (i.e. predicting the color, grape type, or quality). We then used our knowledge of BERT-based models to segment wines into four equal price buckets based on the data's price percentiles, opting to train both a vanilla BERT implementation as well as its more advanced cousin RoBERTa. We ran a variety of experiments given the various features provided in the dataset, injecting variables such as the grape variety and country, into the description to boost accuracy and varying the number of price buckets from four to 2. Ultimately, our models showed a near-equal comparative difference between the baseline SVM model and the BERT-based models as in the papers and approached 60% in accuracy against an objective that had nearly no correlation with the input.

1 Introduction

Due to the subjective nature of individual palates and preferences, as well as the challenge of conveying sensory experiences through words, there are longstanding debates over how informative wine reviews truly are.[1] Some go as far as arguing that smells and flavors are impossible to put into words on a consistent basis, and question whether wine reviews convey any helpful information at all.[2]

Indeed, these reviews by sommeliers are meant to have the ultimate say on the taste and quality of wine, and we felt that if this was the case, then there should be a way to connect the content of such reviews to the overall price.

Essentially, our goal is to find out whether it is strictly necessary that a wine be expensive to have a good taste to a wine connoisseur.

This is challenging because our model has to associate particular words that could describe anything from the color and taste of it to the added flavors and the grape variety that the wine is made from. In addition, this is not an easy task for the untrained human palette: guessing the price of a wine just based on the taste is something that is normally quite a difficult task.

2 Background

Although it is not a popular area of research, the task of extracting meaning from wine reviews has been the subject of some study in the field of natural language processing. One popular paper uses machine learning techniques to explore how 'expert' wine reviews convey information about wine quality (among other characteristics)[3]. Building on previous research into relationships between wine reviews and quality, the authors focus on three classification tasks: color prediction, grape type prediction and country prediction, using support vector machine models from the LIBSVM toolkit by Chang and Lin (2011).

The results are promising, suggesting that descriptions alone convey information about color, grape variety and country of origin. Additionally the authors use regression models to explore the relationship between wine price and average word length, finding a slight correlation between average word length and price (beta estimate 2.61); however, they do not discuss the relationship between the reviews themselves and price in the context of a predictive task. This is of interest to us in particular due to ongoing debates about the ability of sommeliers to predict wine prices, and we aimed to expand on this research by focusing on predictive tasks

[1] Gawel (1997) and Suarez Toste (2007)

[2] Sperber et al., 1975

[3] Lefever et al., LREC 2018

related to price, reviews and other features like country of origin.

A different paper from just last year gave us more inspiration on how to implement this project. Duwani Katumullage and his team took previous work on training SVMs for this and applied bidirection long short-term memory (LSTM), a convolutional neural network (CNN), and then BERT to the problem of interpreting wine reviews through a neural network lens rather than a statistical one. Their approach, specifically that of ordinal classification by way of creating bins from ratings for the model to more simplify categorize wines, helped us think along the path of predicting numerical values based on text alone. Eventually, we confirmed that price was what we wanted to predict as it could be just as if not more practical as quality for users: being able to input an expert's review of a wine and get a rough estimate of the price range could make the decision of buying wine perhaps even simpler than knowing the quality by itself.

From our research using these papers, we thought of fine-tuning one pre-trained BERT model and one pre-trained RoBERTa model for two reasons. First, they are significantly more powerful than previous modeling approaches that used, for example, SVMs. Second, other approaches use it to classify wines into *quality* buckets, while in our approach, we used it to classify wines into *price* buckets. In addition, we are now more familiar with BERT-based models given our coursework in this class.

3 Methodology

3.1 Data and Preprocessing

The dataset we used is from Kaggle, which provides 150,929 examples of sommelier review transcripts that include the wine's country, vineyard, rating, price, province, region, and variety. However, after doing some exploratory data analysis, we noticed that there were a few problems with it: a good chunk of the rows had the exact same description, but had different attributes. For example, a wine review with the words *"A little bit funky and unsettled when you pop the screwcap, but soon it finds its floral, blueberry base. Remains superficial and sweet in the mouth, with candied flavors, vanilla and mild oak. Highly regular; could use more concentration and density."* came up 6 times in our data, but in different rows.

Another problem was that a large amount of rows had NaN values in the price category, which meant that those rows were irrelevant for our experiments. After we had dropped both duplicates and rows with NaN price values, we ended up with 89,109 rows. While this was a big step down from the number of rows we had before this, our dataset was in line with the rest of the papers we read - Lefever's team had a dataset of 76,410 records. However, due to the constraints of Google Colab's usage limits, our team was unable to run more than 20,000 records at a time without the runtime disconnecting. As these charts demonstrate, the performance of our models doesn't increase much with new data, implying that this is close to the overall accuracy our models would achieve with those many records.

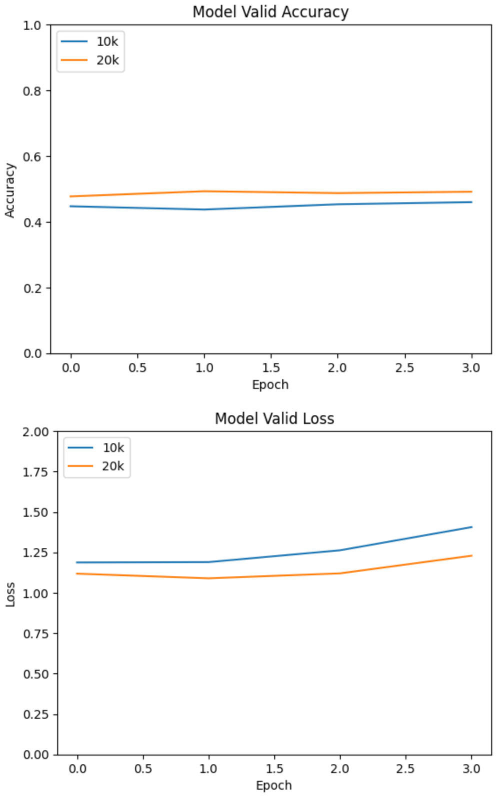

Figure 1: These plots depict the small variance between the model's accuracy with 10,000 records versus 20,000 and the flatness of the learning over epochs as well as the increasing of loss over epochs, demonstrating that our model's accuracy for this type of classification would bottom out around the same numbers.

Additionally, we noticed that many wine descriptions contained information on quality (*"delicious. 80-82 points"*) or even the price itself (*"$10 for this very drinkable Cab? That's crazy."*) As a result, we chose to remove all numbers from wine reviews in order to discourage models from leaning on those features for predictive power and limiting our ability to infer about the relationship between language and price. With this, could truly test the power of our models without any interference from numerical data already in the input.

3.2 Using Bins and Features for Experiments

Since price is inherently variable and not an objective indicator of value, we chose to categorize price into distinct 'bins'. This approach allows us to de-emphasize minor fluctuations in price, which could be influenced by diverse markets, time periods and other factors and instead focus on more significant (and perhaps more meaningful) differences in price. Specifically, we trained models on two separate categorical price features: 4-class ('low', 'low-medium', 'medium-high', 'high')' and 2-class ('low', 'high') as shown in previous papers[4].

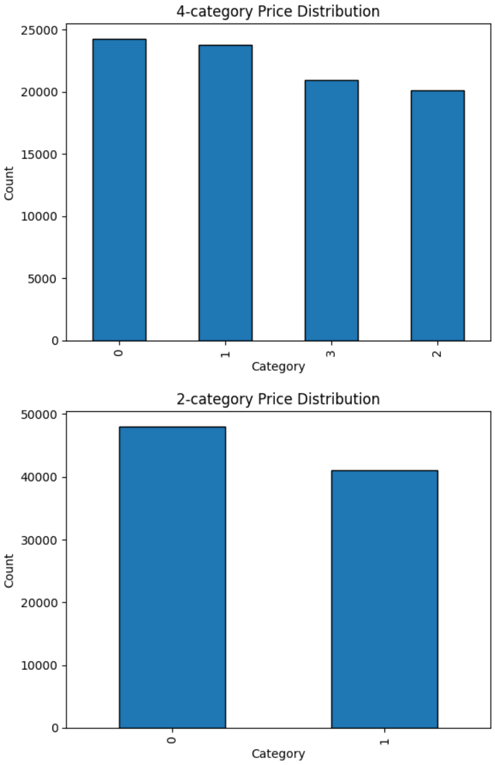

Figure 2: These histograms show the distribution of the still imbalanced, but more equally-sized bins for both types of experiments.

The additional features in the dataset allowed us to think how the model could learn to associate particular varieties, regions, and wine rating with the sentiment of the sommelier's words to better understand how to predict those bins.

As such, we created four separate experiments to isolate the effects that they would have on each type of bin: one to train the model on the description and a sentence that included the feature (i.e. "The country for this wine is" + country) for four classes, another for this and two classes, and vice versa for just the description by itself.

3.3 Baseline Model: SVM

Since we are predicting one of four price categories with roughly equal class distribution, our most naive baseline target is guessing class 1 "low" every time (25% accuracy, in other words).

For a more nuanced baseline model, we used a Support Vector Machine (SVM) classifier similar to Lefever et al., along with a Term Frequency-Inverse Document Frequency (TF-DIF) vectorization technique to convert textual descriptions into numerical features. The TF-IDF approach captures the importance of words within the descriptions while considering their frequency across the entire corpus as a maximum of 5,000 features are selected to represent the text. We then used a linear kernel for the SVM and trained on TF-IDF-transformed training data and the corresponding categorical price labels.

3.4 BERT Model

Next, we explored BERT-based models as seen in other papers to see if we can achieve better performance predicting each categorical price variable. We chose BERT due to our familiarity with it and considering its ability to ingest wide and raw amounts of text data while making sense out of them, which would prove useful when looking at such a variety of reviews across thousands of them. We experimented heavily with the architecture and came up with one that ensured that the model would be fully trainable and had two hidden layers such that the second had half the ones in the first in order to decrease model complexity as it apporaches the output. We had to change the maximum length of the tokens between experiments that did and did not include the added features to account for the increase of tokenized data. In order to do so, we looked at the 80th and 90th percentiles of the word length of the descriptions with and without features and adjusted the "max length" parameter accordingly.

3.5 RoBERTa Model

Having seen the power of BERT-based models first-hand, we turned to a more advanced cousin of BERT to see if it could improve upon our results. Although extremely similar, RoBERTa builds upon BERT by training on a larger dataset of text and has slight editions to its masking pattern and overall training procedure. Prior studies and comparisons have shown that RoBERTa performs very similarly to BERT-based models, but with a marginal improvement in most performance categories. Our RoBERTa model used almost

[4] Katumullage et al., Journal of Wine Economics 2022

identical parameters to the BERT model. Some slight differences include a longer "max length" parameter and different hidden layer sizes.

4 Results and Discussion

4.1 Baseline Model: SVM

Our initial SVM predicted the four-category price bins with an overall accuracy of 0.47 and a weighted average F1 of 0.47. Not surprisingly, the model performed better on the lowest and highest price bins (F1s of 0.56 and 0.57 respectively) compared to the ‘low-medium’ and ‘medium-high’ classes (F1s of 0.40 and 0.33 respectively). For the two-category price prediction task the model had an overall accuracy of 0.81, with nearly identical F1 scores within each class.

4.2 BERT Model

The vanilla BERT model we implemented predicted the four-category price bins with an accuracy of 0.49 without features added and 0.60 with features added, having a weighted average F1 of 0.47 without features added and 0.6 with features added. Just as above, the model performed better on the lowest and highest price bins (F1s of 0.73 and 0.72 respectively) compared to the ‘low-medium’ and ‘medium-high’ classes (F1s of 0.49 and 0.40 respectively). For the two-category price prediction task the model also had an overall accuracy of 0.81.

4.3 RoBERTa Model

Comparing performances, we see that RoBERTa generally performed nearly on par with SVM and similar to the BERT model, reinforcing the comparisons performed in other BERT versus RoBERTa studies. Interestingly, we see that for each model, the lowest and highest price bins were classified most accurately. For the RoBERTa model, that difference is slightly magnified (F1s of 0.59 for both bins) compared to the other models. Although we generally expect RoBERTa models to perform slightly better than BERT base models, the hyperparameters between the RoBERTa and BERT model differed slightly as a result of division of labor. We also would assume that under a larger training size, RoBERTa would perform slightly better given its architecture and slightly more effective training procedures.

4.4 BERT and SVM Accuracy Analysis

Although BERT-based models performed better than our naive baseline, the observed limited improvements of BERT over SVM models deserves careful examination. However, we are not the only ones experiencing this phenomenon: the Katumullage paper explicitly mentions that their BERT model had less than a 2% difference in accuracy compared to the SVM model they saw in their research. Based on our methodology and findings, we suggest several factors that may have contributed to this limited improvement.

First, we must acknowledge the influence of the size of data on the performance of BERT. BERT’s success often depends on having vast amounts of data for robust feature learning. With sampled datasets ranging from 10,000 - 20,000 wine reviews, BERT’s many parameters and sophisticated architecture may encounter challenges with overfitting that hinder its ability to achieve substantial improvements over SVM. Second, the relatively concise nature of wine reviews introduces the concept of text complexity. BERT often succeeds in deciphering intricate linguistic nuances, making it particularly advantageous for tasks involving extensive contextual understanding. In the case of wine reviews, where the language can be straightforward and reviews are succinct, BERT might not have a significant advantage over traditional models like SVM. Finally, the aspect of domain specificity must be considered: BERT’s pretraining on diverse internet text may not align with the unique terminologies present in the specialized domain of wine reviews. Although less intricate in its language comprehension, SVM might still capture the essence of sentiment within these reviews due to its domain-agnostic nature. These insights point to future opportunities to fine-tune BERT-based models on a larger amount of wine review data in order to fully exploit the improvements BERT offers over SVM in many NLP tasks.

5 Conclusion and Future Work

These results suggest that descriptions of wine alone *do* convey some meaningful information about price, challenging previous assumptions to the contrary and supporting previous research primarily focused on predicting color, variety, and country of origin.

Due to the nature of the dataset, we had access to quite a variety of variables, which made us think about how we could further analyze this data. One experiment that arose from this was to understand whether the country of origin for a wine impacted the way the review was worded and could lead to improving price predictions. We chose the BERT model for this task as we wanted to see if a neural network could tell the difference between countries’ wine descriptions rather than a statistical difference by itself. We looked at the top three countries that had wines reviewed in the dataset - the United States, Italy, and France respectively - and created three new datasets from them. Individual BERT models were trained on each dataset, from which we saw that Italy and France improved by 5 and 4% respectively. While they may not seem like much, we now know that we could group this dataset by various variables and potentially boost the accuracy of a particular domain.

Another experiment that came up while defining the problem statement for this project was to compare and contrast ordinal classification, which uses categories that have an ordered rank (i.e. a rating system or binning), with a regression problem context. In particular, since price is a continuous variable, we wondered whether it would make more sense to view this as a regression instead, to which we put BERT to use again with a different loss function (mean squared error or MSE) and metric (root mean squared error or RMSE). While we did see good results from this experiment, one goal of ours for this project was to ensure interpretability and usability for the reader - even if our model did have a low RMSE, the person implementing this project may not see the value of this outside academic pursuits. Due to this, we stuck to completing ordinal classification experiments for this project, but would like to investigate this further in the future.

6 References

Chang, C.-C. and Lin, C.-J. (2011). *LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27:1–27:27. ISSN: 2157-6904.

Els Lefever, Iris Hendrickx, Ilja Croijmans, Antal van den Bosch, and Asifa Majid. 2018. *Discovering the Language of Wine Reviews: A Text Mining Account.* In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Gawel, R. (1997). *The use of language by trained and untrained experienced wine tasters.* Journal of Sensory Studies, 12(4):267–284.

Katumullage, D., Yang, C., Barth, J., and Cao, J. (2022). *Using Neural Network Models for Wine Review Classification.* Journal of Wine Economics, 17(1), 27-41. doi:10.1017/jwe.2022.2

Sperber, D. (1975). *Rethinking symbolism.* Number 11. CUP Archive.

7 Appendix

Table 1: Results for our four types of experiments for three models. We denote the experiment, model, accuracy, precision, recall, and weighted average F1 score for each run. Note the slight improvements in BERT and RoBERTa over the SVM model.

Experiment	**Model**	**A**	**P**	**R**	**WF1**
2-class, description only	SVM	0.81	N/A	N/A	0.81
2-class, description only	BERT	0.74	0.74	0.74	0.73
2-class, description only	RoBERTa	0.73	0.74	0.73	0.73
2-class, description w/ features	SVM	0.8	N/A	N/A	0.8
2-class, description w/ features	BERT	0.78	0.80	0.78	0.78
2-class, description w/ features	RoBERTa	0.79	0.81	0.79	0.79
4-class, description only	SVM	0.47	N/A	N/A	0.47
4-class, description only	BERT	0.47	0.5	0.47	0.48
4-class, description only	RoBERTa	0.47	0.48	0.47	0.46
4-class, description w/ features	SVM	0.55	N/A	N/A	0.55
4-class, description w/ features	BERT	0.6	0.6	0.6	0.59
4-class, description w/ features	RoBERTa	0.58	0.57	0.58	0.57