Predicting Wine Ratings via NLTK and Naive Bayes Classifier: Data Challenge 8

Jacob Farner
Computer Science, COMP331
Occidental College
Los Angeles, California
jfarner@oxy.edu

Leopold Ringmayr
Computer Science, COMP331
Occidental College
Los Angeles, California
lringmayr@oxy.edu

***Index Terms*—Kaggle, Naive Bayes, NLTK, Sentiment Analysis, Wine Ratings.**

I. Introduction

What make the perfect wine? Perhaps one can rely on specific wineries or regions, or perhaps a short description of the wine is enough to predict the quality of the wine. This work seeks to leverage tools such as Natural Language Processing and a Naive Bayes classifier to help shed light on whether the quality of wine can be established based on a short description from a wine taster. By building a machine learning model, one can examine whether the quality of the wine, measured by a point score, is correlated with certain attributes and features–and ultimately whether the wine quality is predictable using pre-existing data. Specifically, this work deals with descriptions of different wines based on wine tasters' opinions. Based on these descriptions, and with help of other notable features (wine price, for example, the goal is to extract features from the text to predict how highly the wine is rated in points. For broader purposes, this analysis could aid in wine recommendation pages or simply in ranking different wines for consumers to choose from. From an economic perspective, this work could aid in product recommendations for consumers and provide new insights into sales analysis for the wine industry.

II. Related Work

This problem of applying computational and machine learning techniques to wine ratings has been explored extensively in academic research. The existing work related to this problem suggests both Natural Language Processing approaches as well as machine learning solutions that do not rely directly on NLP. Hendrickx et al. 2016, for example, examined a very similar problem to this research, as they looked at "de- scriptions produced by a select group of people who have considerable experience naming smells and flavors, i.e., sommeliers and wine journalists." (Hendrickx et al. 2016. This approach places emphasis on the accounts by wine tasters and experts and leverages sentiment analysis to help assess wine quality. Moreover, this particula research had a highly nuanced analysis of different wine features, such as grape variety. Other approaches place emphasis on other wine features, such as year, price, or place of winery. These approaces also make use of different alogorithms and models. For example, Yeo et al. 2015, utilizes Gaussian processes to model the wine prediction. Yeo et al. 2015 rely on time series for their prediction model, and "treat the wine returns in our dataset as a stochastic process" (Yeo et al. 2015). Ultimately, this problem of using machine learning and NLP for qine analysis is an area with comprehensive and active research.

III. Methodology

For this classification problem, the main steps are the reading in and reformatting of the initial .csv file, feature selection and data cleaning, and the use of Naive Bayes to build the classifer.

A. Project Data and Data Preprocessing

The project data stems from the Kaggle Dataset "Wine Reviews", posted by zackthoutt. The data is in a .csv format, and is saved locally within a PyCharm project. The project data consists of a total of 130,000 entries for different wines, providing information on the wine description, country, winery, title, and rating, among other features for each wine. This project will use the wine features to predict the point rating of the wine, thus treating the "points" column of the dataset as the label.

For building the classifier using Python's NLTK, the reviews are read in and each file is assigned the respective classification label, stored in a list of all reviews, although the label column was created by adjusting and normalizing the "points" from the original dataset. Not only are the words in the wine description used as features for the model, but this work also takes takes prices into account. This list is ultimately shuffled for the separation into training and test data.

B. Data Cleaning and Feature Selection

To allow for simple manipulation of the wine data, Python's Pandas library was utilized. Converting the .csv file into a DataFrame format allowed for isolation of individual columns, which was useful for feature selection. Based on domain knowledge and the given dataset, the price column was considered to be valuable, as there was a general trend in the data for more expensive wines to rank higher in points.

As a pre-processing step, prices from the original data were divided into brackets and assigned values between 0 and 4 for the different price brackets. In addition, the label column (points) was modified to show points between 0 and 5 to allow for simpler classification. Here too, certain ranges were established to assign the modified point number (e.g. entries between 84 and 87 points were converted to 2); for example, 5 points would be associated with a very high ranking wine while 1 point would correspond to a wine with a lower rating. The objective here was to create 6 possible labels and allow for simpler training. This new point system (0-5 points) is an adaptation of Olivier Goutay's work on wine ratings from Kaggle.

As a further step, we created a lexicon od words that were strongly linked to the highest ranking wines, as we found these words to be particularly important to use as features. These words were recorded as "praise" words, and included words such as "beautiful". This "praise lexicon" was used to create featuremaps as well as the other feature detection that relied on word frequencies. The converted price was also included as a feature for this model.

C. Classification

Building the model entailed the use of 10-fold cross-validation on the 2000-file dataset. The cross validation allows for a more holistic assessment of the model's performance and accuracy. This model relies on the NLTK Naive Bayes classifier. In addition, the ten most informative features were shown for each fold of the cross validation to give insight into which tokens were most impactful on the classification.

After the ten folds, the average of each of the performance statistics was taken (i.e. average accuracy over ten folds, average precision for the positive label, etc.). The final output consists of total accuracy along with precision and recall and F-score for each of the two labels.

In a separate file (secondmodel.py) a logistic regression model was used on the same dataset a a means for comparison. This model used the same steps for the feature selection and the same mode of tokenization, however, the accuracy of this model was on a test set after a simple training-test split (i.e. no cross-validation).

IV. Results and Analysis

The Naive Bayes Classifier had an average accuracy of 60 percent over the 10 folds–with some folds having a classification accuracy of above 63 percent. The most informative features included tokens such as "beauty", "massive", "gorgeous", "impressive". These features stemmed mainly from higher rated wines in comparison to very low rated wines; for example, a wine rated with 5 points compared to a wine rated with 2 points. In addition, a price category of 3 (higher priced wines) was listed as a strong feature for the model.

In the separate file with the logistic regression model on a simple training and test data split, the accuracy was 77 percent as well. Thus, this result is not significantly higher or lower than the Naive Bayes classifier.

V. Threats to Validity

It is unclear whether a different classifier, such as a Logistic Regression approach, would have proven more suitable for this wine rating prediction problem. A comprehensive test using different models could provide a broader overview of how well the chosen features, and especially the wine description, predict the point rating of the wine. In addition, domain knowledge plays an important role with this work as well: With more expertise on which features (certain words in the wine description, or maybe a particular wine taster) have a heavier impact on the quality of the wine, this model could have been even more refined. More domain knowledge is required for a more in-depth analysis. In addition, one could make the argument that the wine descriptions are subjective and linked to a specific wine taster. While this is true, the model takes this subjectivity into account through its large dataset. The large vocabulary captured from the wine descriptions should capture these differences that may exist between certain wine tasters' descriptions.

VI. Conclusion

Our results suggest that this logistic regression approach is overall a suitable method to predict wine rating; however, with certain restrictions. Overall, the model may be suitable for predicting a very high ranking wine from a wine with a particularly low rating, but some of the more nuanced differences between mid-ranking wines may not have been entirely captured in this model. This can be seen in the most informative features: With words such as "superb" or "promises" or "massive" being among the most informative features, it becomes clear that these features mainly distinguish a label 5 wine from a label 2 or 3 wine. This suggests that it may be easier for the model to predict a very highly rated wine. It also interesting that a high price (price class of 3) was also listed as an informative feature, suggesting that price can be a good indicator of wine quality. Perhaps for future work, other wine features should be taken into account to supplement the text descriptions, although wine price have been figured into this model as a feature. This would allow for a more nuanced categorization of unseen wine descriptions.

Acknowledgment

Thank you to Professor Chen, who taught us NLTK fundamentals and the theoretical and practical underpinnings of natural language processing.

References

[1] I. Hendrickx, E. Lefever, I. Croijmans, A. Majid, and A. van den Bosch, "Very quaffable and great fun: Applying NLP to wine reviews," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 2016, pp. 306–312, doi: 10.18653/v1/P16-2050.

[2] Kaggle.com. 2020. Wine Reviews. [online] Available at: ¡https://www.kaggle.com/zynicide/wine-reviewswinemag-data-130k-v2.csv¿ [Accessed 24 April 2020]. www.statista.com/statistics/379112/e-commerce-share-of-retail-sales-in-us/

[3] M. Yeo, T. Fletcher, and J. Shawe-Taylor, “Machine Learning in Fine Wine Price Prediction,” J Wine Econ, vol. 10, no. 2, pp. 151–172, Nov. 2015, doi: 10.1017/jwe.2015.17.