# Wine Classification with Deep Learning

**Kendra Gedney**
M.S. Analytics
Georgetown University
Washington, D.C.
kg729@georgetown.edu

## Abstract

This paper explores text classification methods for predicting a wine's variety from its tasting note, or description. It finds that a linear SVM and a simple deep learning model achieve the highest accuracy and are the most efficient. In a separate but related task, word embeddings are fit on the wine tasting notes using the Word2Vec algorithm. These reveal interesting clusters and relationships of wine-related words.

## Introduction

Wine is a complex product, complete with detailed and descriptive wine tasting notes. The notes are professionally written to provide details to buyers. They tend to be florid and poetic in style. The task of automatically identifying characteristics about wine could be of interest to the large industry surrounding wine. This research focuses on building text classifiers to predict the type of wine based on the tasting note. Additionally, a low dimensional vector space representation of the words used in wine tasting notes is created using the Word2Vec algorithm. These word vectors provide a rich representation of the language that can then be explored.

Previous research (Levefer, 2018) has focused on classifying a wine's region, color, and grape type, based on one of the same datasets used in this research. Their best model was a Support Vector Machine (SVM) variant, and they did not apply any deep learning methods to their classification tasks.

## 1. Datasets

The data used in this research are written wine tasting notes and their corresponding wine variety. Wine tasting notes are written by professional wine tasters which include descriptions of a wine's color, aroma, and taste. For example, the following is a tasting note from the Wine Enthusiast dataset describing a chardonnay:

*"The aromas entice with notes of wet stone, honeysuckle, chamomile and stone fruit. The palate is generous in flavor and feel, showing a fine sense of balance."*

A wine's variety is the type of grape used to make the wine, for example, chardonnay or merlot. In general, wine tasting notes follow a similar format and lexical style. They are generally short in length, containing just a few sentences or fragments.

### 1.1 Wine Enthusiast Dataset

The primary dataset was collected from the Wine Enthusiast Magazine that was made available on Kaggle. The full dataset was subsetted to include the top 28 wines, each of which has at least 1,000 tasting notes, for a total dataset size of 110,000 rows. Blends (Red Blend, White Blend, Sparkling Blend) were excluded as they combine multiple grape varieties. The resulting classes are not balanced, with Chardonnay, Pinot Noir and

Cabernet Sauvignon making up the majority of tasting notes.

### 1.2 Wine Cellar Insider Dataset

A second dataset was scraped from thewinecellarinsider.com. This dataset was subsetted to contain four wines, with 1,200 rows of tasting notes. This dataset is used to test the transferability of models fit on the primary dataset.

### 1.3 Combined Dataset

The combined dataset, from both sources without any filtering, contains 167,000 rows. This full dataset is employed when fitting word embeddings as described in Section 3.2.

### 1.4 Leakage Concerns

Both datasets may include some leakage. Sometimes the writer may directly reference the variety of wine in the tasting note. Additional leakage may come from winery and vineyard names. In subsequent research, named-entities should be removed prior to modeling.

## 2. Methods

This research consists of two experiments. The first is a text classification task, to predict the variety of wine from its tasting note based on the primary dataset. After fitting, the transferability of those models using the secondary dataset was tested.

Secondly, word embeddings were fit using the Word2Vec algorithm.

### 2.1 Text Classification

The primary dataset was randomly split into a training set (80%) and test set (20%). Four models were tested in the text classification task: one linear Support Vector Machine (SVM) and three deep learning models. The deep learning models fit are:

- Long Short Term Memory (LSTM)
- Bidirectional LSTM (Bi-LSTM)
- Simple Pooling

The architecture for Simple Pooling model is based on the model proposed by Joulin, et. al. (2016), which presented a simple, efficient model that achieved high accuracy results. The architecture of the Simple Pooling model used in this research is as follows:

- Embedding layer of size 256
- 1D Global Max Pooling layer
- Dense Layer of size 100
- Softmax output layer

### 2.2 Word Embeddings

Word embeddings are the learned representation of words as low-dimensional (*eg.* 128-dimensions) dense, vectors which capture relationships among the words (Mikolov 2013). The wine word embeddings were fit on the combined dataset of tasting notes using *gensim's* implementation of the Word2Vec algorithm. In particular, the continuous bag of words (CBOW) version of Word2Vec was used with a window size of 5.

## 3. Results

Evaluation of the text classification tasks was done with the test accuracy rate. For the word embeddings, results are displayed through a tSNE visualization.

### 3.1 Classification Results

The Simple Pooling model performed the best, achieving 84% accuracy on the 28-class classification task. However, all models' performance fell into a tight range, between 82%-84% accuracy.

| Model | Test Accuracy |
|---|---|
| SVM | 0.83 |
| LSTM | 0.82 |
| Bi-LSTM | 0.83 |
| Simple Pooling | 0.84 |

Table 1: Results from text classification task.

Training time was much faster for the SVM and the Simple Pooling models, making them the more efficient models as well as the most accurate.

Results varied by the variety of wine. Chardonnay is identified the best, but it is also the majority class, so it is possible there is some bias.

Italian red wines, Sangiovese, Nebbilio, and Barbera also fared well.

| Variety | Test Accuracy |
|---|---|
| Chardonnay | 0.93 |
| Nebbiolo | 0.92 |
| Sangiovese Grosso | 0.91 |
| Pinot Noir | 0.89 |
| Sauvignon Blanc | 0.89 |
| Barbera | 0.88 |
| Riesling | 0.86 |
| Portuguese Red | 0.86 |
| Corvina, Rondinella, Molinara | 0.85 |
| Rosé | 0.85 |
| Zinfandel | 0.84 |
| Shiraz | 0.83 |
| Bordeaux-style Red | 0.83 |
| Sangiovese | 0.83 |
| Cabernet Sauvignon | 0.81 |
| Rhône-style Red | 0.81 |
| Grüner Veltliner | 0.81 |
| Syrah | 0.81 |
| Tempranillo | 0.80 |
| Pinot Grigio | 0.78 |
| Merlot | 0.77 |
| Viognier | 0.76 |
| Pinot Gris | 0.75 |
| Cabernet Franc | 0.75 |
| Port | 0.74 |
| Champagne Blend | 0.74 |
| Bordeaux-style White | 0.73 |
| Malbec | 0.73 |

Table 2: Results by class from best performing model, Simple Pooling model.

### 3.2 Transferability Test Results

The secondary dataset was used to perform a transferability or generalization test on the models fit in Section 2.1.

The tokenizer and weights from the Simple Pooling model were applied to the secondary dataset, scraped from the Wine Cellar Insider website. This test was a four-class classification task for Chardonnay, Pinot Noir, Cabernet Sauvignon and Cabernet Franc.

The results were poor. The Simple Pooling model trained on the new dataset achieved 40% accuracy. A linear SVM trained directly on the new dataset achieved 95% accuracy.

| Model | Accuracy |
|---|---|
| SVM | 0.95 |
| Simple Pooling | 0.40 |

Table 3: Results from transferability test. The tokens and weights from the best performing model were applied to the secondary winecellarinsider.com dataset. A linear SVM was fit on the new dataset too.

The best performing wine was Chardonnay, which was the also the largest class in the primary dataset.

### 3.3 Word Embeddings Results

After fitting the word vectors, results were projected down to two dimensions using the tSNE method (van der Maaten, 2008). The relationships among the wine words can now be both visualized and computed.

The resulting image, as shown in Figure 1, the points represent words. The clustering reveals how similar words are embedded close together. For example, there are clusters of words to describe the foods to pair wine with, events to pair wine with, and colors to describe the wine.

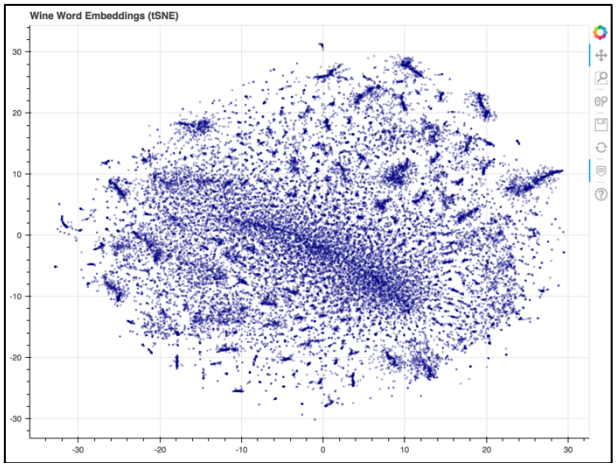


Figure 1: tSNE projection of wine word embeddings. Interactive version of the plot available here: https://github.com/kgedney/anly580-wine-project/blob/master/interactive_tsne.html

Another cluster includes interesting adjectives used to describe wine: thicker, softer, creamier, lighter, rounder, fruitier, deeper.

A particularly dense cluster includes a large amount of fruit-related words, which are commonly used to describe wine: passionfruit, honeysuckle, pear, nectarine, apricot, honeyed.

Vector math is possible with the resulting word vectors using gensim's built in functions that computes distances between the word vectors. For example, the most similar words to "sweet" can be computed, which are displayed below in Table 4.

| Word | Distance |
|---|---|
| coated | 0.53 |
| jammy | 0.51 |
| sweetened | 0.51 |
| ripe | 0.51 |
| overripe | 0.48 |

Table 4: Top five most similar words to "sweet" as determined by the Word2Vec model fit on wine tasting notes.

Additionally, more complicated vector math can be computed. In this example, the vectors for "oak, oaky" are subtracted from the vector for "chardonnay". The closest results are other white wines that are unoaked.

| Word | Distance |
|---|---|
| erbaluce | 0.45 |
| riesling | 0.45 |
| vermentino | 0.43 |
| pigato | 0.42 |
| falanghina | 0.41 |

Table 5: Top five most similar words to "chardonnay" subtracted from "oak" and "oaky" as determined by the Word2Vec model fit on wine tasting notes.

## 4. Discussion

Overall, classification accuracy rates were high. This suggests that the professionally written tasting notes, however obscure they may seem at times, do contain rich information on which classification tasks can be performed with high accuracy.

It also suggests that there are distinguishable differences among wines, either in appearance, aromas, or taste that are present in the language used to describe wine. A red wine is not simply a red wine in the world of professional wine tasting notes.

The tests revealed that simple models worked best for text classification in this domain. There were no improvements from the baseline SVM or the Simple Pooling model by using the complex, high-capacity LSTM models.

However, the poor transferability results suggest that these models may be overfit on the training data or contain too much bias to be generalizable. Having balanced classes may be important in training such models, and further research could prove or disprove the idea.

Leakage is also a concern in these tests. Italian red wines fared well in classification. Italian wine tasting notes may include Italian names of winemakers or vineyards, which the model could be using to help make its classifications. Further research could test the impact of removing named-entities.

Finally, the word embeddings provide a high-level overview of the words used in wine tasting notes. They could be used in writing new tasting

notes and finding new ways to describe taste and aromas. For a wine-drinker, the vector math could be a way to discover new wines with certain attributes for example, a wine similar to a Chardonnay without the oakiness. Further experiments could include the names of the variety of wines to solidify their position among the wine-related words.

## Conclusion

To conclude, this research has shown that wine classification based on its tasting note alone is possible to a high level of accuracy. Wine word embeddings can be used to reveal relationships and spur creativity in crafting new tasting notes.

## References


L.J.P. van der Maaten and G.E. Hinton, "Visualizing High-Dimensional Data Using t-SNE", *Journal of Machine Learning Research*, 2008.

T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *ICLR Workshop Papers,* 2013.

E. Lefever, I. Hendrickx, I. Croijmans, A. Bosch, and A. Majid, "Discovering the Language of Wine Reviews: A Text Mining Account", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, "Bag of Tricks for Efficient Text Classification", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2016.


## Appendix

All code and visualizations are available here:

https://github.com/kgedney/anly580-wine-project