# Deep Learning Wine Grape Variety Recognition Final Report

APS360 Applied Fundamentals of Machine Learning
Instructor: Yani Ioannou

Prepared by:

Group 18

| Chen, Ryan | Deng, Sara | Tian, Yunying (Kyra) | Wang, Yining |
|---|---|---|---|
| 1003912992 | 1003218109 | 1005975768 | 1005728134 |


August 13, 2021

Word Count: 2400/2500

# Contents

# 1 Introduction

This project is about recognizing wine types (only for those fermented with grapes) through their existing taste descriptions via machine learning technology, specifically the LSTM model. The thought of doing such a topic is from the interest in wine by the group members and many other people in the world, especially Canadians [1]. It is the numerous alcohol lovers that make the recognition and appreciation of the different types of wines an interesting topic. This project can help online stores manage and classify their storage at a more rapid speed. It can also help vendors make recommendations on wine type to buyers based on their taste preferences.

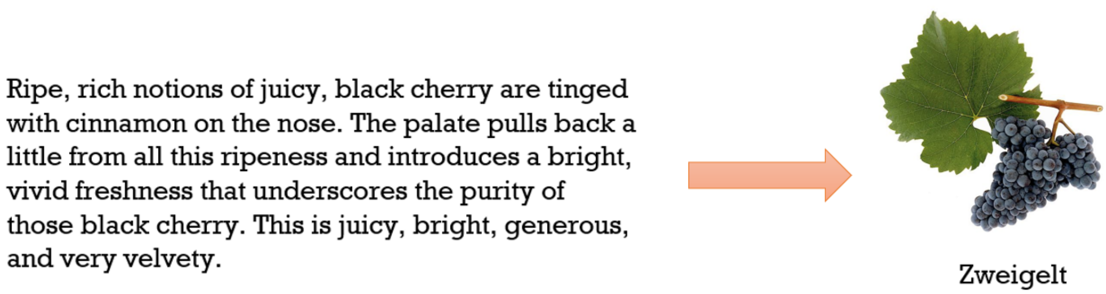



Figure 1: Example input/output pair of the wine grape variety recognition model.

Since the problem requires complex and large datasets with the need for proper data processing, analyzing, training, and testing, machine learning can be a useful tool in solving this task. Besides, no clear predetermined solution to the problem is available, so leveraging machine learning which can execute with no reliance on any pre-established equation can be a good choice. The CNN model, RNN model and LSTM model, which can proceed with sentences as their input data and train on top of pre-trained word vectors for sentence-level recognition and classification, can be alternatives that could be used in dealing with the problem.

| | description | variety |
|---|---|---|
| 0 | Whiffs of freshly dug truffle and rose petals lend seductive perfume to preserved black plum and cherry in this unctuous blend of Grenache (75%), Syrah (20%) and Mourvèdre (5%). It's a spicy, smoky sip w | Rhône-style Red Blend |
| 1 | Aromas of raspberry and boysenberry mingle into savory sage and thyme on this decadent, perfumed blend of Grenache (80%), Syrah (15%) and Mourvèdre (5%). The palate offers rich, ripe layers of plum | Rhône-style Red Blend |
| 2 | The nose of this blend of Grenache, Syrah and Mourvèdre wafts dried strawberry and cherry against a backdrop of herbes de Provence, smoked meat and anise. While plush and rounded, the palate offers c | Rhône-style Red Blend |
| 3 | Cutting mineral notes contrast plush yellow apple and pear in this rich, almost glycerol white. A blend of Grenache (70%), Roussanne (28%), Bourboulenc (1%) and Clairette (1%) sourced entirely from sandy | Rhône-style White Blend |
| 4 | This wine comes from a single vineyard just downhill from the famed windmill of Moulin-à-Vent. It is a rich wine, aged in large barrels for a year that give it a perfumed character. Black fruits and warm tannins | Gamay |
| 5 | Layers of orange cream and dulce de leche lend weight to crisp yellow peach and apple flavors in this blend of six regional white grape varieties. It's a joyously rich, creamy white that lingers on the palate wit | Rhône-style White Blend |
| 0 | Ripe, rich notions of juicy, black cherry are tinged with cinnamon on the nose. The palate pulls back a little from all this ripeness and introduces a bright, vivid freshness that underscores the purity of those bla | Zweigelt |
| 1 | Tender green apple notes mix with hints of wet oakmoss on the subtle nose. The palate then homes in on a supple texture that seems alive with salt and shimmering white pepper amidst savory glints of sage an | Grüner Veltliner |
| 2 | Meaty berry aromas come with notes of herbs, wild flower and oaky wood spice. This reserva from a warm year is solid as a rock and medium-plus in mouthfeel and structure. Ripe black-fruit flavors are back | Tempranillo Blend |
| 3 | This snappy and well-focused wine practically bursts with vivid cranberry and sour-cherry flavors tempered by riper black cherry and hints of earth and mushroom. It is well balanced, offering light tannins anc | Mourvèdre |
| 4 | This blend of Cabernet Franc, Merlot and Cabernet Sauvignon is a distinctive bottling, highly herbal and savory, with umami and sanguine elements. Green herbs, soy sauce and pencil lead are all noticeable | Red Blends |
| 5 | Zest, oakmoss and a hint of crushed fennel seed are all subtle on the wine's nose, but already promise savoriness. The slender but concentrated palate then unites the vivid, pervasive tingle of finely ground wh | Grüner Veltliner |
| 0 | After three years in bottle, this sparkling combines 50% Chardonnay with 40% Pinot Noir and a touch of Pinot Gris. Earthy mushroom and muddled lime, lemon and apple flavors highlight a smooth, structured | Sparkling Blend |
| 1 | Traces of hazelnut and cream lend richness to crisp pineapple and pear in this satiny blend of 60% Roussanne and 40% Bourboulenc. An opulent, fleshy wine aged in large-format oak barrels and stainless st | Rhône-style White Blend |
| 2 | Voluptuous and unabashedly forward, this full-bodied blend of Grenache, Mourvèdre and Syrah drenches the palate with intensely ripe, rich blackberry and boysenberry flavors. It's a creamy, unctuously textu | Rhône-style Red Blend |
| 3 | Delicate swirls of vanilla, caramel and nuts lend richness to crisp yellow apple in this opulent, elegant white. A blend of Grenache Blanc (60%), Roussanne (30%) and Clairette (10%) fermented and matured s | Rhône-style White Blend |
| 4 | This is a plump, hedonistic blend of Grenache, Syrah and Mourvèdre that pulsates with luscious strawberry, raspberry and fig flavors. It's a sultry, supple sip marked with just a hint of leather and edged by fine | Rhône-style Red Blend |
| 5 | Scintillating strawberry and red plum flavors abound in this lip-smacking Grenache-dominant blend sourced primary from sandy soils. Compared with the more heaving, densely packed reds from this appellati | Rhône-style Red Blend |
| 0 | Refined aromas of blue flowers, ripe dark-skinned berry and spice shape the nose. It's elegantly structured, featuring red cherry, orange zest and licorice flavors set against taut, fine-grained tannins. Bright a | Red Blends |
| 1 | Made with organically grown Sangiovese, this opens with aromas of leather, camphor and black-skinned berry. The savory palate offers blackberry, green peppercorn and tobacco alongside fine-grained tann | Sangiovese |
| 2 | A compact nose with blackberry and integrated oak aromas leads to a flush, round palate that's balanced and whole. Blackberry, black plum and cassis flavors are ripe and pleasing, with a touch of raw oak in | Tempranillo |
| 3 | Sourced from Tualatin Estate Vineyard, which dates back to the 1970s, this wine offers plush, palate-filling flavors of raspberries, blueberries and plums. Aged 16 months in 39% new oak, it rolls into a smooth | Pinot Noir |
| 4 | Old-vine Pommard clone vines dating back to the 1970s punch this wine forward with lovely, detailed and ripe cherry fruit. It's notched with streaks of oranges and nectarines, yielding a juicy, immaculate and | Pinot Noir |
| 5 | Notes of fresh yeast blend with hints of cut green pear on the appetizing nose. The palate has that same salty, yeasty tanginess but adds a whole load of white pepper spice. All is held in a tight, textured frame | Grüner Veltliner |
| 0 | This opens up with scents of sweet spices and barrel toast, leading into a burst of strawberry fruit. It stiffens in the back end, with a phenolic bite to the tannins, and residual notes of bitter herbs. It's a muscula | Pinot Noir |
| 1 | This blend of Merlot, Petite Sirah, Malbec and Cabernet Sauvignon is a satisfying sipper for a fair price. Aromas of black cherry, dark cocoa, leather and charred beef lead into a creamy palate of espresso a | Red Blends |
| 2 | Pure, jubilant black cherry and mulberry flavors burst from this unoaked, fruit-focused red. A spicy, smoky 60-40 blend of intensely ripe Grenache and Syrah vinified with no added sulfur and unfined, it's delic | Rhône-style Red Blend |

Figure 2: Data samples, including index, description and variety.

# 2 Background & Related Work

A discovered prior work is a paper named *Convolutional Neural Networks for Sentence Classification* [2]. It investigated how the Convolutional Neural Network (CNN) can be applied to a series of experiments to train on top of pre-trained word vectors for sentence-level classification tasks [2], [3].

A simple CNN is trained with one layer of convolution on top of word vectors acquired from an unsupervised neural language model [2]. 100 billion words of Google News were taken to train the vectors [3]. Despite little tuning of hyperparameters [4], the simple CNN with one layer of convolution performed remarkably well on multiple benchmarks [2]. These include detecting positive/negative reviews from movie comments with one sentence per comment, which is in nature close to detecting grape variety from wine descriptions. The results suggest that the pre-trained vectors are good, universal feature extractors and can be utilized across datasets for various classification tasks [2]. These results proved that pre-trained word embedding models can be utilized with CNNs to classify sentences into keywords as output which serves as a feasible approach for the project.

# 3 Data Processing

The raw data is collected from Wine Enthusiast which has over 100,000 bottles of wine with information on vintage, variety, country, description, etc [5]. A Python scraping code is used to download and store the raw data in a csv file with only two fields: variety and description, based on the design of the system [6]. The description field is further transformed from sentences to lists of keywords that are comprehensible for the Natural Language Processing model [7].

First, the description (an entire string) is formatted into lowercases and then nltk.tokenize is used to split sentences into individual words (a list of strings). Then, GloVe is used to calculate the word embedding for each description where the converted vectors would be the input of the model. On the other hand, the distribution of grape varieties is unbalanced based on the diagram of grape variety from the dataset: 9% and 10% of the wine are made from Chardonnay and Pinot Noir with a total of 652 varieties as shown in Figure 3 below. Thus, the team decided to include only the top 26 varieties with the highest frequencies in the model. There are about 50,000 data samples remaining after the selection. One-hot encoding is used to convert wine varieties into labels of the model. In addition, class weight is added to the loss function to reduce the impact of an unbalanced dataset. Lastly, the training, validation and testing data is split with a ratio of 8:1:1.

```
Pinot Noir                  8268
Chardonnay                  6119
Bordeaux-style Red Blend    4535
Red Blends                  4432
Cabernet Sauvignon          4326
                            ...
Symphony                       1
Sauvignon-Sémillon             1
Feteasca Alba                  1
Sauvignon-Semillon             1
Kotsifali-Syrah                1
Name: variety, Length: 652, dtype: int64
```

Figure 3: Variety Distribution of the Entire Dataset.

# 4 Architecture

During the project proposal drafting phase, the only suitable network architecture we were exposed to for this task is the CNN. As a result, the CNN architecture was proposed as the primary model for the project. However, under Section 4.4 of the project proposal, it is indicated that we are willing to explore other neural network architectures if they are deemed to be more promising after gaining more insight into them. As a result, we were able to identify two additional candidate model architectures after the submission of the project proposal: the Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). In order to select the most adequate model for this task, the team decided to evaluate all three candidate models. The model with the best performance will be selected as the primary model for the project. Ultimately, the LSTM architecture was chosen as the primary model moving forward since it significantly outperformed the CNN and RNN models. An illustration of the LSTM model is shown in Figure 4 below.

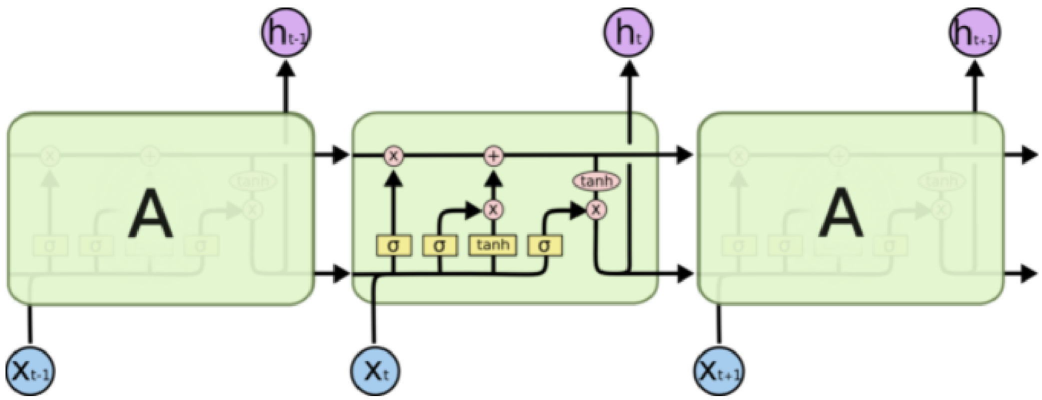


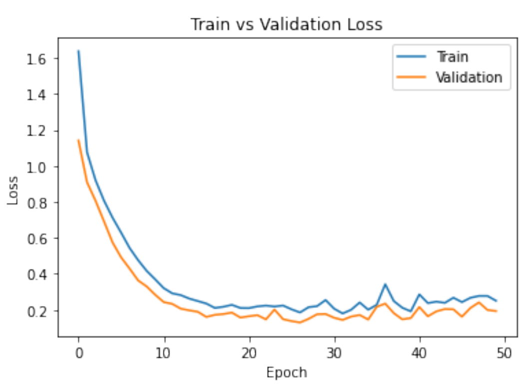

Figure 4: The architecture of a standard LSTM [8].

## 4.1 Input/Output

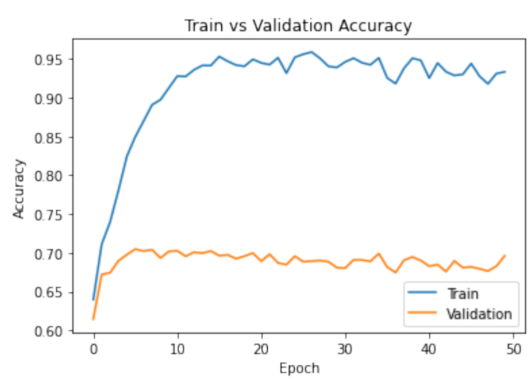
As previously described in Section 3, the input to the deep learning model will be processed wine reviews. The output of the deep learning model will be one-hot encoded labels of grape varieties.

## 4.2 Model Setup

Firstly, an `nn.Embedding` layer was introduced prior to the LSTM to allow the lookup of multiple words simultaneously. Secondly, the model was built with `input_size = 50` and `hidden_size = 200`. Lastly, the max-pooling and average-pooling of the LSTM outputs were concatenated and passed through a fully connected layer for classification. The LSTM is trained using the cross-entropy loss function (with class weights inversely proportional to the class frequency) and Adam optimizer with a batch size of 64, learning rate of 5e-3, and over 50 epochs. The training curves are shown in Figure 5 below.




(a) Loss of LSTM Model




(b) Accuracy of LSTM Model

Figure 5: Training/validation loss and accuracy of the LSTM model.

# 5 Baseline Model

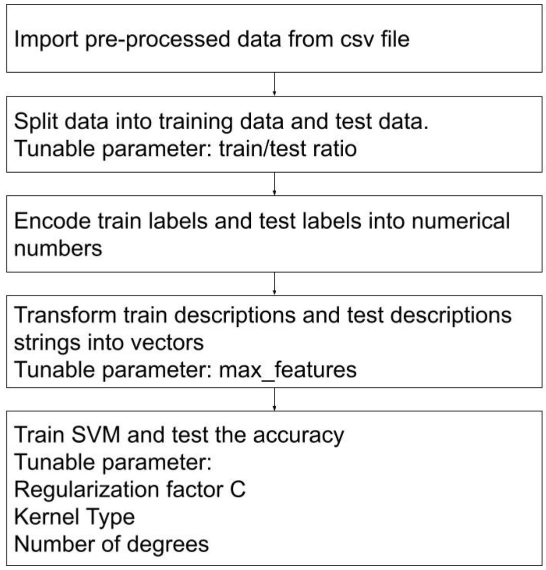



Figure 6: Flow diagram of baseline SVM model.

We were looking for a baseline model that is easy to implement and compare with. The baseline model chosen to compare the neural network against is a support-vector machine (SVM). SVMs are supervised learning models based on the Structural Risk Minimization principle. They are also proven to be well-suited for text classification problems prior to the prevalent usage of deep learning for the task [9]. Thus, a SVM can be utilized as a reasonable baseline model to gauge the performance of the neural network. Figure 6 shows the flow of our baseline model. The model first imports the pre-processed data, then splits data, encodes labels and transforms natural sentences into vectors. The model is built up with a Python tool kit called scikit-learn, which contains a built-in SVM model. Hyperparameters such as train/test data ratio, SVM kernel types and SVM kernel degrees were well tuned for the best validation accuracy.

# 6 Quantitative Results

## 6.1 Accuracy

Table 1: Validation and Testing Accuracy of Evaluated Models

| Accuracy | SVM | CNN | RNN | LSTM |
|---|---|---|---|---|
| Validation Accuracy | 67.00% | 23.52% | 54.61% | 69.61% |
| Test Accuracy | 69.11% | 21.18% | 56.06% | 69.91% |

In terms of quantitative results, the validation and testing accuracies of the utilized models are summarized in Table 1 above. The accuracies of the models are calculated by comparing the most confident output prediction made by the model with the true label. If the predicted class and the true label are the same, then it is counted as one correct prediction. As shown in the table, the LSTM model has the best validation and testing accuracies compared to other deep learning models. However, the baseline support vector machine model is also able to perform quite well with just a slightly lower accuracy compared to the LSTM model.

## 6.2 Confusion Matrix

Predicted Label

Real Label

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [[65 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 0 | 1 | 1 | 3 | 4 | 0 | 7 | 0 | 20 | 1 | 0 | 1 | 0 | 1 | 3 | 0 | 3] |
| [ 0 | 200 | 0 | 0 | 0 | 1 | 3 | 9 | 0 | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 0 | 13 | 0 | 0 | 1 | 6 | 0 | 0 | 0] |
| [ 0 | 0 | 96 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 44 | 0 | 0] |
| [ 1 | 0 | 0 | 158 | 0 | 2 | 6 | 0 | 0 | 1 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 10 | 0 | 2 | 0 | 0] |
| [ 0 | 0 | 0 | 0 | 65 | 4 | 0 | 11 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 9 | 3 | 0 | 0 | 1 | 0 | 0 | 0] |
| [ 2 | 1 | 0 | 4 | 3 | 194 | 25 | 6 | 6 | 0 | 4 | 0 | 6 | 2 | 3 | 0 | 2 | 1 | 11 | 2 | 2 | 2 | 19 | 2 | 1 | 0] |
| [ 2 | 2 | 2 | 10 | 1 | 16 | 650 | 6 | 10 | 0 | 10 | 15 | 2 | 12 | 0 | 7 | 2 | 35 | 9 | 0 | 11 | 11 | 10 | 3 | 4 | 5] |
| [ 0 | 11 | 0 | 0 | 7 | 6 | 2 | 177 | 1 | 4 | 1 | 0 | 0 | 1 | 1 | 1 | 2 | 1 | 41 | 9 | 0 | 2 | 4 | 0 | 1 | 0] |
| [ 0 | 0 | 0 | 0 | 1 | 3 | 12 | 0 | 97 | 0 | 0 | 0 | 3 | 9 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1] |
| [ 0 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 37 | 0 | 0 | 0 | 0 | 2 | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0] |
| [ 7 | 0 | 15 | 9 | 0 | 8 | 22 | 0 | 0 | 0 | 271 | 14 | 3 | 10 | 1 | 4 | 0 | 30 | 2 | 0 | 4 | 4 | 1 | 41 | 0 | 7] |
| [ 0 | 0 | 0 | 0 | 0 | 0 | 23 | 3 | 0 | 0 | 7 | 67 | 0 | 0 | 0 | 3 | 0 | 18 | 3 | 0 | 0 | 5 | 0 | 1 | 0 | 1] |
| [ 4 | 0 | 0 | 1 | 2 | 1 | 8 | 0 | 10 | 0 | 0 | 0 | 106 | 17 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 0] |
| [ 9 | 0 | 0 | 3 | 0 | 3 | 16 | 0 | 6 | 0 | 11 | 2 | 8 | 381 | 0 | 0 | 0 | 17 | 4 | 2 | 6 | 2 | 1 | 0 | 0 | 3] |
| [ 0 | 1 | 0 | 0 | 0 | 1 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 61 | 0 | 4 | 0 | 10 | 0 | 1 | 0 | 2 | 0 | 0 | 0] |
| [ 9 | 0 | 0 | 1 | 0 | 0 | 8 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 1 | 47 | 0 | 7 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 1] |
| [ 0 | 8 | 0 | 0 | 1 | 2 | 2 | 6 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 52 | 0 | 14 | 0 | 0 | 0 | 4 | 0 | 1 | 0] |
| [10 | 0 | 0 | 3 | 0 | 0 | 41 | 1 | 0 | 0 | 8 | 24 | 1 | 11 | 1 | 7 | 0 | 259 | 4 | 0 | 15 | 16 | 1 | 7 | 0 | 19] |
| [ 1 | 8 | 0 | 0 | 5 | 4 | 16 | 39 | 2 | 6 | 0 | 0 | 0 | 2 | 10 | 1 | 6 | 2 | 464 | 7 | 0 | 3 | 16 | 0 | 12 | 1] |
| [ 0 | 0 | 0 | 0 | 1 | 3 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 6 | 39 | 2 | 0 | 2 | 0 | 0 | 0] |
| [ 4 | 1 | 0 | 0 | 0 | 0 | 14 | 1 | 1 | 0 | 2 | 0 | 3 | 4 | 0 | 0 | 0 | 14 | 0 | 0 | 31 | 1 | 0 | 0 | 0 | 2] |
| [ 7 | 1 | 1 | 10 | 0 | 3 | 24 | 0 | 1 | 0 | 1 | 14 | 1 | 1 | 0 | 5 | 1 | 18 | 0 | 0 | 0 | 116 | 0 | 1 | 0 | 1] |
| [ 1 | 3 | 0 | 1 | 0 | 5 | 4 | 3 | 0 | 3 | 0 | 0 | 1 | 1 | 2 | 0 | 1 | 0 | 12 | 0 | 0 | 0 | 93 | 2 | 3 | 0] |
| [ 1 | 0 | 15 | 1 | 0 | 0 | 7 | 0 | 0 | 0 | 13 | 2 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 1 | 0 | 121 | 0 | 0] |
| [ 0 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 12 | 0 | 58 | 0] |
| [ 5 | 0 | 0 | 1 | 0 | 1 | 19 | 0 | 0 | 0 | 3 | 6 | 1 | 8 | 0 | 2 | 0 | 21 | 1 | 0 | 6 | 2 | 0 | 2 | 0 | 32]] |

Figure 7: Confusion Matrix of LSTM Model.

In order to gain more insight into the performance of the LSTM model, the confusion matrix is also calculated and generated for predictions made over the test set as shown in Figure 7 above. The precision and recall values of each class are also calculated based on the confusion matrix, which will be further discussed in Section 9.

# 7 Qualitative Results

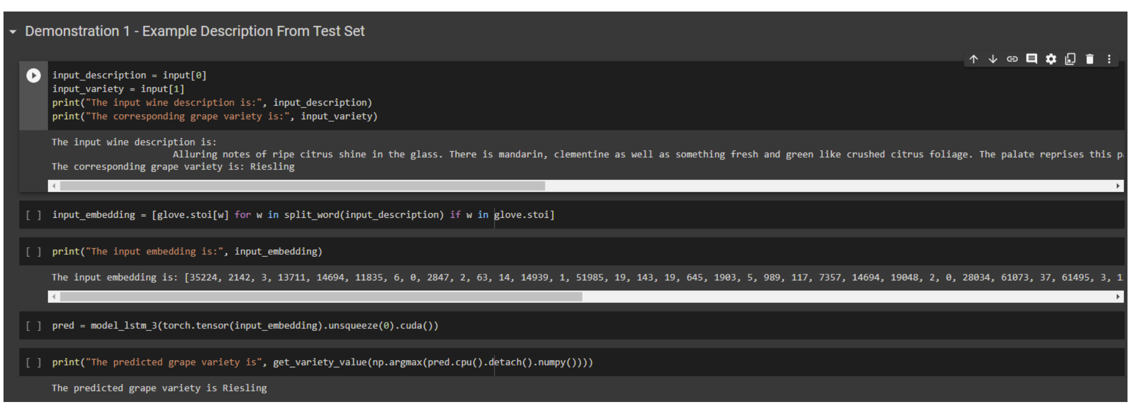



Figure 8: Sample output of the LSTM model given an input wine description from the test set.

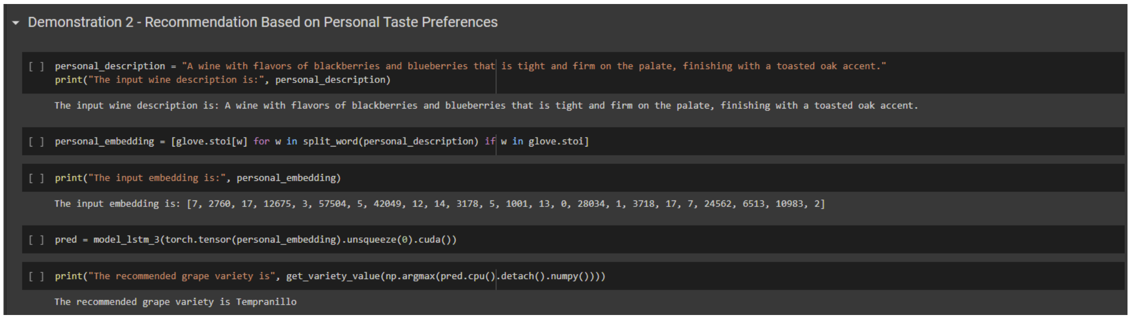



Figure 9: Sample output of the LSTM model given a personal wine taste preference.

The functions and characteristics of different deep learning models are evaluated and compared with each other. Based on the test accuracies generated by the CNN model, RNN model, and LSTM model separately, the LSTM model is finally chosen for its highest validation and testing accuracy (about 70%). Figures 8 and 9 are demonstrations of sample output from the LSTM model based on two different types of input. The outcome (final test accuracy) is not surprising since LSTM is remarkable for memorizing past data in memory and is a modified version of RNN, which is a generalization of feedforward network that has an internal memory [9]. These two architectures are efficient in word and sentence prediction. They can use their internal state memory to manage and process sequences of input and each sample can be reckoned as dependent on previous ones. LSTMs add a long term memory to RNNs to improve accuracy and are able to outperform the RNN model given longer sequences, which is the reason why it has comparatively higher accuracy.

# 8 Model Performance on New Data

5,000 new testing data were used to examine the LSTM model that was ultimately produced, aiming at using a new (unused) dataset to test the trained model. There are two types of sample inputs and outputs when testing on new data. For the first type, the descriptions the team got from the website were transformed into GloVe embeddings and can be checked by seeing if the outputs are of the correct type. For instance, a description of Riesling was transformed into GloVe embedding and put into the `get_variety_value` function and got the correct outcome: Riesling. In order to amplify the functionality of the model, the team designed a second type of input. The second type is to input the personal description of a certain type of wine and let the model determine which type of wine it should recommend to the user, which is a common need for both general wine consumers and wine enthusiasts, further broadening the applications of the trained model in practice. Also, based on the nature of the project: predicting wine type based on descriptions written by the tasters, the accuracy could be directly calculated after running the model and require no further steps unlike other projects that utilize generative models.

# 9 Discussion

The CNN model was evaluated before the final model was determined and had a bad performance in accuracy. CNN is prominently used in image content processing and higher image representation extraction. In the CNN model, 2 one-dimensional convolutional layers, one maxpool layer, and two fully-connected layers were used. CNN is more likely to be used in image processing and prediction, acting poor in word/sentence processing or prediction. A possible reason for this performance is that CNN's nodes do not form a cycle and are not good at memorizing related contexts. Also, CNN has no overfitting protection and only has fixed input size with no order considered.

The baseline model is performing higher than expected in test accuracy. SVM is a linear model for classification and regression problems. It is proven to be well-suited for text classification problems [9]. One of the advantages of SVM is that it can learn independent of the dimensionality of the feature space. When learning text classifiers, one has to deal with very many features. In fact, there are about 20,000 to 30,000 common English words used by a native English speaker, and this doesn't include some professional terms that are used in wine description. Since SVM has overfitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces and outperform initial expectations.

Table 2: Precision and Recall of Sample Grape Varieties

| Grape Variety | Precision | Recall |
|---|---|---|
| Merlot | 30.3% | 36.4% |
| Cabernet Franc | 36.8% | 41.0% |
| Riesling | 86.0% | 79.0% |
| Rhône-style Red Blend | 90.0% | 78.8% |

Some interesting results can also be interpreted from the confusion matrix of the LSTM model shown in Table 2 above. For example, predictions made on the grape varieties Merlot and Cabernet Franc have the lowest precision and recall values out of all the classes. This could be a result of Cabernet Franc being the parent grape of Merlot, and both grapes having relatively mild and similar tastes compared to each other, which can be difficult for the deep learning model to distinguish based on taste descriptions. On the other hand, grape varieties Riesling and Rhône-style Red Blend have the highest precision and recall values out of all grape types. This is due to the fact that their distinct flavors can be easily detected based on taste. Wine made with Riesling tends to have high acidity that resembles flavors of citrus such as lemons. While wine made with the Rhône-style Red Blend tends to have rounded warm tastes with hints of red fruits such as cherries and strawberries.

# 10 Ethical Considerations

Based on the source website, this system is only valid for English since the description of wines are in English and the Natural Language Processing model that was used only works for English according to the logic behind tokenization texts. Any users that do not understand English would be excluded from using this model due to language barrier. In addition, the system is limited to the grape varieties that exist among the 26 varieties with highest frequency from the website. If a wine with a new variety appeared, the accuracy of the model would be affected since the output would never match the label.

# 11 Project Difficulty/Quality

One of the most difficult parts of the project is the highly unbalanced dataset with a large number of classes. There are over 600 labels in the original dataset where the largest class occupied 10% of the data sample as mentioned in the Data Processing section. If all the labels are used in the model with one-hot encoding, this would significantly increase the number of parameters and decrease the efficiency of the model. In addition, it would be impossible to produce output that matches with the label with smaller classes that occupied less than 0.002% of the dataset. As a result, the team only includes the 26 classes with the largest size. However, the data distribution is still highly unbalanced after the selection. The team chose to add class weight in the loss function of the model to improve the reliability of the model performance.

The team trained 3 different models for this project other than the baseline model: CNN, RNN and LSTM. The CNN model has a poor performance which matches the expectation since it is not suitable for sequence learning. LSTM has a better performance compared to RNN since it is designed to have a better "memory" of the input in the previous states. Overall, LSTM achieved the highest testing accuracy among all models. The final performance is about 70% which meets the expectation of the team for this project.

From data processing to tuning hyperparameters for different models, the team uses knowledge from lectures and labs and works beyond the requirement of labs by collecting data directly from websites, modifying architecture of the models, and tuning different hyperparameters that are relevant for these models.

## 12 Colab Link

The link to the Colab file containing the final LSTM model is: https://colab.research.google.com/drive/1HgCRYfdITAMwa86Ay1GAAnQO9bQcYIrX?usp=sharing

## References


[1] J. Conway, "Wine market in canada: Statistics and facts," Dec 2020. [Online]. Available: https://www.statista.com/topics/2996/wine-market-in-canada/#:~:text=Retail%20sales%20of%20wine%20in,in%20Nunavut%20purchase%20the%20least

[2] Y. Kim, "Convolutional neural networks for sentence classification." Jun 2021. [Online]. Available: https://arxiv.org/pdf/1408.5882.pdf.

[3] M. Etal, "Google code archive for pre-trained vectors." Jun 2021. [Online]. Available: https://code.google.com/archive/p/word2vec/.

[4] M. D. Zeiler, "Adadelta: An adaptive learning rate method," Dec 2012. [Online]. Available: https://code.google.com/archive/p/word2vec/.

[5] "Best wine ratings amp; reviews online: Wine enthusiast magazine."

[6] Z. Thoutt and V. Mukhtarulin, "Zack thoutt - wine deep learning," July 2020. [Online]. Available: https://github.com/zackthoutt/wine-deep-learning/blob/master/scrape-winemag.py

[7] G. Bedi, "Simple guide to text classification(nlp) using svm and naive bayes with python," July 2020. [Online]. Available: https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34

[8] C. Olah, "Understanding lstm networks," Aug 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[9] "Text classification [U+200A]—[U+200A] rnn's or cnn's?" 2020. [Online]. Available: https://towardsdatascience.com/text-classification-rnns-or-cnn-s-98c86a0dd361