IDA - Department of Computer and Information Science

Linköping Unviersity

Text Mining Project 2021

Taste of wine analysis and wine classification in different regions

Weng Hang Wong

wenwo535

March 17, 2021

Abstract

This study is to explore how the features and characters are described among different regions. And the next step is to create a model to classifiy wine region bases on the wine reviews. The project used TF-IDF vectorizer and Count vectorizer of 130k samples from the top 10 regions to analysis adjectives and nouns in unigram dn bigrams that are represented the characteries of the region. Statistical models as Navie Bayes and Logistics Regression, as well as nueral netword model as Multilayer Preceptron and Convolution Neural Networds, are used to classify region from the tasting reviews. The result indicated that the unigram model has a higher performance among all the classifiers to classifiy the region, and also the insights of wine features are extracted by comparing unigram and bigram models.

Total words: 3687

Table of Content

1. Introduction

Wine selection is often complicated since it is a subject of expertise and experience. The enormous choices and diversity of wine can be overwhelming and confusing to choose, especially for those who do not have in-depth knowledge of wine. Wine experts are capable of describing the characteristics of wine, such as the details of smell and flavor, that can provide a general idea in words to normal people. The basic characteristics of wine mainly categorized into the level of sweetness, fruitiness, tannin, acidity and light/full body.

We are interested in the relationship between flavor and country, and how to select wine based on personal flavor. Put it in other words, we would like to have a wine recommendation from an expert according to the preference of our wine flavor. Given the wine reviews from the experts, we will classify how wine flavor diverse in different countries using text mining techniques.

The aim of this study is to analyze the wine characteristics among different countries, predict wine region using wine review text.

2. Data Description

The data is downloaded from the website: https://www.kaggle.com/zynicide/wine-reviews, the owner scraped the data from WineEnthusist during 2017. The data is a csv format which contains 130k wine reviews with 14 features such as country; description; designation; points; price; province; region1; region2; taster_name; taster_twitter_handle; title; variety; winery. In this study, the variable ‘country’; ‘description’; ‘points’ are used for the training. ‘country’ is where the wine are produced; ‘description’ is the reviews that written by different wine experts, it contains around 30-50 words for each, which can be considered precise in terms of describing the characteristics of wine. ‘points’ is the score of wine that given by the experts with range 80 – 100.

2.1 Data preprocessing

In order to have a relatively balanced data, the dataset is extracted with 10 most frequent countries, which are Italy, Portugal, US, Spain, France, Germany, Argentina, Chile, Australia and Austria.

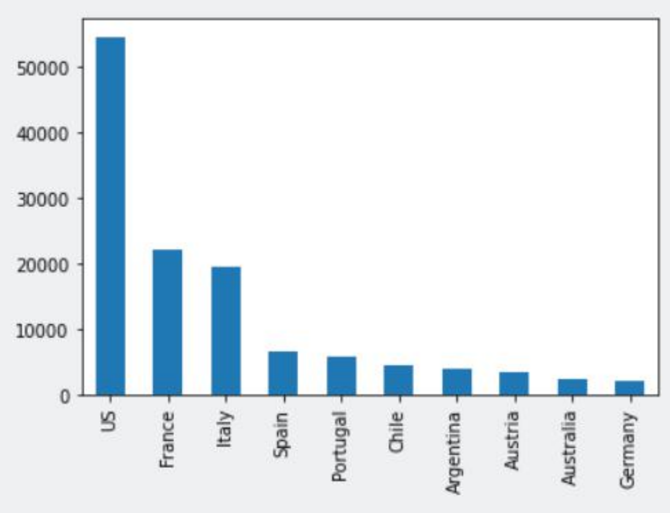


Figure 1: Instances of 10 most frequent countries

Second, the description text is tokenized and lemmatized using Spacy through the 12k rows of data. In order to keep only the adjectives and nouns of the description, not only the stop words, but also the verbs are also removed from the text.

Example:

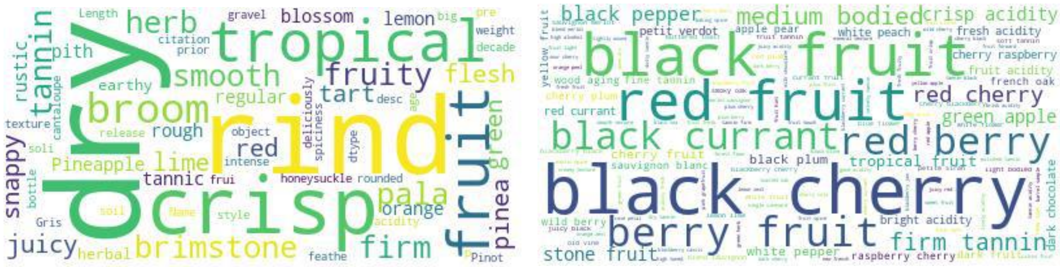
Original text: "Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity."

Preprocessing text: 'aroma tropical fruit broom brimstone herb palate overly expressive unripened apple citrus sage alongside brisk acidity'

2.2 Data investigate with unigram and bigram

To investigate how the tasting review looks like, we can take a look into the frequency of the description words. From the generated wordcloud, some words that most frequent appeared are: dry, rind, crisp, fruist, tropical, tannin, jucy, etc. These are the keywords that mostly used to describe the characteristics of wine (See figure 3.1).

To further analyze the information in the description, n-gram phrase is taken in place includes the sequences of adjacent words (See figure 3.2). N-gram analysis
At the flrst glance, the bigram wordcloud looks more accurate and informative than unigram wordcloud. For example, instead of ‘fruit’ in unigram, there are ‘black_fruit’, ‘red_fruit’, ‘berry_fruit’ in bigram analysis.



(1) unigram (2) bigram

Figure 2: WordCloud of most frequent unigram and bigram words

3. Theory

Texting mining is useful to extract the valuable information and find the pattern from large amount of text documents. Text mining techniques are commonly used in different types of domains, such as natural language processing(NLP), information retrieval, text classification and text clustering[i].

3.1 Preprocessing techiques

Text preprocessing is essential before starting text mining process, removing the irrelavent text can improve the efficiency of the process and increase the accuracy rate for the algorithms.
There are multiple techniques of preprocessing:

1. Tokenization

A method to tokenize the content into individual word[i].

2. Stop word removal

A method to remove words that is frequent occurred but is not useful for the text mining application. Stop words often make the text heavier and they are less important for analysis[i].

3. Lemmatization
Lexme means a set of word forms that have the same fundamental meaning. Lemmatization assures the word only retains it's original form in order to have a better performance on processing.

3.2 Term frequency-inverse document frequency (tf-idf)

TF-IDF is a statistical measure used to evaluate the importance of a word by collecting the weighted frequency in a document[ii]. The idea of TF-IDF is that the more ocurrancy of the words, the more important the words are, but meanwhile, the TF-IDF scores decrease if the frequency is too high. TF-IDF is a score for a term *t* in a document *d*, it is computed by multiplying term frequency matrix *tf(t,d)* and inverse document frequency matrix *idf(t).* Inverse document frequency targets the rare words, a bigger score closer to 1 means the word is appear rarely and vice versa.

The tf-idf weight is computed as:

$$tf - idf(t,d) \; = \; tf(t,d) \bullet log \; \frac{N}{df(t)}$$

$$tf(t,d) \; = \; log(1 + freq(t,d) \;)$$

where t and d denotes a term t in a document d, N denotes the number of documents in the collection.

3.3 N-gram

N-gram is a N-continue sequence phrase, phrases are essential in natural language processing and text mining. Phrase of words often contains meaning, each n-gram phrase can be a new component of corpus for text mining. For example, "black_berry" is a 2-gram phrase that frequent used on documents. N-gram analysis is useful to extract text data and it helps to associate phrases with topics in a more acurrate way.

3.4 Bags of Words (BoW)

Bags of Words is a common method to extract text for model used, it is a representation of text and also descriibes the occureence of words in a document.[iii]

3.5 Naive Bayes (NB)

Naive Bayes is a linear classifier based on Bayes Theorem. Assume a data set X with n features $X_1 \; ... X_n$ and Y is the label $Y_1 \; ... Y_n$.

Naive Bayes Classifier:

$$\mathrm{P}(Y|X) \; = \; \frac{P(Y) \prod P(X_i|Y)}{P(X)}$$

$$\hat{y} \; = \underset{n \in \{1..n)}{argmax} P(Y) \; \prod_{i=1}^{n} p(X_i|Y)$$

where $\hat{y}$ is the prediction label.

3.6 Multinomial Logistics Regression

Multinomial Logistic regression is a simple extension of logistics regression model, while logistic regression is only a binary classifier. Multinomial logistic regression is a predictive analysis that explains the relationship between one and more independent variables.[iv] It is modified by using the softmax function[v].

$$softmax(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

where x_i is the input data and n is classes.

The model including the softmax function is :

$$\hat{y} = softmax(xW + b)$$

where W is the weight matrix and b is the bias term.

L2 norm penalty is a popular penelty that used for mulinomial logisctic regression, which adds the weighted sum of the squared coefficients to the loss function.

3.7 Multilayer Preceptron (MLP)

A MLP is a feed-forward artficial neural netword model with one or more layers between input and output layer. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. The layers of nodes in a directed graph, with each layer fully connected to the next one. MLP can solve problems that are not linearly separable[vi].

Assume input ${x^{l}}_{m}$ is in l layer m nodes, ${y^{l}}_{m}$ is the ouput value and ${b^{l}}_{m}$ is the bias value, ${w^{l-1i}}_{m}$ is the weight.

$$x_m^l \quad = b_m^l \sum_{i=1}^{k} w_{im}^{l-1} \bullet y_i^{l-1}$$

$$y_m^l = f(x_m^l)$$

Loss function is :

$$E = E(y_1^l, \dots., y_h^l) \ = \sum_{j=0}^{h} y_j^l - t_j^2$$

where lth output layer, t_j is j th node's expected output.

Since MLPs are fully connected, each node in one layer connects with a certain weight w_{im} to every node in the following layer.

$$w_{im}^{l-1} = w_{im}^{l-1} - \varphi \bullet \frac{\partial E}{\partial w_{im}^{l-1}}$$

where φ is the learning rate.

3.8 Convolutional Neural Networks (CNN)

One of the deep neural networks is convolutional neural networks. A convolutional neural networds contains one or more convolutional layers. Each layers consists of three stages: convolutions to compute linear combination z, vector stage to compute activation y, and pooling function.[vii]

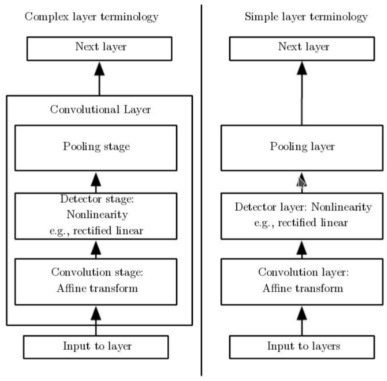


Figure 3: The components of a typical convolutional neural network layer[viii]

CNN use the convolution kernel to slide the image for information extraction, and control the legth and width of the image by strides and padding.

Kernel controls the amount of learning parameters.They are the small window slide over the input matrix. In this case of text classification, a convolutional kernel will be a sliding window and look at embeddings for multiple words.

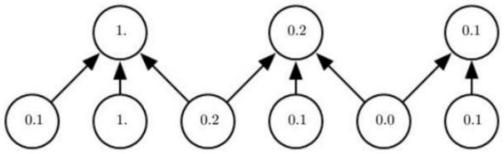


Figure 4: Pooling with downsampling[ix]

Strides decides how far the kernel slide over the input matrix. Pooling layer section reduces the amount of parameters if the input size is too large, which can improve statistical efficiency and reduce. In this case, max pooling is applied on the convolutional networks to take out the maximum from a pool, which can interpretate the input matrix into a small pooling matrix with a certain stride value and kernel.

Activation function is often applied in order to activate non-linear transformation on the input signal, and ReLU function computes the function f(k) = max(0,k). In this case, ReLU activation function is used in the CNN model.

4. Methodology

4.1 Compare wine characteristics with word frequency

As data preprocessing part mentioned, it is preceed with the 10 most frequent appeared countries in the dataset, which contains the US, France, Spain, Italy, Porgutal, Chille, Germany, Argentina, Austria and Australia. Wine reviews wrttien by the wine tester are critical information for understand the features of the wine. To compare the main features of each country's wine, TF-IDF weighted is used with counting word-frequency. Figure 6 and Figure 7 shows the unigram and bigram analysis using TF-IDF scores, the terms with highest TF-IDF determined the main character of wine in each region.

4.2 Classifications on wine regions with unigram and bigram models

In this classification experiment is to build a model to classifiy regions and perform a prediction based on the wine reviews. The dataset contains around 130k data, which is split into training data and test data with each is sampled 40% and 10% from the dataset.

The baseline model in this study is using Naive Bayes method. It is is used to predict the country labels of wine reviews by putting a MultinomialNB with countvectorizer and tfidfvectorizer respectively in a pipeline. CountVectorizer in skleran module can convert a collection of text doucments to a matrix of token counts. TF-IDF can convert the raw documents to a matrix of TF-IDF scores matrix, while weighting the more relevant terms in a lower score. For both word-count and tf-idf, unigram and bigram variations are test in this baseline method, and the better n-gram method is chosen according the result. To evaluate the result, we can look at the F1-score to do the comparison different N-gram methods and differnet vectorizers.

To compare the result with Navie Bayes method, here we use traditional machine learning method as logistic regression; and also Neural Netword methods, such as MLP model and CNN model. After a few experiement on applying n-gram on the above models, unigram method has the better result than ungiram; so we only show the unigram on the following result.

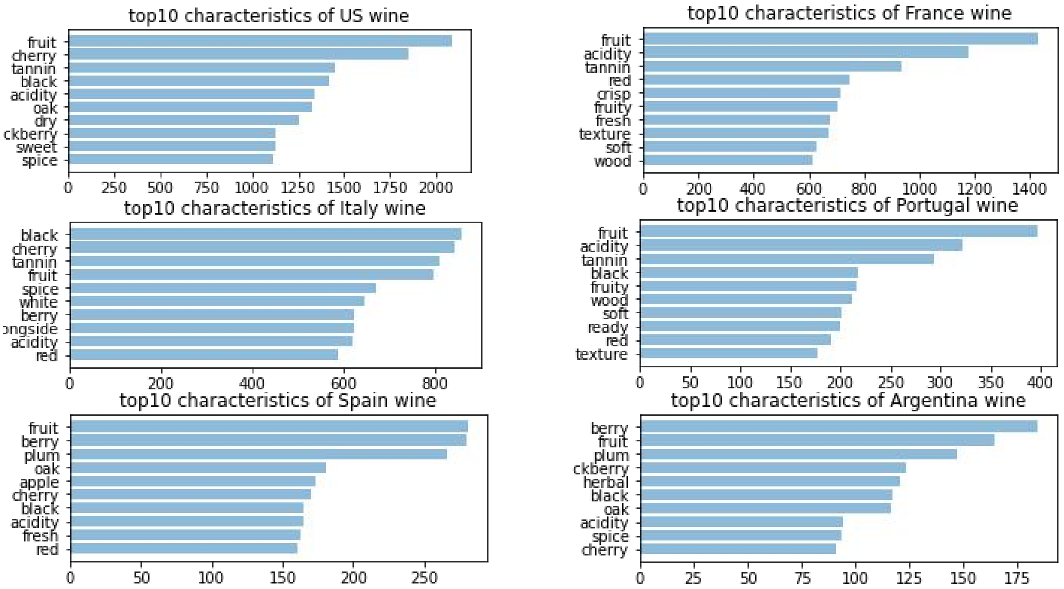
In logistics regression, the multinomial classes is applied in order to predict 10 classes efficiently. Logistics regression and MLP method are putted into a pipeline respectively with CountVectorizer, so that the text transformation process can be more efficient and have less time costs .

In CNN model, the training and test data needed to re-tokenized and reshaped in order to fit in the CNN model. The data are toenized using pad_sequences and labels are also reshaped with to_categorial. Embedding layer, convolution1D layer and maxpooling1D layer are used for building the CNN architecture.

5. Result

5.1 Comparison of unigram and bigram

Unigram score:



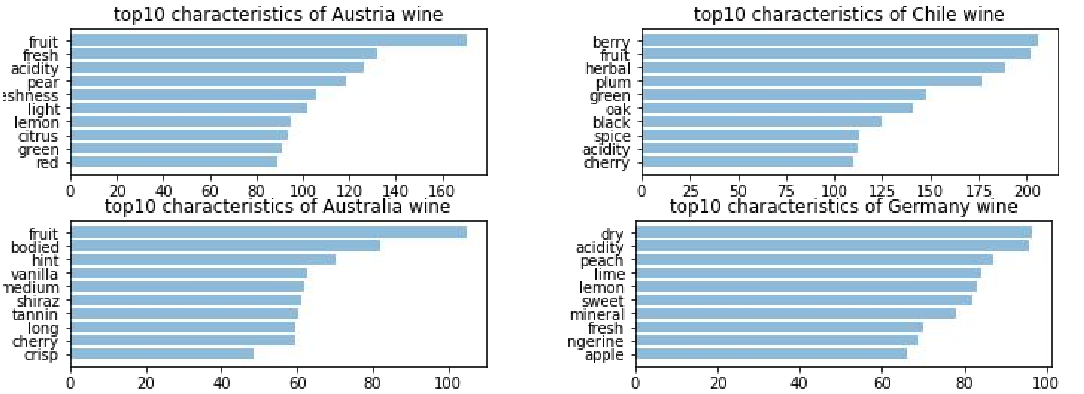


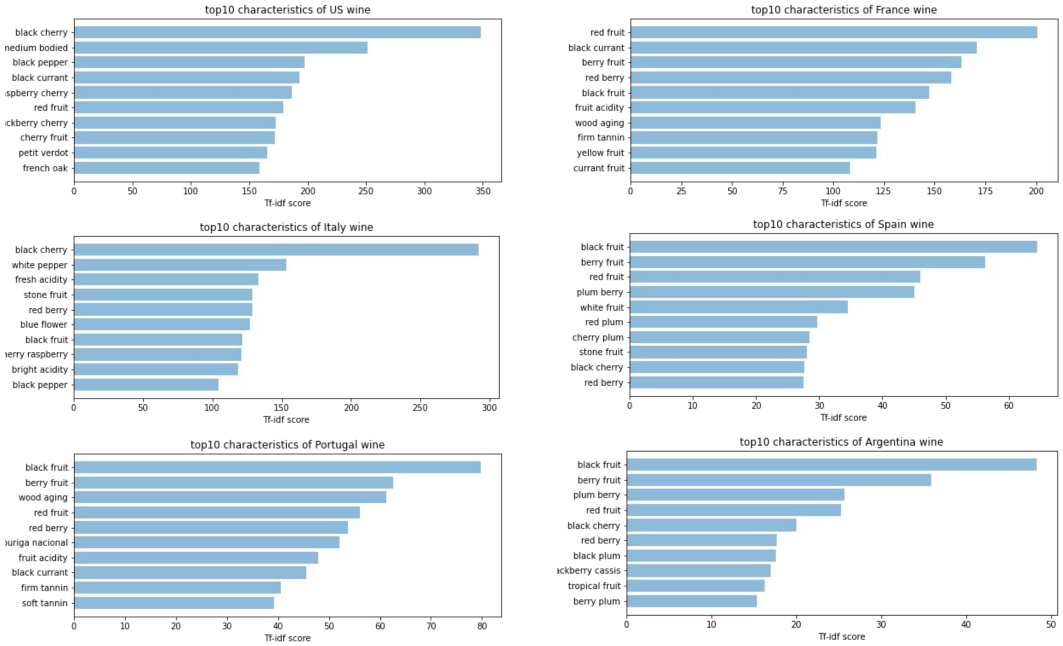
Figure 5: y axis is the top 10 corpus with highest frequency in 10 countries. x axis is TF-IDF scores.

The 10 hightest counts indicate the crucial features of the wine in different wine such as ‘berry’, ‘tannin’, ‘plum’, ‘spice’, ‘oak’, ‘vanilla’. However, the disavantage of unigram is that it is ambious in terms of the word ‘fruit’, ‘black’, ‘texture’. See Figure 5.

From the result of word frequency, the wine in the US, France, Italy, Spain Australia and Portugal are quite similar to each other. They contain the word as ‘cherry’, ‘tannin’, ‘acidity’, and ‘black’, which implys the wine is tend to be medium to full body with strong taste of berry fruits.

The words in Argentina and Chille contain ‘plum’, ‘oak’ and ‘herbal’ in aroma or taste. It indicates the wine tend to have higher acidity and in the range of light to medium body. Austria and Germany wine are tend to be light body with high acidity. The words ‘acidity’ has a high score in both country, as well as ‘lemon’, ‘pear’, ‘lime’, ‘apple.

Bigram TF-IDF score:



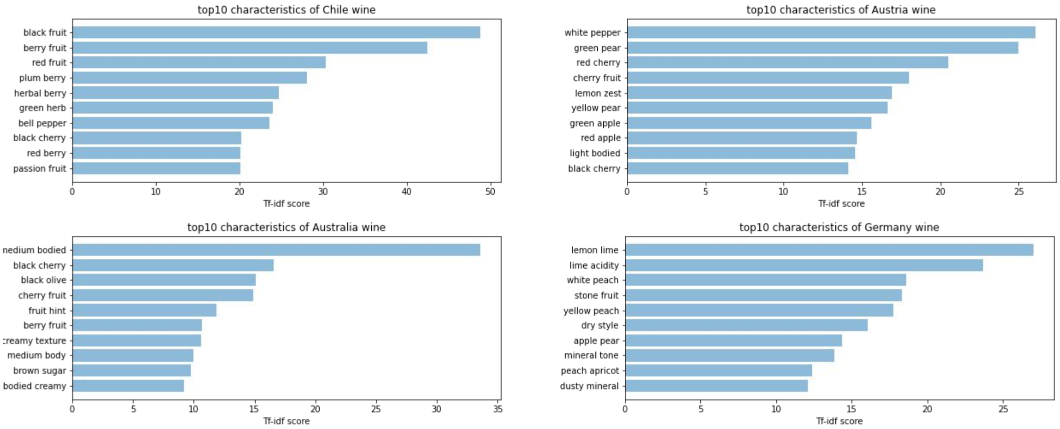


Figure 6: y axis is the top 10 corpus with highest frequency in 10 countries. x axis is TF-IDF scores

The bigram terms show more precise information in each country than unigram, such as ‘red fruit’, ‘black currant, ‘black cherry’ instead of ‘fruit’ or ‘berry’ in unigram showed. However, the bigram terms in word counts mainly focus on fruit terms and igorned some useful terms as ‘dry’, ‘tannin’, ‘oak’, etc.

Basically, in the regions as France, the US, Italy, Australia and Progutual; the unigram and bigram features are contains many ‘black fruit’, ‘berries’, ‘tannin’, so we can conclude that the wine in these wine are relatively strong in taste and have a fruity aroma.

In the contrast, the features are concentrated on ‘fresh’, ‘dry style’, ‘lemon’ in description, meaning that wine in Germany and Austria are relatively dry and have a fresh and higher acidity character.

Argentina and Chille are in the middle between the US group and the Germany group in terms of wine strongness in taste.

5.2 Classification models on unigram and bigram

Naive Bayes model

	Unigram	Bigram
Count words	**0.84**	0.79
TF-IDF	0.75	0.69

Table 1: F1-score comparison

The prediction with using naive bayes model has a good accuracy of 0.84 on CountVectorizer with unigram parameter. The TF-IDF result is comparatively worse than the word frequency result.

Models Comparison using CountVectorizer

	Unigram	Bigram
Logistics regression	**0.86**	0.80
MLP	**0.86**	0.81

CNN	0.77	0.78

Comparing with three models, logistics regression and MLP words well on the prediction of the 10 classes. Notice that MLP took longer time process on training the model, logistics regression model is more efficient in this case. Both of them have similar performace on both unigram and bigram predictions.

CNN model has the worst performance among all models with 0.77 and 0.78 on unigram and bigram prediction. Both of the models are trained with 20 epochs and 200 batch size. The learning rate of the ‘adam’ optimizer is 0.1. As the result was time-consuming and low accuracy, CNN model in text prediction is not recommanded.

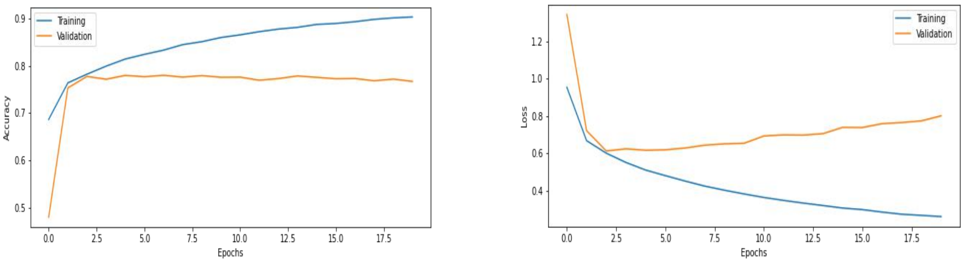


Figure 7: Accuracy and Loss figure of CNN model in 20 epochs (Unigram)

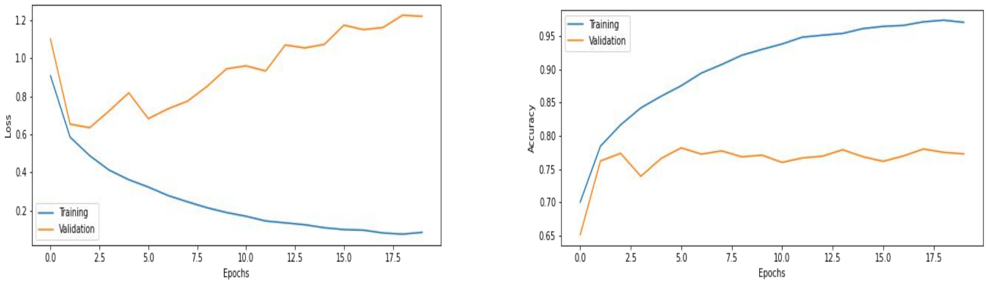


Figure 8: Accuracy and Loss figure of CNN model in 20 epochs (Bigram)

6. Conclusion

Thought the text mining analysis on wine reveiws, the main characters are also extracted and analyzed by unigram and bigram models with TF-IDF scores; and the wine regions are successfully predicted by statistical models and neural network models.

From the wine character analysis by bigram and unigram models, the features of wine are extracted among the top 10 countries. In conclude, if a person like to have a higher tannin and fruitness wine, the US, Italy, Fance, Porgutal and Australia wine are recommended. Especically the wine is more balanced in France and the US. Spain, Argentina and Chilie are quite smiliar in taste, they are suit for the people who likes fruity wine more not so favored in tannin. Last but not least, wine in Germany and Austria are in the flavoured of fruitiness with higher acidity, which often comes with dryness in these wine; it is recommended to people who like light wine with a hint of fresh fruit.

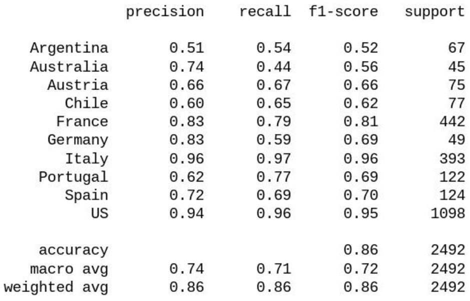
7. Discussion

The wine character analysis by bigram and unigram models can be extended to become a wine recommendation system. The frequency of feature scores can used to leverage the level of the charateristics, and word-embedding can be applied here to calculate the distance of words in order to categorized feautures into: tannin, sweetness, dryness, fruitness and acidity. Therefore, a wine recommendation system can based on the scores to recommend the wine that match with the costomzied categorized-features.

The classification of the wine region can be improved by categories their variety at the beginning. The variety means the grapes of the wine and the color red, white, rose of the wine. Since each variety of the wine can be diffierentiate better among different regions, the wine characters in text will be obvious to classify. Therefore, the classification will be improve and more accurate.

In this project, the wine regions are successfully classified. However, the CNN model did not perform well as expected. Sometimes, the hyperparameters in CNN model are the essential part to adjust. The unsatisfied result of the CNN model could be because of the failure to optimaized and tuned the hyperparameters.

8. Appendix

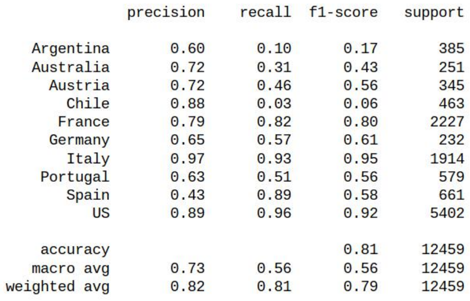


```
              precision    recall  f1-score   support

   Argentina       0.51      0.54      0.52        67
   Australia       0.74      0.44      0.56        45
     Austria       0.66      0.67      0.66        75
       Chile       0.60      0.65      0.62        77
      France       0.83      0.79      0.81       442
     Germany       0.83      0.59      0.69        49
       Italy       0.96      0.97      0.96       393
    Portugal       0.62      0.77      0.69       122
       Spain       0.72      0.69      0.70       124
          US       0.94      0.96      0.95      1098

    accuracy                           0.86      2492
   macro avg       0.74      0.71      0.72      2492
weighted avg       0.86      0.86      0.86      2492
```

Table 2: MLP unigram F1-score

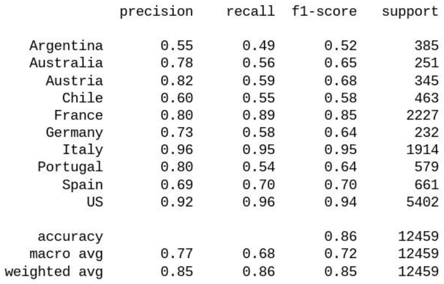


```
              precision    recall  f1-score   support

   Argentina       0.60      0.10      0.17       385
   Australia       0.72      0.31      0.43       251
     Austria       0.72      0.46      0.56       345
       Chile       0.88      0.03      0.06       463
      France       0.79      0.82      0.80      2227
     Germany       0.65      0.57      0.61       232
       Italy       0.97      0.93      0.95      1914
    Portugal       0.63      0.51      0.56       579
       Spain       0.43      0.89      0.58       661
          US       0.89      0.96      0.92      5402

    accuracy                           0.81     12459
   macro avg       0.73      0.56      0.56     12459
weighted avg       0.82      0.81      0.79     12459
```

Table 3 MLP bigram F1-score

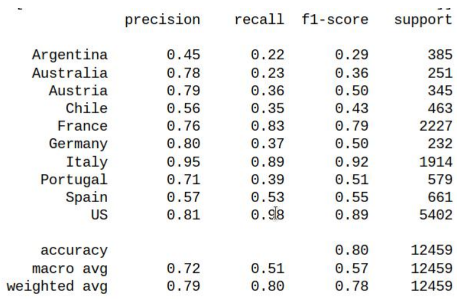


```
              precision    recall  f1-score   support

   Argentina       0.55      0.49      0.52       385
   Australia       0.78      0.56      0.65       251
     Austria       0.82      0.59      0.68       345
       Chile       0.60      0.55      0.58       463
      France       0.80      0.89      0.85      2227
     Germany       0.73      0.58      0.64       232
       Italy       0.96      0.95      0.95      1914
    Portugal       0.80      0.54      0.64       579
       Spain       0.69      0.70      0.70       661
          US       0.92      0.96      0.94      5402

    accuracy                           0.86     12459
   macro avg       0.77      0.68      0.72     12459
weighted avg       0.85      0.86      0.85     12459
```

Table 4: LR unigram F1-score



```
              precision    recall  f1-score   support

   Argentina       0.45      0.22      0.29       385
   Australia       0.78      0.23      0.36       251
     Austria       0.79      0.36      0.50       345
       Chile       0.56      0.35      0.43       463
      France       0.76      0.83      0.79      2227
     Germany       0.80      0.37      0.50       232
       Italy       0.95      0.89      0.92      1914
    Portugal       0.71      0.39      0.51       579
       Spain       0.57      0.53      0.55       661
          US       0.81      0.98      0.89      5402

    accuracy                           0.80     12459
   macro avg       0.72      0.51      0.57     12459
weighted avg       0.79      0.80      0.78     12459
```

Table 5: LR bigram F1-score

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 40, 300)	4883700

```
conv1d (Conv1D)                 (None, 40, 128)         115328
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 14, 128)            0
_________________________________________________________________
conv1d_1 (Conv1D)               (None, 14, 64)          24640
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 5, 64)              0
_________________________________________________________________
conv1d_2 (Conv1D)               (None, 5, 32)           6176
_________________________________________________________________
flatten (Flatten)               (None, 160)             0
_________________________________________________________________
dropout (Dropout)               (None, 160)             0
_________________________________________________________________
batch_normalization (BatchNo (None, 160)                640
_________________________________________________________________
dense (Dense)                   (None, 128)             20608
_________________________________________________________________
dropout_1 (Dropout)             (None, 128)             0
_________________________________________________________________
dense_1 (Dense)                 (None, 10)              1290
=================================================================
Total params: 5,052,382
Trainable params: 5,052,062
Non-trainable params: 320
_________________________________________________________________
```

Table 6: CNN model parameters

9. References

[i] Vijayarani, S., Ms J. Ilamathi, and Ms Nithya. "Preprocessing techniques for text mining-an overview." International Journal of Computer Science & Communication Networks 5.1 (2015): 7-16.

[ii] Kumar, Mukesh, and Renu Vig. "Term-frequency inverse-document frequency definition semantic (TIDS) based focused web crawler." International Conference on Computing and Communication Systems. Springer, Berlin, Heidelberg, 2011.

[iii] Goldberg, Yoav. "Neural network methods for natural language processing." Synthesis lectures on human language technologies 10.1 (2017): 1-309.

[iv] Bayaga, Anass. "MULTINOMIAL LOGISTIC REGRESSION: USAGE AND APPLICATION IN RISK ANALYSIS." Journal of applied quantitative methods 5.2 (2010).

[v] Ren, Yazhou, et al. "Robust softmax regression for multi-class classification with self-paced learning." Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017.

[vi] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.

[vii] Goodfellow, Ian, et al. Deep learning. Vol. 1. No. 2. Cambridge: MIT press, 2016.

[viii] Goodfellow, Ian, et al. Deep learning. Vol. 1. No. 2. Cambridge: MIT press, 2016.

[ix] Goodfellow, Ian, et al. Deep learning. Vol. 1. No. 2. Cambridge: MIT press, 2016.