UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

The ChatGPT of wine: to what extent can NLP models be as accurate as empirical models in the context of a wine variety prediction task?

Appendix

Auteur : De Ro Corentin, Verleyen Simon
Promoteur(s) : Vande Kerckhove Corentin
Année académique 2022-2023
Master [120] : Ingénieur de Gestion à finalité spécialisée

Contents

I Python codes

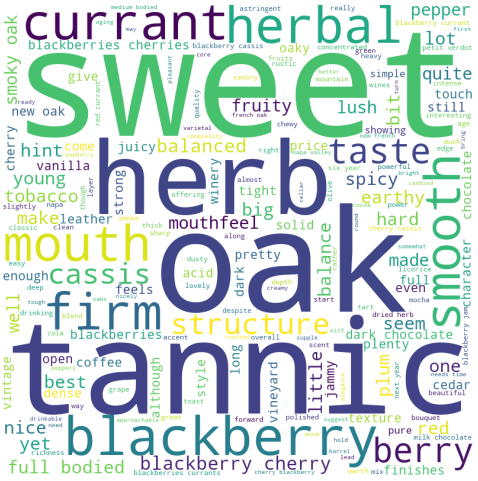
The python codes related to this thesis are all available here. Note that most of them are under the *.ipynb* format. We advice you to download the ZIP of the whole project and then drag and drop it on Jupyter. Then you can run the following code to unzip the folder :

```
import zipfile as zf
files = zf.ZipFile("Master-Thesis-main.zip", 'r')
files.extractall()
files.close()
```

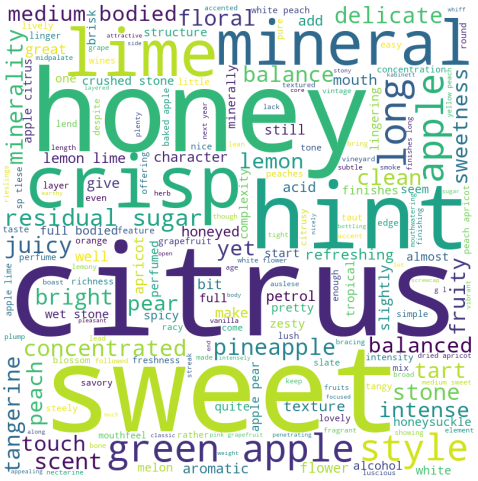
This will unzip the folder and you will have access to all our codes. A small *README.md* file is available as well. Also, please do not rerun the BERT models, it takes almost 36 hours to run without any GPU. These models are already uploaded with their results so no need to rerun them.

II List of Additionnal Figures

II.I Cleaned Wordclouds



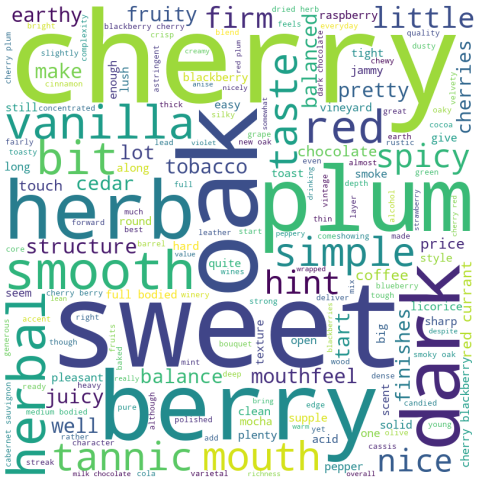
(a) Cabernet



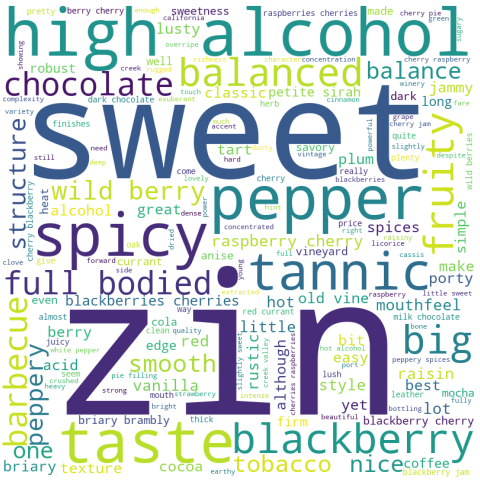
(b) Riesling

(c) Sangiovese

(d) Malbec

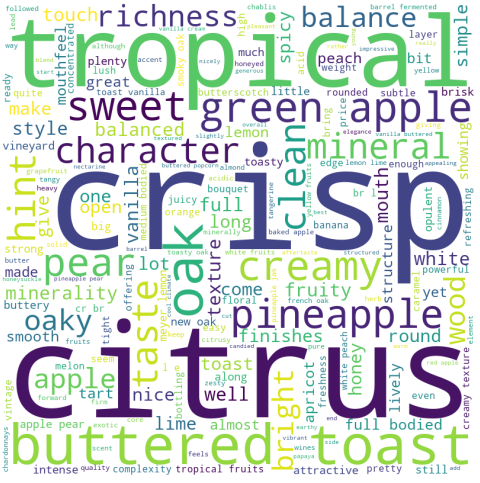


(e) Merlot



(f) Zinfandel

(g) Pinot Noir



(h) Chardonnay

Figure 1: Wordclouds by variety

II.II Power of wines by variety

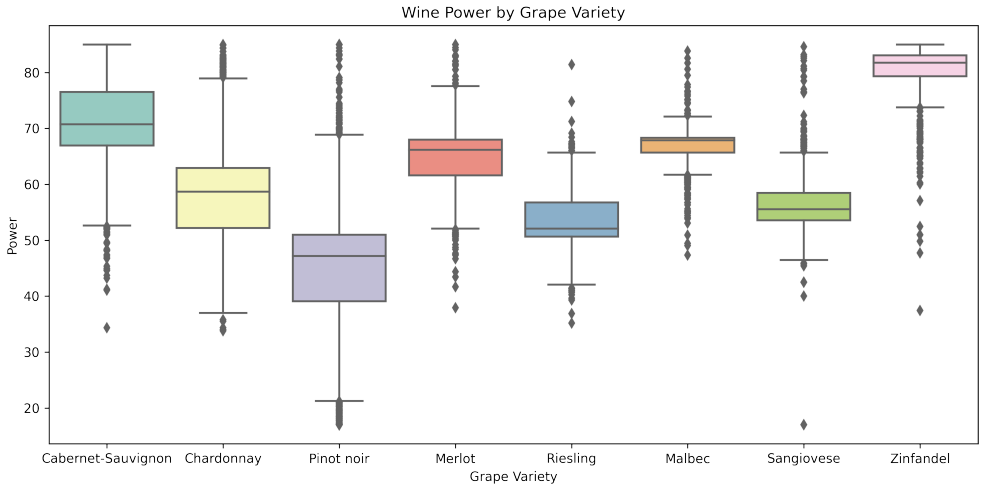


Figure 2: Power of wines by variety

II.III Tannicity of wines by variety

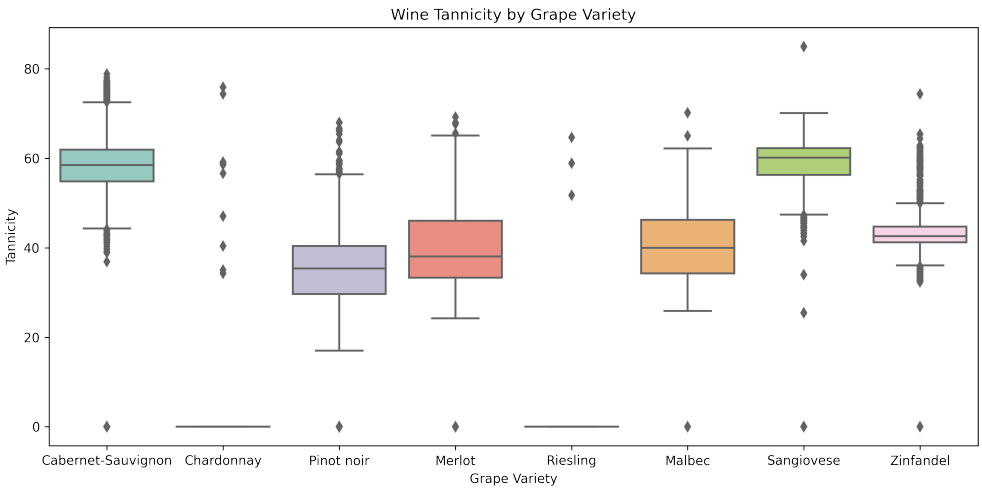


Figure 3: Tannicity of wines by variety

II.IV Sweetness of wines by variety

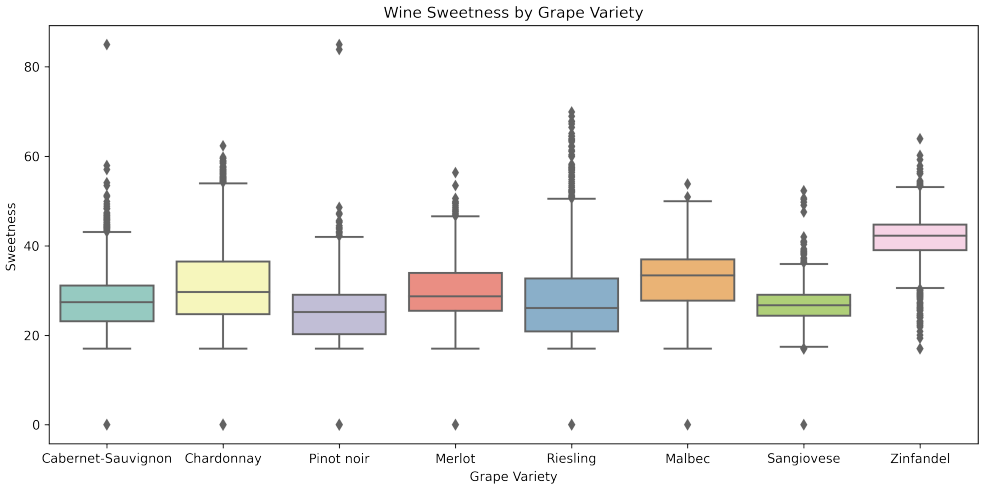


Figure 4: Sweetness of wines by variety

II.V Acidity of wines by variety

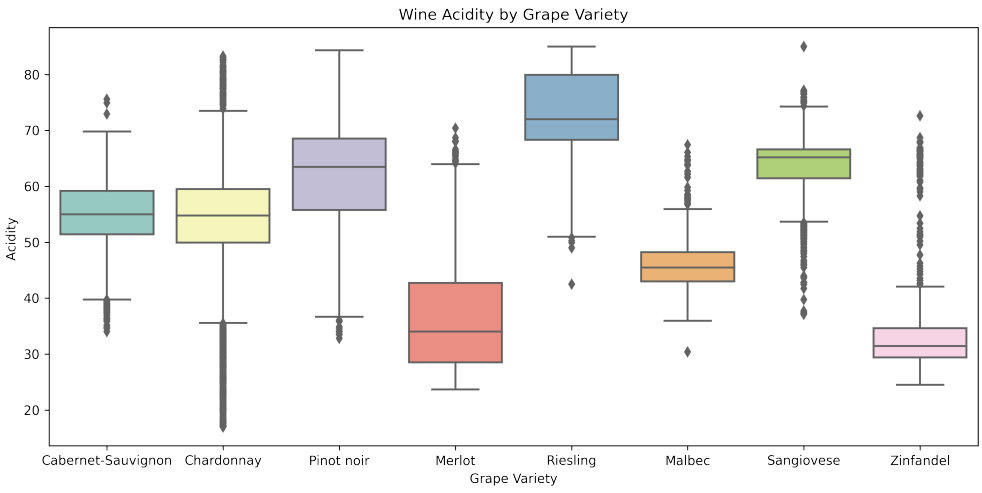


Figure 5: Acidity of wines by variety

II.VI Pairplot

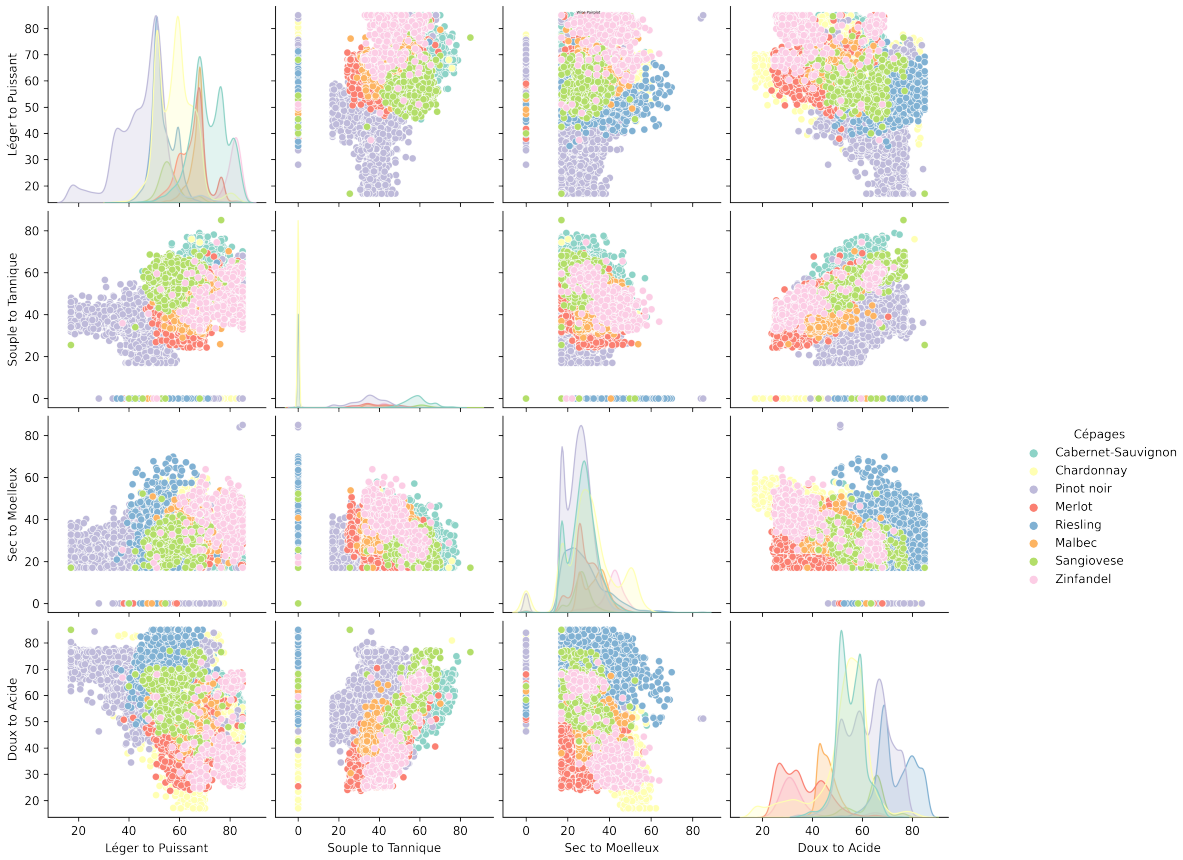


Figure 6: Pairplot

II.VII Power density

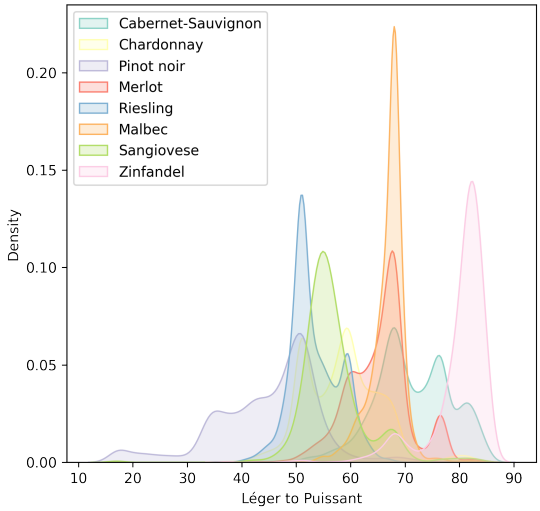


Figure 7: Power density by variety

II.VIII Densities

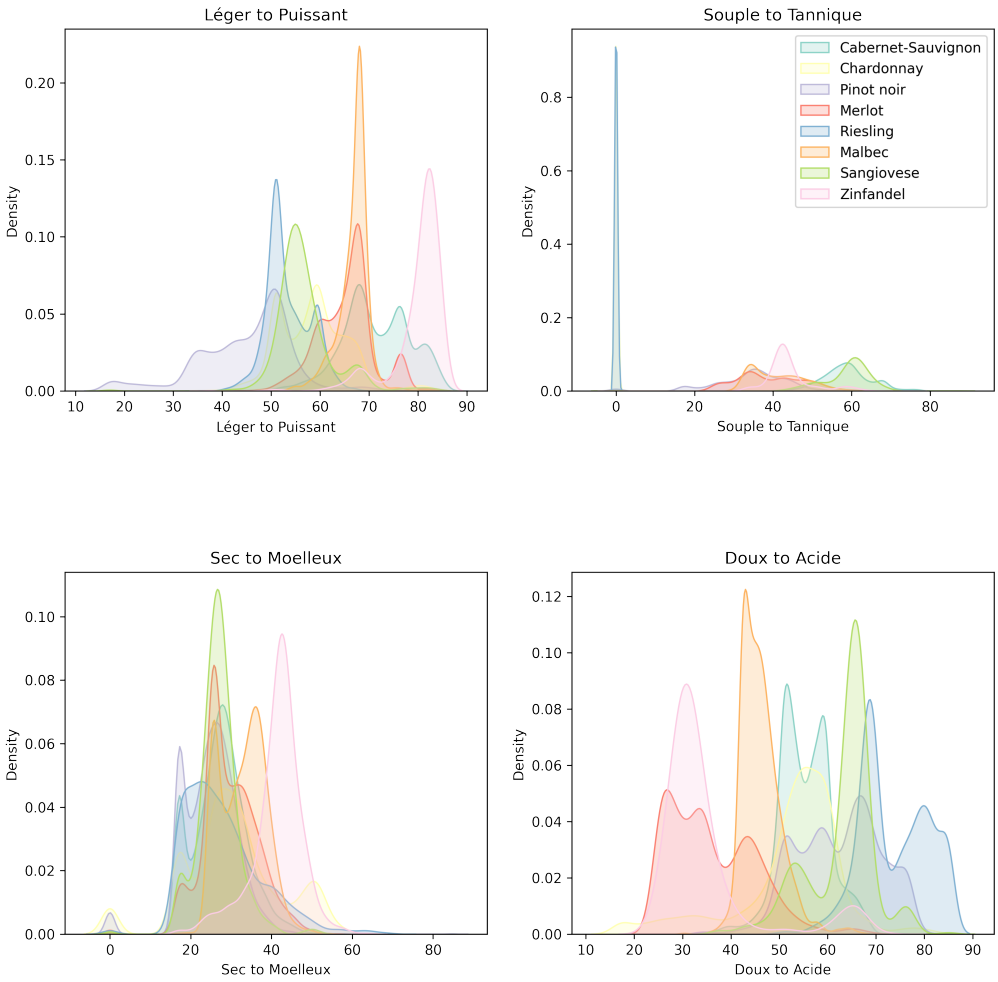
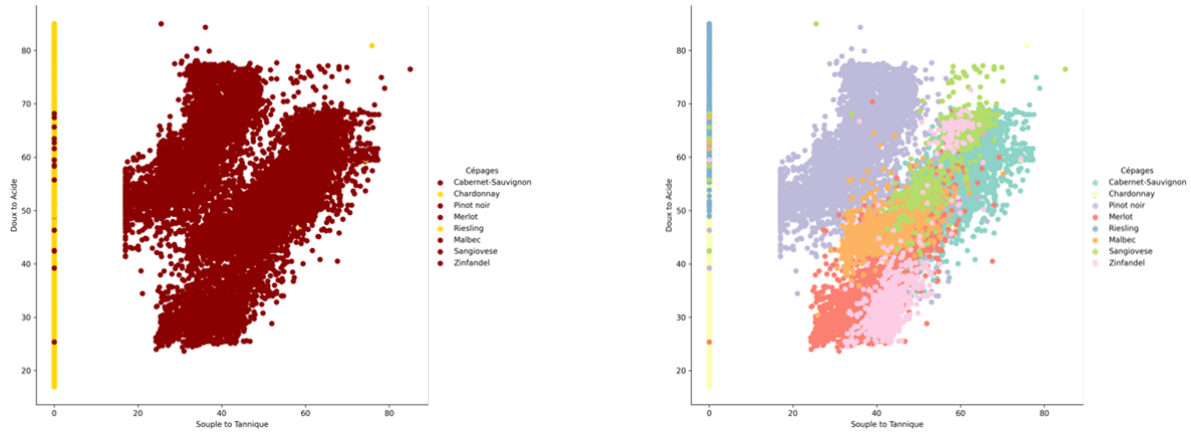


Figure 8: Densities by variety

II.IX Tannicity and Acidity of wines by variety



(a) Red and White wines based on their tannicity and acidity

(b) All varieties wines based on their tannicity and acidity

Figure 9: Tannicity and Acidity of wines by variety

II.X Most represented aromas by variety

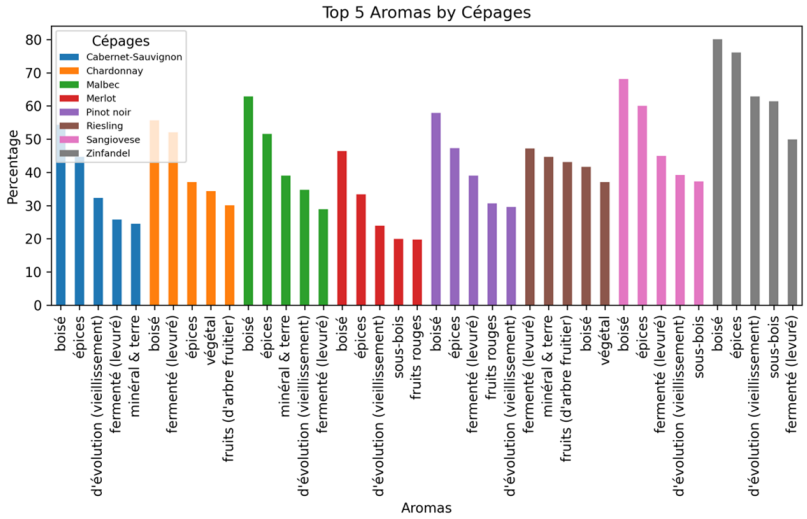


Figure 10: Most represented aromas by variety

II.XI Chi-Square Test for Feature Selection

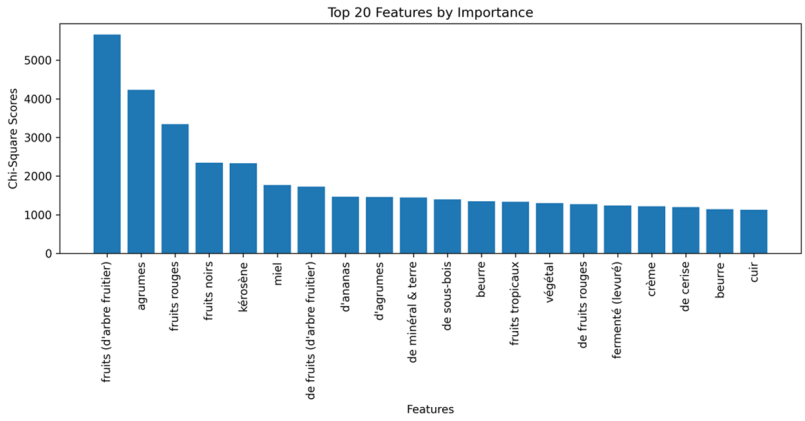


Figure 11: Most Important Aromas

II.XII Percentage of fruits aroma by variety

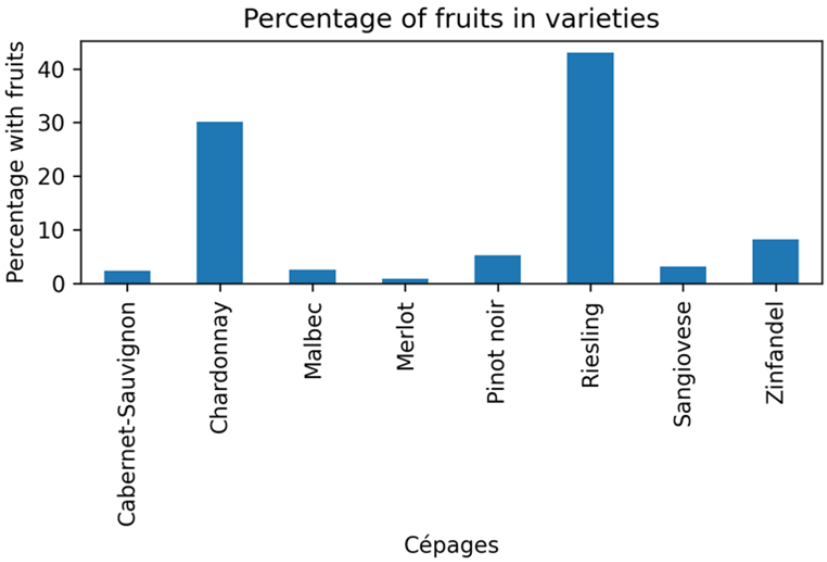


Figure 12: Percentage of fruits aroma by variety

II.XIII Chi-Square Test for Feature Selection in Red Wines

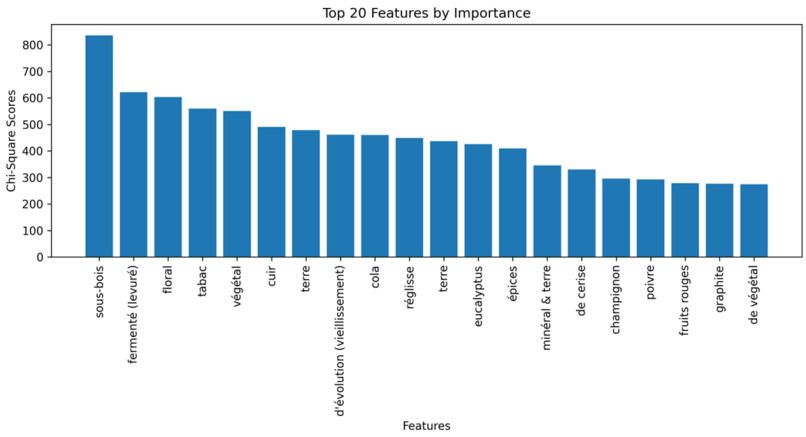


Figure 13: Top 20 aromas in red wines

II.XIV Percentage of sous-bois aroma in red wine varieties

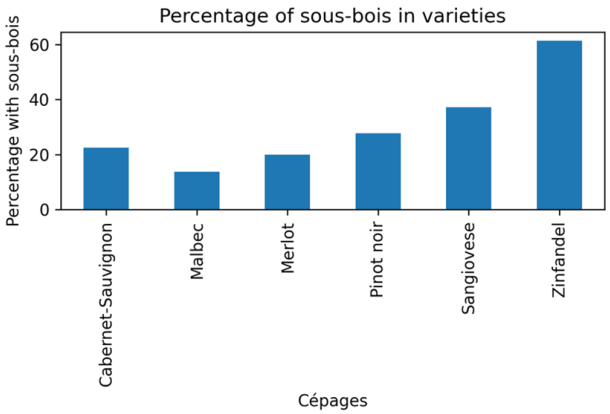


Figure 14: Percentage of sous-bois aroma in red wine varieties

II.XV Most Represented aromas specific to a single variety

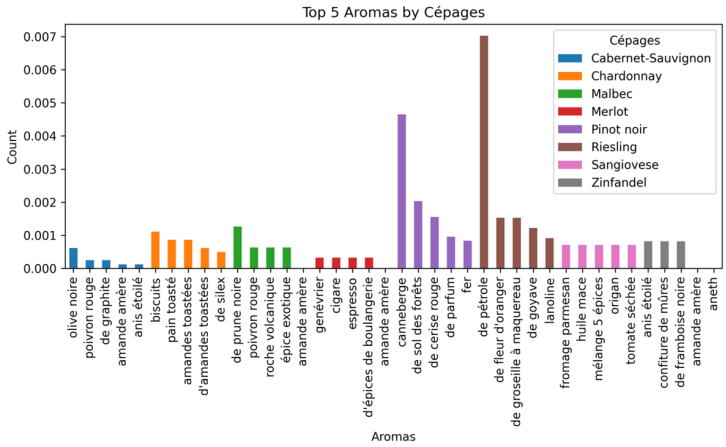


Figure 15: Most Represented aromas specific to a single variety

II.XVI Most Represented aromas specific to a maximum of 2 varieties

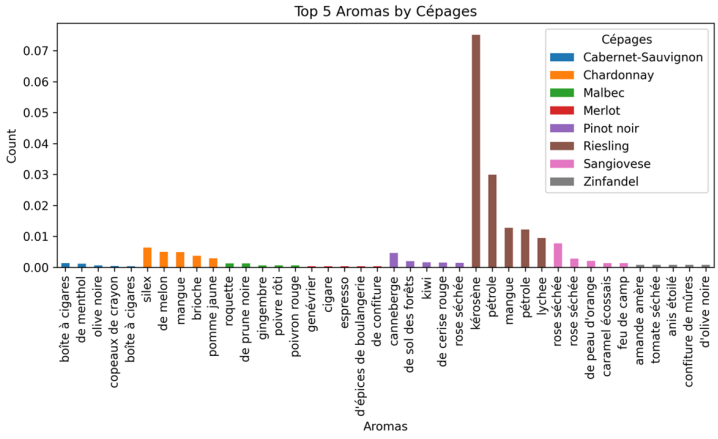


Figure 16: Most Represented aromas specific to a maximum of 2 varieties

III Google Forms

The Google Forms that we sent to some experts is available here.

IV Models

IV.I BERT Model

Was the review sufficient to predict the variety of a wine? This is the question we will be able to answer in this subsection. After testing different models, we discovered that the Bert-base uncased was the best one with impressive accuracy. Here is the classification report.

	Precision	Recall	F1-Score	Support
Cabernet-Sauvignon	0.79	0.87	0.83	440
Chardonnay	0.95	0.98	0.96	569
Malbec	0.73	0.69	0.71	118
Merlot	0.74	0.63	0.68	216
Pinot Noir	0.86	0.93	0.89	561
Riesling	0.97	0.91	0.94	294
Sangiovese	0.94	0.64	0.76	103
Zinfandel	0.83	0.71	0.77	94
Accuracy			0.87	2395
Macro Avg	0.85	0.79	0.82	2395
Weighted Avg	0.87	0.87	0.86	2395

Table 1: BERT Classification Report

At the end of the day, we reached an accuracy of 87%. At first glance, it looks promising, the overall performance is quite surprising. It also shows that new NLP models can show really incredible results.

We can also notice that, according to this report, our models perform differently according to the varieties. The recall indicates that our model captures a larger proportion

of the relevant results, regardless of the inclusion of some irrelevant ones. On the other hand, precision reflects the model's ability to provide more pertinent results compared to impertinent ones.

Also, when looking at the f1-score, we can observe that the unbalanced dataset has an impact on our results. Indeed, the f1-score is less good for the varieties with the smaller number of wines. This is because it considers both precision and recall, allowing it to capture the effects of imbalances on both metrics. That also explains why the weighted average is always higher than the average for every metric.

Here is the confusion matrix of the same model. It is interesting as we can compute some statistics according to specific varieties. For example, we can say that, 95.25% of the white wines were correctly predicted, meaning that it is quite easy for BERT to detect if it is a white or a red wine. Also, among the 4.75% that is not predicted correctly, we can see that 78% are still classified as white wines. That means that overall, 98.95% of the time we can classify a wine as white when it is effectively white.

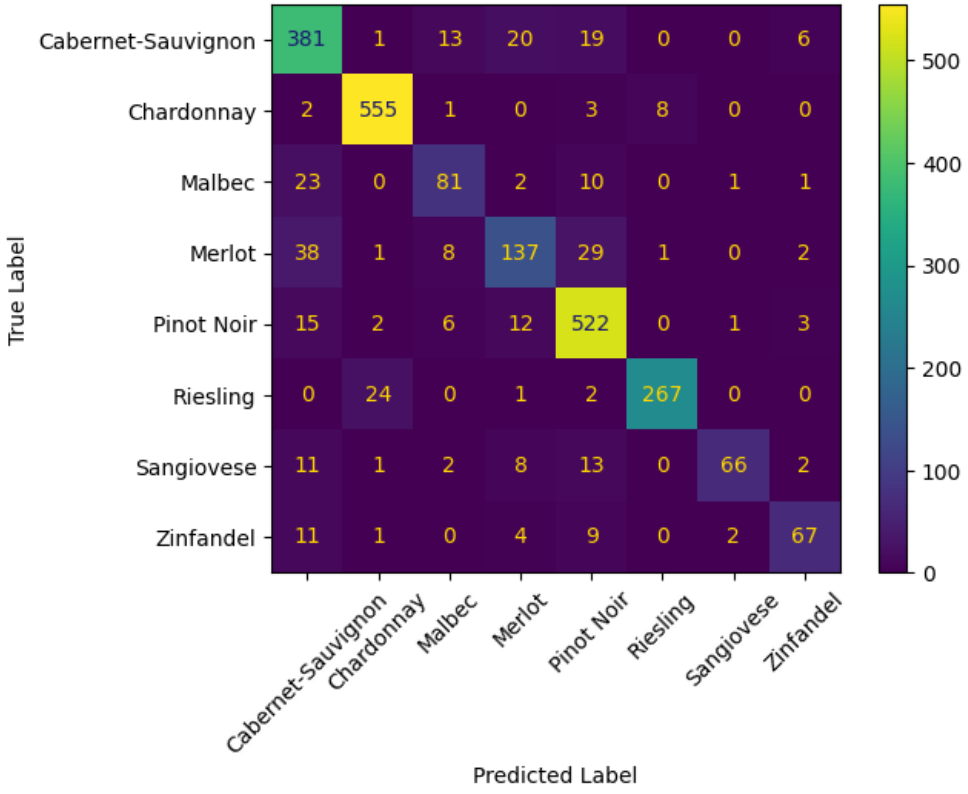
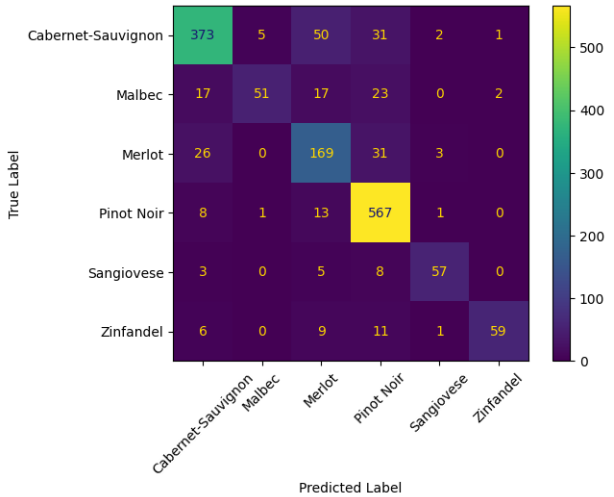


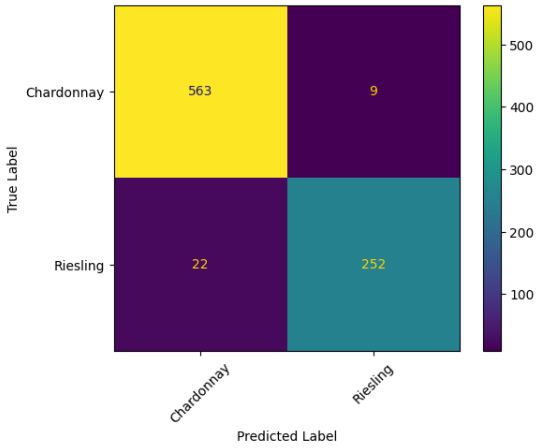
Figure 17: Confusion Matrix

We can do the same for the red wines, but we can suppose it to be a bit more complex as there are 6 varieties of red wines. Nevertheless, only 0.45% of the red wines were classified as white meaning that there is really a strong difference between red and white within the descriptions. Based on these results, we asked ourselves a question. Is there a difference in the accuracy of our model if we keep only red or white wines. To go as deeply as possible in this reflection, we also created a model on red wines only one one

hand, and on white wines only on the other hand. There were then two trainsets, two validation sets and two test sets. Here are the confusion matrixes.



(a) Figure 1



(b) Figure 2

Figure 18: Confusion Matrix by color

Now, on the right screenshot, we can see that only 3.66% of the wines are not predicted correctly. That means that we have slighty improved the results (it was 4.75%) within the category of white wines. Bert has then learnt only on white wines and understand them better as well. Nevertheless, it is still interesting to observe that we did not improve a lot neither. The addition of red wines do not impact that much the whites ones. For bert, it is not considered as noise to classify the white wines. In conclusion, BERT and its accuracy offered us a really good baseline for the next numerical models. We now expect all of them to be at least 87% accuracy.

IV.II SVM Results

SVM will be the first model tested. Of course, we expect this model to perform better than the NLP one thanks to the use of structured data. According to the analysis that we did in the Datasets Analysis section, it seems that SVM could reach great performance. The question mark of this model relies in its ability to handle the imbalanced dataset. We know that SVM could show some weaknesses when facing strong imbalance. However, it could be compensated for by its ability to handle lots of variables.

Also, for this model, we used the Gridsearch function to optimize the best hyperparameters. However, this pre-implemented function already creates a validation set behind the scene. Consequently, we do not have to keep our 10% validation set anymore. In the end, we will then train on 90% of the data and test on the last 10% totally unknown. This is the equivalent of the 80/10/10 with the epoch's parameter. Here is the classification report.

	Precision	Recall	F1-Score	Support
Cabernet-Sauvignon	0.86	0.94	0.90	427
Chardonnay	0.95	0.98	0.96	590
Malbec	0.86	0.87	0.87	110
Merlot	0.90	0.79	0.84	228
Pinot Noir	0.99	0.95	0.97	597
Riesling	0.96	0.95	0.96	275
Sangiovese	0.86	0.83	0.84	75
Zinfandel	0.93	0.89	0.91	93
Accuracy			0.93	2395
Macro Avg	0.91	0.90	0.91	2395
Weighted Avg	0.93	0.93	0.93	2395

Table 2: SVM Classification Report

As expected, SVM overtook BERT with an accuracy of 93%. At first glance, this model seems to be a bit less impacted by the imbalanced dataset. Indeed, the gap between well represented and less represented classes is lower than in BERT. Here is how the confusion matrix looks like.

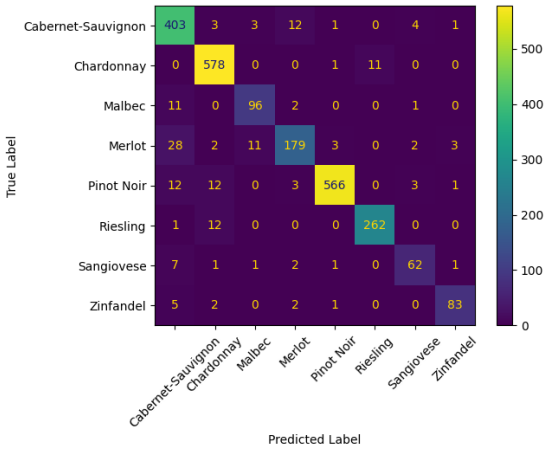
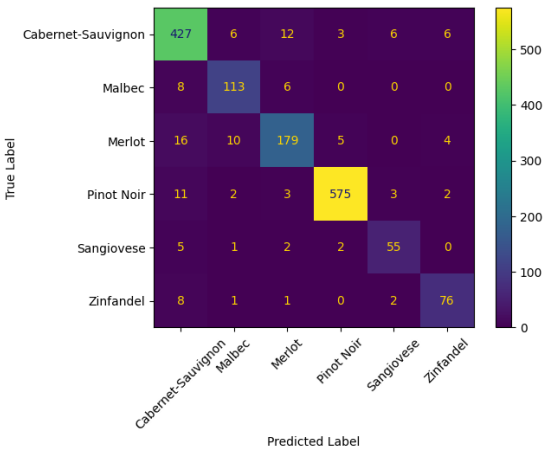
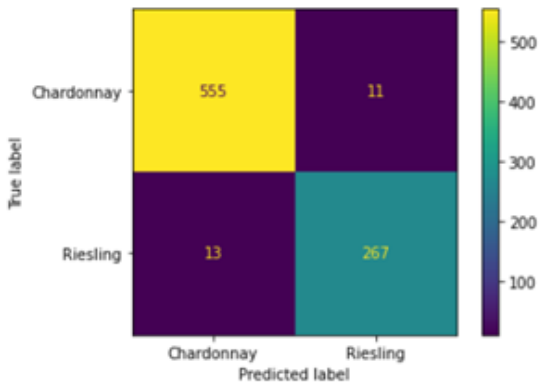


Figure 19: Confusion Matrix

In this model, a bit more than 97% of the white wines were correctly classified. Among the 3% that were not correct, only 8% were classified in a red wine category. At the end of the day, 99.8% of the white wines were recognized as white, which is almost perfect. On the other hand, 90% of the red wines were correctly predicted. This time, among the errors, it is around 13% that was classified as white wines. We can then deduct a strong capacity to detect the color of the wine. From this analysis, let's separate the red and white wines to analyze them separately and see how the model behaves.



(a) Figure 1



(b) Figure 2

Figure 20: Confusion Matrix by color

As far as the white wines are concerned, we retrieve the exact same probability (97%),

meaning that even though the model was specialized on white wines, it did not improve itself to correct the mistakes. However, for the red wines, we went up to almost 92% compared to 90% of the mixed model. That means that the model really improved when training only on red wines. This improvement could be explained by many reasons: imbalanced dataset, class overlapping or maybe just the complexity of the model.

To investigate a bit more our hypothesis, we also decided to test our model without our 635 aromas. We indeed found them not especially interesting and useful in the Boolean analysis. There was in the end only a few words that were useful to separate classes. It was rare to find a word that belonged to one and only one variety. Here is the classification report without all this potential useless information.

	Precision	Recall	F1-Score	Support
Cabernet-Sauvignon	0.86	0.85	0.86	427
Chardonnay	0.95	0.93	0.96	590
Malbec	0.77	0.68	0.72	110
Merlot	0.85	0.72	0.78	228
Pinot Noir	0.81	0.96	0.88	597
Riesling	0.92	0.88	0.90	275
Sangiovese	0.86	0.64	0.73	75
Zinfandel	0.95	0.74	0.83	93
Accuracy			0.87	2395
Macro Avg	0.87	0.80	0.83	2395
Weighted Avg	0.87	0.87	0.87	2395

Table 3: SVM without aromas Classification Report

We lost 6% of accuracy, meaning that these aromas were in the background quite useful to distinguish our varieties. Even though it was not obvious, they are still useful in our prediction task. Without them, we have exactly the same accuracy than the NLP model by the way. Diverse methods such as Tf-Idf could have been used in the dataset analysis to explore more the capacity of a word to discriminate varieties.

IV.III Random Forest Results

Another model that could perform well in our specific case is the random forest. As for support vector machine, we decided to use the pre-implemented GridSearch function. Hence, it will directly produce the best model possible. The goal is to know if it will perform better than SVM to pursue our analysis. Here is its classification report when running with a 5-folds method :

	Precision	Recall	F1-Score	Support
Cabernet-Sauvignon	0.87	0.92	0.89	460
Chardonnay	0.96	0.98	0.97	558
Malbec	0.79	0.73	0.76	120
Merlot	0.85	0.82	0.83	227
Pinot Noir	0.97	0.96	0.96	589
Riesling	0.97	0.97	0.97	278
Sangiovese	0.89	0.65	0.75	75
Zinfandel	0.92	0.91	0.91	88
Accuracy			0.92	2395
Macro Avg	0.90	0.87	0.88	2395
Weighted Avg	0.92	0.92	0.92	2395

Table 4: RFs Classification Report

The global accuracy is slightly lower than SVM (only 1 %) but it overall worked pretty well as well. As we want to compare both best models in NLP and in ML and we are interested in the loss of accuracy more than boosting our models, we will not investigate further the subdivision between red and white. This model is then put aside.

IV.IV Artificial Neural Networks Results

This model is extremely computationally intensive and time-consuming. After running *GridSearch* with a 5-folds method this is our classification report:

With a 92% of global accuracy, we can say that our model worked pretty well. Comparing to SVM, the f1-score for the less represented varieties went down quite a bit. It seems that this model has been more impacted by the imbalance of our dataset. Moreover,

	Precision	Recall	F1-Score	Support
Cabernet-Sauvignon	0.88	0.90	0.91	439
Chardonnay	0.96	0.97	0.97	595
Malbec	0.82	0.76	0.79	127
Merlot	0.86	0.85	0.86	233
Pinot Noir	0.96	0.93	0.94	580
Riesling	0.93	0.96	0.94	258
Sangiovese	0.75	0.78	0.76	80
Zinfandel	0.86	0.86	0.86	83
Accuracy			0.92	2395
Macro Avg	0.88	0.88	0.88	2395
Weighted Avg	0.92	0.92	0.92	2395

Table 5: ANN Classification Report

the pace of this model is really slow as well comparing to the support vector machine. Overall, this model has nevertheless reached great performance. We will not detail the same analysis as for SVM as the goal is not to compare our different models but to investigate the loss of accuracy when switching from structured to unstructured data.

IV.V ChatGPT

IV.V.1 F1-score

	F1-Score
Cabernet-Sauvignon	0.58
Chardonnay	0.86
Malbec	0.25
Merlot	0.27
Pinot Noir	0.67
Riesling	0.88
Sangiovese	0.54
Zinfandel	0.66

IV.V.2 Comparison with other models

	BERT	SVM	ANN	RF	ChatGPT
Accuracy	0.87	0.93	0.92	0.92	0.62

IV.V.3 ChatGPT Confusion Matrix

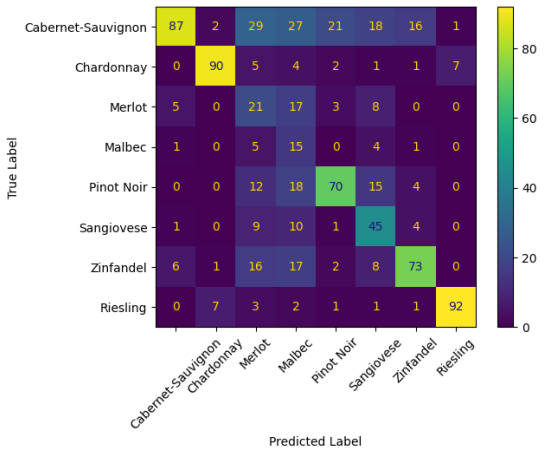


Figure 21: Confusion Matrix