

**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?**

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

## Abstract :

In this increasingly digitalized world, many innovative AI tools are emerging. Fueled by data, they arouse interest in fields like Natural Language Processing and Machine Learning. Although these two disciplines have made significant progress on their own, we have yet to witness real synergies between them. Moreover, bridging these disciplines makes sense since leveraging human language could help fasten the adoption of AI technology by making it more accessible. This work aims at better understanding the gap between human language and empirical data.

The research methodology involves two datasets: one with quantitative properties of wines and the other with textual reviews. The quantitative data include factors like acidity, sweetness, tannicity, power, and aromas. Textual reviews provide subjective descriptions of taste and aromas. To predict the wine variety, Machine Learning algorithms and NLP models are trained and evaluated on 8 different possibilities of white and red varieties: Cabernet-Sauvignon, Chardonnay, Merlot, Malbec, Pinot Noir, Riesling, Sangiovese & Zinfandel.

According to the findings, BERT model is promising in predicting wine variety but has poorer accuracy when compared to empirical models. However, the limits and constraints of NLP models are well offset by the user-friendliness and speed that this model provides to wine lovers. The thesis also emphasizes the significance of training NLP models on specific datasets, as this has a major impact on their ability to discover patterns related to wine data. The comparison of the BERT model with the ChatGPT model clearly shows that the latter's results suffer greatly from the lack of task-specific training.

Extending our findings into less lexically rich sectors and investigating hybrid models to improve accuracy and resilience could be future research objectives that will contribute to existing studies on the computerization of taste.

In conclusion, this master's thesis offers useful insights into the accuracy of NLP models in the wine variety prediction challenge. It emphasizes the possibilities and limitations of NLP models in comparison to empirical models, underlying the importance of additional improvements and domain-specific training. The findings add to ongoing research on the computerization of taste and inspire future investigation regarding to the possibilities of NLP models in wine and related fields.

## Acknowledgements

*First of all, we would like to especially thank our supervisor, Professor C. Vande Kerckhove. His expertise, his availability and his numerous constructive feedbacks have greatly contributed to the development of this thesis.*

*Secondly, we would also like to thank all the people who were involved in this paper. Their reviews, advices and opinions have been greatly appreciated throughout this project.*

*Finally, we would like to thank our respective families for their moral support throughout this work but also throughout our 5 years of study.*

# Table of Contents

|          |                                                |           |
|----------|------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                            | <b>5</b>  |
| <b>2</b> | <b>State of the Art</b>                        | <b>7</b>  |
| 2.1      | Literature review . . . . .                    | 7         |
| 2.2      | Research Question . . . . .                    | 29        |
| 2.3      | Business Incentives . . . . .                  | 31        |
| <b>3</b> | <b>Data and Methods</b>                        | <b>34</b> |
| 3.1      | Datasets . . . . .                             | 34        |
| 3.2      | Datasets Analysis . . . . .                    | 40        |
| 3.3      | Methods . . . . .                              | 46        |
| <b>4</b> | <b>Experimental Results</b>                    | <b>51</b> |
| 4.1      | Comparison of the different models . . . . .   | 51        |
| 4.2      | Comparison with ChatGPT . . . . .              | 59        |
| <b>5</b> | <b>Evaluation of our models</b>                | <b>64</b> |
| <b>6</b> | <b>Conclusion, limitations and future work</b> | <b>68</b> |
| 6.1      | Conclusion . . . . .                           | 68        |
| 6.2      | Limitations . . . . .                          | 70        |
| 6.3      | Future work . . . . .                          | 73        |
| <b>7</b> | <b>References</b>                              | <b>75</b> |
| <b>8</b> | <b>Appendix</b>                                | <b>86</b> |

## List of Figures

|    |                                                   |    |
|----|---------------------------------------------------|----|
| 1  | Genetic distances and dendrogram ([74]) . . . . . | 27 |
| 2  | NLP Dataset layout . . . . .                      | 34 |
| 3  | Wine varieties proportions . . . . .              | 36 |
| 4  | Quantitative taste structure of a wine . . . . .  | 37 |
| 5  | Aromas of the wine . . . . .                      | 37 |
| 6  | Technical Details of the wine . . . . .           | 37 |
| 7  | ML Dataset Layout . . . . .                       | 38 |
| 8  | White & Red wines proportions . . . . .           | 39 |
| 9  | Wordclouds by variety . . . . .                   | 41 |
| 10 | SVM & BERT Confusion Matrices . . . . .           | 54 |
| 11 | Red & White Confusion Matrices . . . . .          | 56 |
| 12 | White wines confusion matrices . . . . .          | 56 |
| 13 | Confusion Matrix of ChatGPT . . . . .             | 60 |
| 14 | Dendrograms . . . . .                             | 65 |

## List of Tables

|    |                                                                                                                      |    |
|----|----------------------------------------------------------------------------------------------------------------------|----|
| 1  | Varieties proportions . . . . .                                                                                      | 36 |
| 2  | Comparison of performance between the quantitative models . . . . .                                                  | 51 |
| 3  | Comparison of performance between the NLP models . . . . .                                                           | 52 |
| 4  | Comparison between the two best-performing models . . . . .                                                          | 53 |
| 5  | SVM Classification Report . . . . .                                                                                  | 53 |
| 6  | BERT Classification Report . . . . .                                                                                 | 53 |
| 7  | BERT Classification Report detailed . . . . .                                                                        | 55 |
| 8  | SVM Classification Report detailed . . . . .                                                                         | 55 |
| 9  | Comparaison between the models predicting all the varieties and the models<br>predicting only the red ones . . . . . | 57 |
| 10 | F1-Score by variety of ChatGPT . . . . .                                                                             | 59 |
| 11 | Comparison of Accuracy between ChatGPT and other models . . . . .                                                    | 60 |

|    |                                             |    |
|----|---------------------------------------------|----|
| 12 | Major mistakes comparison . . . . .         | 64 |
| 13 | Small mistakes comparison . . . . .         | 66 |
| 14 | Distances matrix based on experts . . . . . | 67 |
| 15 | Number of outliers per variety . . . . .    | 70 |

# 1 Introduction

The Internet has undergone significant growth since its creation, with some experts even describing this growth as exponential. This has led to the World-Wide-Web becoming a huge network for exchanging a wealth information in complex ways (Huberman & Adamic ,1999) [44]. As a result, there has been a steady increase in the amount of data available, making it an interesting subject for study (A. Seydtaghia ,2020) [77]. In response to this digital expansion, various disciplines such as Data Science, Machine Learning and Recommender Systems have experienced significant developments as they all involve working with and manipulating data (Sarker, 2021)[75]. Over time, these areas of study have become increasingly integrated into the business world as well. Nowadays, they often play a crucial role in generating value and decision-making in certain organizations (Vidgen, 2017) [87]. Business Analytics, defined as the data science for business problems, will be a key discipline of this master’s thesis. (Walter R. Paczkowski, 2021) [65].

To narrow down our scope of research, we are going to focus on a fast-growing field in the viticultural world called Wineinformatics. It is a new data-driven application with focus on understanding wine through artificial intelligence (B. Chen, 2022) [17]. In this paper, artificial intelligence will therefore play an important role. The history of AI spans over 70 years of research, development, and innovation to create machines that can perform tasks that typically require human intelligence, such as learning, reasoning, perception, and natural language understanding. This thesis aims at better understanding of the winery world at the hands of lexicon computerization and data exploitation. Wine will consequently be the subject of all our tests.

For thousands of years, wine has been the most consumed beverage by humans (Oliveira, 2021) [62]. Now ranked second behind beer, this highly diversified beverage (P. This & al., 2006) [83] has given birth to new jobs such as sommelier. This man of the art of taste discusses, selects, appraises, buys, preserves, resells, elaborates, recommends, adjusts his advice, tastes the wines to offer the best experience to his wine-loving peers (S. Olivesi, 2015) [63].

This work aims to enhance the computerization of taste description using textual data, with all its challenges: how will machines handle human language?

## 2 State of the Art

In this section, we will review the existing literature. This step represents an important part of our work since this thesis is at the confluence of several fields. We will cover a wide range of topics in the first instance to narrow down our interests in the second instance. In this exploration phase, we will proceed from a general context description that will lead, down the line, to a definition of our research question. In the course of the topics, we will therefore evolve through the progress and challenges of various disciplines linked in one way or another. To structure this state of the art, we will first outline the evolution leading to today's automated and digitalized world. After years of research and developments, efficient algorithms are now capable of interpreting human language. The latter turns out to be extremely rich in some specific domains, including wine, therefore making the crossing with NLP exciting. Overall, progress, challenges and opportunities will be addressed to make you arouse interest in our research question.

### 2.1 Literature review

In today's fast-paced and automated world, algorithms keep expanding. Their extreme diversity can even lead them to take many different forms, from fraud detection in finance (P. Prudhvi, 2021) [66] to a better management in agriculture ( D. Raj, 2022) [71]. Their scope is extremely wide. Algorithms, defined as the sequences of successive moves providing an output based on defined input (K. Kwon, 2021) [48], are in this respect considered as the fundamentals of this automated and digitized world. Indeed, for decades, humans have been automating their working place. Automation or more accurately, the use of machines and computers that can operate without any human interference (Cambridge University) [6], is nowadays an essential aspect in all kind of business, including SME's (Chavez, 2021) [16]. Indeed, benefits flowing from automation are extremely valuable for companies : improved accuracy, reliability and productivity (S. Dunlap, 2023)[29]. This progress has consequently revolutionized industries leading to the well known Industry 4.0, more scalable.

More recently, this automation has started to integrate Artificial Intelligence as well, making the range of opportunities even wider. Intelligent Automation was born. As defined by IBM [92], Intelligent Automation is “*the use of automation technologies – artificial intelligence (AI), business process management (BPM), and robotic process automation (RPA) – to streamline and scale decision-making across organizations.*” There are, as a result, clear benefits for companies to invest in these technologies.

As previously explained one of the component of the Intelligent Automation, is Artificial Intelligence. It is defined as “*a branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence.*” (Manaware, 2020) [55]. This has become hot topic today thanks to the widespread use of the tool ChatGPT. Before going into details about what artificial intelligence is, let’s take a brief look at its complex and consistently evolving history.

The term “artificial intelligence” was first used by John McCarthy in 1956. However, the roots of AI could be traced back to Turing’s work in 1950. He was the first to study the ability of machines to think and whether it was possible for them to differentiate humans from machines (J. Curry, 2022)[26].

Thirty years later, AI really began to arouse interest at the hands of a system capable of playing high-level chess games. Since then, AI has not stopped growing and has developed different subfields such as Machine Learning and Natural Language Processing (J. Curry, 2022)[26].

Over time, AI has considerably expanded its range of activities. Nevertheless, some common characteristics have remained over years such as its ability to solve complex problems, make predictions or even undertake tasks that would have needed human intervention. Overall, its history is mainly characterized by constant innovation and development, both leading to make machines as human as possible (J. Curry, 2022)[26].

Within the landscape of AI, two particular subfields have emerged as powerful tools: Machine Learning and Natural Language Processing. Let's delve into these two topics.

Defined as a subfield of artificial intelligence that focuses on developing algorithms and models that enable computers to learn from and make predictions or decisions based on data, Machine Learning has recently become very popular with both professionals and individuals (Alexander L. Fradkov, 2020) [32]. This widely accepted definition highlights the common base of all subfields in AI, data. Data are an exponentially growing gold mine for all data-driven disciplines (Huberman, 1999) [44].

The 21st century is considered as key moment in the development of machine learning. Three major trends are considered to have contributed to its real boom: Big Data, multithreading and new deep machine learning algorithms (Alexander L. Fradkov, 2020) [32]. Moreover, its concrete applications were particularly appealing to us. Recommender systems, demand and stock levels predictions, customer classification, all these tasks that could help develop a business more effectively. Various investigations of this kind have of course already been carried out, allowing us to understand that ML tools such as stacking integration can improve recommendations (W. Mu, 2022) [58] among numerous examples.

Over the years, machine learning has proven to be useful for three main tasks which are classification, regression, and clustering (K. Jolly, 2018) [46]. In this work, we will focus on the classification task. The goal of classification task is to build a model capable of classifying new observations into particular classes, as humans could do.

In this paper, the classification task will take place in a supervised environment. Therefore, the model will know the target variable for the set of observations on which the machine trains.

To accomplish such classification tasks, variety of categorization models were created including, among others, Artificial Neural Networks (ANN), Random Forests, and Support Vector Machines (SVM). Each of these models has advantages and drawbacks that make them appropriate for various datasets and applications. It is clear that, these three models are not the only ones in machine learning but they turned out to be efficient in many cases. Reviewing them would already give some insights about the functionnings of these commonly shared models.

Other models have been left out, such as Decision Tree model because some studies (Pranckevicius & al, 2017) [69] show that Random Forests perform better. Analyzed more closely, a tree model is what composes a Random Forest. There is consequently no need to analyze both of them. Similarly, a model as Naïve Bayes has also not been considered. Andrew Y. Ng (2001) [61] showed that Naïves Bayes is quickly caught up by other models. Its assumption stating that all features should be independent is hardly respected in real life applications.

Let's get deeper into the details of each of these three models now. Firstly, to pursue on the interaction human-machine, Artificial Neural Networks have been developed in the 50's. These mathematical models have been inspired by the human brain. The first ANN, called Perceptron, was actually a mathematical representation of human neurons (Fitch, Warren, McCulloch & Pitts, 1944) [56] . Nevertheless, this first promising model rapidly faced difficulties and interest began to wane (Werbos, 1974) [91].

In the 80's, backpropagation helped rekindle interest in ANNs. This learning technique for multi-layer neural networks had the potential to get over the Perceptron's drawbacks. The Backpropagation Network, Radial Basis Function Network and Hopfield Network were, among others, successful ANNs that resulted from this algorithm (Broomhead & Lowe, 1988) [11]. In summary, this algorithm was used to boost models thanks to some mathematical optimizations techniques, using notably forward and back propagation (Rumelhart & al., 1988b) [72].

During the 1990s and 2000s, ANN's were used for a variety of tasks, such as speech recognition, picture processing, and natural language processing (LeCun, Bottou, Bengio, et al., 1998) [50]. Their numerous layers allowed them to learn complex patterns and representations from data.

A second turning point arised when deep learning was developed in the mid 2000s (LeCun et al., 2015) [49]. This discipline has facilitated the development of autonomous vehicles, virtual assistants, and a lot of other AI applications. Their combination opened up then many new doors and opportunities.

One of ANNs' key benefits is their capacity to learn from noisy input and handle variety of data types, including categorical and continuous variables. In tasks like image and speech recognition, ANNs excel at capturing non-linear relationships between features (Heaton, 2017b) [38]. However, the development and training of ANNs, especially deep networks with multi-layers networks, can be extremely computationally expensive. Additionally, ANNs may be vulnerable to overfitting if the training data is insufficient or the network architecture is overly complex.

Despite these challenges, ANNs are widely used in different fields. Researchers and practitioners are continually exploring innovative techniques like regularization and dropout to improve performance and mitigate overfitting. ANNs are still expected to play a significant role in machine learning and artificial intelligence applications, driving innovation and progress across industries.

Secondly, Random Forest is a type of ensemble learning method. It combines the results of single models to improve the accuracy and stability of predictions. For Random Forest, each single model is a decision tree. It is a supervised ML technique used to categorize new observations based on binary conditions on specific features. Ho first suggested the basic concept of mixing numerous trees for classification and regression in 1995 (Ho, 1995) [42]. Later, Breiman would be the first to name it “Random Forest” (Breiman , 2001) [10].

In his work, Breiman designed two significant improvements of Random Forest. First, Random feature selection, helping to reduce overfitting and enhancing generalization by picking up a subset of characteristics at each decision node. Second, bootstrapping helping in decorrelating the variables of the tree as well as lowering its variance (Breiman , 2001) [10].

Random Forest sets up multiple decision trees by training them on a randomly picked subset of observations and features. Each of these trees make their own predictions and a majority vote is used to provide the final prediction. This approach reduces overfitting and allows a huge dataset management as well.

Thanks to its various applications, Random Forest has gained in popularity over the years and is now a commonly used machine learning technique. It can perform a lot of miscellaneous tasks such as feature selection, clustering, etc. It can also create randomized trees, which in the end improves predictability and decreases bias (Siroky, 2009) [79].

Its capacity to successfully handle high-dimensional data is one of its significant advantages. Random Forest is effective at handling datasets with lots of characteristics since it minimizes the overall number of features needed for predictions. This can be particularly helpful in circumstances where feature reduction or selection is sought. Additionally, it performs especially well with datasets that contain duplicate features.

However, Random Forest also has significant drawbacks. The computational expense of the method is one of its main constraints, especially for large datasets when training many decision trees is necessary. Additionally, it may not perform as well as other methods if the dataset has considerable class imbalance. The majority vote aggregation could indeed foster the dominant class. Despite these drawbacks, it can still be an interesting model for specific applications. Furthermore, some hyperparameter tuning and feature engineering could even further boost its global performance.

We will end the investigation of diverse machine learning models by the Support Vector Machine (SVM). SVM was first presented in 1992 by Boser, Guyon and Vapnik at the COLT conference (Boser et al., 1992) [9]. Its initial approach relied on identifying hyperplanes that maximize the margin between two classes where the support vectors of these classes are actually the nearest data points to the decision boundary. (Vapnik, 1963) [85].

Over the years, SVMs have been subject to substantial research and development, expansions and adjustments to improve their efficiency for various types of data and applications. The introduction of multi-class SVMs, kernel approaches, and non-linear SVMs are some examples. Nowadays, they remain widely used in machine learning for classification, regression, signal processing, computer vision and genetic algorithms (Cervantes et al., 2020) [15].

SVM first tries to maximize the margin between classes by ensuring that all instances of one class are on one side of the boundary and instances of other classes are on the other side. Some techniques can efficiently handle this problem (Platt, 1998) [68]. When linear separation is not possible, SVM can apply kernel functions such as polynomial, radial basis function (RBF) or sigmoid functions to map the data into a higher-dimensional space for linear division. This choice depends on the data specifically (Kasnavi et al., 2018) [47]. The regularization parameter could significantly influence its performance. It implies a trade-off between a wider margin and the tolerance of certain training errors.

One of SVM's main benefits today is its adaptability: by utilizing kernel functions, it can handle both continuous and categorical data, high-dimensional data and non-linear connections between features. (Cervantes et al., 2020) [15]. These kernel functions are actually used to map data from one space to another, allowing then to locate the hyperplanes. Along with binary and multi-class classification tasks, SVM is also robust against datasets full of outliers and is able to detect complex decision boundaries in data as well. (Cervantes et al., 2020) [15].

Nevertheless, SVM also presents drawbacks, reliant on the parameters used to train the model, like the kernel function. This weakness could become a significant problem when resources are not sufficient to perform a hyperparameter optimization. To obtain the best results, parameters must then be properly tuned (Cervantes et al., 2020) [15]. The difficulty to manage imbalanced datasets is also part of its drawbacks. In such situations, minority classes might not be fairly represented compared to the majority ones. To lessen the effect of class imbalance on SVM performance, appropriate strategies can be applied, such as class weighting, oversampling, or undersampling (Wang & Shi, 2008) [90].

Therefore SVM is a strong and flexible classification technique that can handle a wide range of variables and non-linear connections between features. However, it can be highly demanding to optimize in real life applications (Cervantes et al., 2020) [15].

In conclusion, the development of diverse statistical models has undoubtedly led to significant progress in this field. This nevertheless raises the question of whether the comprehension of human language by machines has been developed at the same pace as these models. Human language, heart of our conversations, will remain the easiest way for humans to embrace technologies. The interaction human-machine shows up again. To elaborate on this reflection, we will retrace the history of the comprehension of human language by machines. This research field is called Natural Language Processing and belongs to the same broader discipline as Machine Learning, Artificial Intelligence.

Defined as *“a branch of artificial intelligence that helps computers to understand, process and analyze large amounts of natural human language data”* NLP has become one of the most popular and indispensable technologies for organizations (Bhattacharjee, 2022) [8]. Siri, Alexa, Google search and more recently Chat-GPT, make Natural Language Processing one of the most challenging hot topic in AI (Corp. C, 2021) [23]. To get a better grasp on how computers can understand human language, let’s explore its history.

The roots of the NLP date back to the 1950s. In its early stages, NLP was first used for translation tasks. Linguistics, mathematics and computer science laid the first foundations of upcoming NLP technologies.

For three decades, AI made remarkable leaps forward, focusing mainly on improving human-machine dialogue. This period marked a significant shift in Natural Language Processing. Researchers explored the possibilities of machines communicating with humans without any mathematical proof. This period is still considered as the golden years of NLP (Corp, 2021) [23].

After that, in the 1990s, the apparition of vast amounts of data and the development of internet (M. Srivastava, 2014) [80] prompted the interest in NLP. Probability methods and models markedly developed the range of possibilities offered by NLP. Hundreds of new doors were opened for further investigations (Corp, 2021) [23].

NLP has seen a significant increase in popularity in recent years. Driven by data, algorithms and computing power, the whole topic of AI has entered a period of rapid development. As a result, research has been strengthened, accelerating progress. Overall, this progress showed three main characteristics. First, the availability of data and diverse knowledge has fueled the advancements. Second, the emergence of deep learning has enhanced progress as well. Third, NLP has reached human-level abilities for some tasks, which represents a real milestone in its progress (Corp. C, 2021) [23].

Among these numerous improvements, new pre-trained models have particularly revolutionized the field of NLP, pushing the state of the art for several tasks. In parallel with recent advances in neural architectures such as Transformers, variants of these models have been developed to achieve impressive results with cutting edge technologies (Chernyavskiy, 2021) [19].

Nevertheless, this promising discipline obviously faces numerous challenges. First, machines still struggle to understand real-life situations and common sense, relying heavily on massive data rather than deeper knowledge. Second, assessing an NLP model and improving its robustness against noise still remain difficult. Third, AI ethics is still failing due to the lack of reinforcement mechanisms (Hagendorff, 2019) [37]. Ethics is a branch of philosophy that determines right and wrong by analyzing and defending different concepts of values (Gilbert, 2023) [35]. For now, machines are not able to act according to a certain set of values. Furthermore, deviations from a potential code of ethics are currently not penalized. Therefore, compliance and ethics in AI become a real problem.

All in all, NLP remains one of the most interesting topics for AI enthusiasts. Notwithstanding challenges, opportunities but also dangers, the discipline has undoubtedly become the subject of all covetousness at all levels. From compliance to improvement of results and extension of achievable tasks, NLP is the trendy subject, in which machines and humans become one.

To get a better grasp of what an NLP model consists of, let's have a look at "BERT", a Bidirectional Encoder Representations from Transformers (Vaswani & al.) [86] .

BERT algorithm is quite recent (2018) and belongs to a larger group of models called Transformers. For now, they are the best performing models achieving state-of-the-art results in many NLP tasks (Ganesh, 2021) [34]. In this paper, we will only focus on the BERT algorithm and some of its variants (Acheampong, 2021) [2]. All of them follow a multi-layer bidirectional Transformer encoder architecture, which is particularly powerful for NLP tasks (Xu, 2022) [96].

BERT and its variants are actually pre-trained language models. This pre-training phase allows these algorithms to first learn patterns from a huge amount of data and to understand how words and sentences work together. Throughout this process, BERT and its variants analyze sentences and predict masked words based on their context, which is called Masked Language Modeling (Acheampong, 2021) [2]. Over the years, some variants have been implemented for scientists to make a trade-off between execution speed and performance but also to test other tokenization techniques.

Tokenization aims to break down words into smaller tokens. These tokens will then be used to represent a word in a space, compulsory step in the algorithm. Over time, different tokenizers have been developed beyond the classic BERT tokenizer such as Roberta or DistilBert.

The classic BERT tokenizer employs WordPiece tokenization, breaking words into subword units based on frequency in training data to handle out-of-vocabulary words. It includes special tokens like [CLS] and [SEP] for input sequence marking, although this is not specific to a dataset. In this thesis, we will try the basic model but also the larger one, trained on a bigger amount of data.

DistilBERT, a smaller and faster version of BERT, also uses the WordPiece tokenizer but in a simplified way, with fewer layers. While this may be useful for large data sets, its impact will be tested in our work.

RoBERTa, on the other hand, uses Byte-Level BPE tokenizer, which excels in handling non-Latin characters and special symbols. This variant will also be tested in this thesis.

Once the pre-training and tokenization have been completed, the algorithm can train on a particular dataset. This second training process is generally called fine-tuning among scientists (Acheampong, 2021) [2]. Vectorization occurs during the second training process. Also called word embedding in the NLP world, this process aims to represent the words by vectors. Different techniques such TF-IDF, CBOW do exist but we will not detail them in this work (Adewumi, 2020) [3].

Overall, BERT presents some great benefits that make it an effective tool for processing written language. Understanding a context or sentiment analysis are good examples. As a rule, BERT trains rather quickly compared to other NLP models and has been used for many different language tasks, making it a versatile tool (Acheampong, 2021) [2].

However, BERT also has some drawbacks. For example, it can only work with one language at the time, and it does not process well long sentences. At times, it may also have trouble understanding the hidden meaning of a sentence, which can lead to errors (Acheampong, 2021) [2]. Additionally, as these models have billions of parameters, they are sometimes too computationally intensive to suit low capability devices (Ganesh, 2021) [34]. New variants have also been developed to fix this issue as detailed in the study of Mohammed (2021) [57].

One cannot address NLP without mentioning the most advanced chatbot, ChatGPT. This AI system has revolutionized its industry in the last few months making Natural Language Processing an even actual topic.

This website accessible to the general public is able to generate answers to questions, hold conversations with humans and much more. It has become in a very short time the new revolution in the world of artificial intelligence. Its user-friendliness allowed the general public to access ChatGPT, which is a turning point for OpenAI, the owner company.

Founded in December 2015 as a non-profit organization, its goal was at that time “*to promote the development of AI in a safe and beneficial way for everyone*”. In 2019, Open AI decided to set up a second entity, which was meant to make profit to fund the research of the first one. They first signed a \$1 billion agreement offering exclusivity of the GPT-3 model to Microsoft. Aware of GPT’s potential, Microsoft invested later another \$10 billion.

Since its creation, OpenAI has released several versions of their GPT model. In 2018, the first GPT-1 model was presented as a language model with 117 million parameters trained on a dataset of approximately 8 million web pages. GPT-2 increased this number of parameters to 1.5 billion in 2019, trained on 40GB of text. A third upgrade was implemented in 2020 with GPT-3 which would use 175 billion parameters and 570 GB of text from the Internet.

The public launch of ChatGPT occurred 2 years later, in November 2022. This newly released technology provides an updated version of the previous one, GPT-3 to fix some issues. Finally, the most recent language model from OpenAI, GPT-4, was ultimately released in March 2023. This state of the art model claims deductive reasoning skills and increased precision in tackling challenging issues.

In conclusion, this review of literature leads us to comprehend the diversity of tasks performed by natural language processing. From the ingenuity of BERT to the user-friendliness of ChatGPT, NLP enjoys a great diversity, as does Machine Learning. However, these two disciplines have been analysed in isolation, with no real bridge between them. As we have seen, making machines more cognitively human would facilitate many applications. A widespread application of human language would indeed allow to cross the borders of the scientific community. In this world of Intelligent Automation, what amount of information goes astray when we use human language instead of quantitative data?

To research this question, it is important to imagine a context and a task that would suit both empirical data and human language. While surfing on the internet, we discovered Winedexer [95] website, proposing a quantitative methodology to describe wine. Its tannicity, power, acidity were indeed translated onto a numerical scale. As wine seemed to be the perfect match to challenge NLP, we decided to compare empirical and NLP models related to wines. Its diversity, but especially its lexical richness led us to consider the relation between a specific lexicon and the performance of NLP models. Experts accurately describe wines with an extreme precision using human language, pushing it to its limits. The NLP model responses to this word abundance were quite uncertain. At this point, we asked ourselves whether this would help NLP models or, on the contrary, challenge them. All in all, this brainstorming process gave us the opportunity to explore the world of wine, so vast and peculiar.

Wine and its industry have a millennium-long history that can be traced back to ancient civilizations. Wine has played an important role in society throughout history, as a much appreciated beverage, symbol of cultural heritage, and social status indicator (Ingilis, 2022) [45]. Taste studies have largely participated in understanding the complexity of wine, turning it into an exquisite. These studies have gone deeply into wine’s sensory qualities, exploring its flavors, smells, textures, and the delicate interplay between them (De Mets et al., 2017) [27].

Wine appreciation and knowledge evolved alongside civilizations. Wine regions arose, each with its own distinct climate, soil, and grape varieties. It gave rise to the concept of terroir; the notion that the characteristics of the place are linked to the wine it produces. Even in recent years, the diversity of grapes has continued to increase (P. This, 2006) [83]. This richness created various expressions of wine, allowing emergence of sommeliers and wine connoisseurs who are passionate about investigating and analyzing the nuances of this ancient elixir.

To solve the mystery of wine, a specialized vocabulary was created, consisting of concepts that encapsulated the multiple character of the beverage (B. Chen, 2018) [18]. Wine terminology evolved to include a variety of aromas like earthy, fruity or floral notes, up to represent the effects of oak barrel aging. This lexicon developed a rich list of terminology, ranging from descriptions like “crisp” and “velvety” to more esoteric terms like “bouquet” and “legs,” allowing specialists to communicate the numerous qualities of wine with clarity and eloquence.

As the market developed and the vocabulary grew, a gap between wine experts and standard consumers started to appear (Cruz, 2018) [25]. For those unfamiliar with the language of wine, the complex terms and nuanced descriptions employed by experts may appear overwhelming and unintelligible. People who wanted to continue to learn about and enjoy wine, without feeling ousted or frightened by the complexity of the lexicon, observed this divide as a barrier creating some appreciated sense of exclusivity.

However, as the wine business was developing, it became clear that it was crucial to cross this gap and increase access to wine. As a result, the terminology used to describe wine was not really simplified (Zubek et al., 2016) [100] but opportunities where one can learn and explore at its own pace have been set up. Oenology courses, beginner-friendly tastings sessions, blogs or amateur magazines such as “Wine Spectator” or “Wine Enthusiast” have all evolved to give consumers information and assurance needed to navigate the wine industry.

In short, the creation of a specialized language and taste research have helped us to better comprehend this extraordinary beverage throughout the history of wine and the wine industry. A shared appreciation for the creativity and joy that wine can provide is fostered by the continual commitment to inclusivity and education, which bridges the gap that the complex vocabulary is creating between experts and regular consumers.

On a higher view, the whole wine industry is currently going through a period of digital transformation. In this highly competitive and rapidly changing sector, digitalization becomes a game changer and provides a competitive edge when it comes to strategic decisions. The wine industry’s automation has also transformed how wine is produced, promoted, and consumed. Technology has cleared the path for novel solutions that improve efficiency and accessibility across the whole supply chain. Digital tools have enabled winemakers to make documented decisions and optimize their processes. It goes from vineyard management systems that use IoT sensors to monitor soil conditions and weather patterns (Hnatiuc et al., 2022)[41] to precision winemaking procedures guided by data analytics. E-commerce platforms and online marketplaces have made it easier for consumers to discover a wide range of wines from across the world, which are frequently accompanied by thorough descriptions, user reviews, and personalized suggestions.

Furthermore, smartphone apps, internet platforms dedicated to wine education and sommelier-guided experiences have enabled wine fans to broaden their knowledge and tastes. As an example, vin.co [88] is an application on which small and large merchants and wineries can sell their products online. It offers tasting sheets that customers can download to make this complex field a bit more accessible (Nation, 2023) [60]. The digital era has not only improved the wine industry’s operational features, but it has also opened up new doors for wine appreciation and global connectivity between producers and consumers. As a result, AI, chatbots and other technologies have now become an unmissable tool in many companies.

Wine language and digitalization provide an opportunity for Natural Language Processing investigation. NLP has shown to be a powerful tool for understanding and extracting meaning from textual data, thanks to its computational methodologies and algorithms. However, the distinctive characteristics of the wine industry present major challenges to the application of NLP. Wine vocabulary includes a wide range of terminology, covering adjectives for flavors, aromas, textures, and even the emotions induced by the tasted beverage. Wine descriptions also come with their own set of challenges, often perceived as

vague and overly literary prose (Caballero, 2007) [13]. The difficulty stems from the subjective nature of odors and flavors, which are thought to be difficult to express precisely (Levinson Majid, 2014) [52]. This throws doubt on NLP’s ability to accurately capture and understand the complexities of sensations represented in wine reviews.

To begin a literature exploration at the junction of NLP and the wine industry, we should first examine the existing corpus of research. It will allow us to gain useful insights into the potential and limitations of NLP when confronted with the complex wine language by reading previous studies.

The main goal of this investigation will therefore be to explore the gap between wine related language and NLP’s computational abilities. Some presented studies will attempt to provide deeper insights into consumer tastes, comprehend wine complex scents, or even enlighten wine business decision-making processes. Furthermore, experiments have been conducted to investigate how NLP models can recognize the complex meanings of wine-related terms and phrases (Lefever et al., 2018) [51]. Therefore, these studies give light on the extent to which NLP can capture the complex variety of the wine vocabulary, from separating the details of regional wine variations to understanding the tiny differences among various grape varieties.

By reading the findings and insights from existing literature, we will gain comprehensive understanding of the progress made thus far in applying NLP techniques to the wine industry. Above all, we will be able to identify gaps in current research, paving the way for our future investigations.

Natural Language Processing techniques in the wine industry have resulted in substantial advance in interpreting and categorizing taste nuances. These advancements have enormous potential for both wine specialists and wine consumers, promoting a deeper awareness and comprehension of the subtle components of wine flavors.

The IT framework of taste in the wine industry has nevertheless already a small historical background, starting its development with B. Chen in 2022 [17]. Chen’s work was built upon the foundation of the Wine Aroma Wheel, initially created by Ann C. Nobel, which aimed to categorize fragrances and flavors experienced when tasting wine. However, the initial version of the Aroma Wheel was insufficient for computers to comprehend reviews as it only focused on describing tangible attributes, neglecting adjectives and wine body (B. Chen, 2022) [17].

Chen employed the widely used bag-of-words technique to improve the Aroma Wheel. This technique, which involves tokenization, normalization, removal of stop words, and analysis of N-grams, proved to be highly efficient in processing wine reviews. Recognizing the effectiveness of this process, Chen gradually refined the framework, dataset by dataset, leading to the development of the Comprehensive Wine Wheel (CWW) 2.0. This CWW aimed to capture all dimensions of wine that experts could describe, offering a more comprehensive representation of taste with an extended range of adjectives.

In a recent development, Q. Tian, a graduate from the University of Central Arkansas, further improved upon the CWW 2.0 (Tian, 2023) [84]. Following B. Chen’s suggestion, Tian incorporated reviews by renowned expert Robert Parker into the framework, enriching its content and insights. Additionally, Tian evaluated the consistency between reviews from Wine Spectator and Robert Parker, discovering methods to effectively combine the two perspectives and achieve optimal results. These advancements have opened up possibilities for further improvements. One potential avenue is the inclusion of reviews from other sources or wine publications to enlarge the comprehensiveness of the wheel. Another promising area of improvement involves incorporating related word groups that may not be adjacent to each other but share semantic connections. Furthermore, while the current focus of the wheel remains on traditional and outstanding wines, there is potential to expand its scope to encompass a negative lexicon.

However, it is essential to recognize that while the aroma wine wheel serves as a useful framework for organizing aromas, its composition of adjectives displayed in a wheel structure does not provide a complete and comprehensive understanding of this complex matter. It does not represent the true ability of NLP models to fully comprehend and interpret the nature of wine descriptions. Despite being a useful tool, the wheel only depicts a small portion of the complexity of wines. Additional analytical and modeling techniques are required to comprehend wine characteristics. This includes investigating sophisticated NLP approaches and digging into complex interactions of language, context, and sensory experiences.

These complexities can be found, for example, in the NLP analysis of wine reviews on the Internet. Businesses could again gain significant insights about client preferences and expectations by investigating user-generated reviews on platforms like Yelp or TripAdvisor. For example, in a study of wine bars in Busan, South Korea, text mining and network analysis approaches were employed to find significant variables associated with positive online ratings (Fu et al., 2022) [33]. With this insight, wineries can devise strategies to increase consumer satisfaction and loyalty. Wine reviews are known to be the biggest source of information about wines (Sadahiro, 2021) [73]. A thorough analysis of their semantics is therefore essential in our work (Cao, 2011) [14]. Some experts claim that the descriptions of sommeliers, artisans of taste as defined by S. Olivesi (2015) [63], are even the best in terms of describing the quality of a wine (B. Kim, 2022) [70].

An effective review processing could actually become an important stake within the industry. As an example, the efficiency of a model like TourBERT in the travel industry (Arefeva & Egger, 2022) [5] could be hint at the viability of developing such models in the wine industry. These models could train on wine-related text data and refine for a variety of tasks such as sentiment analysis, topic modeling, or text classification. Wineries that use such models will benefit from significant insights and updated decision-making tools. As Hendrickx (2020) and Lefever [24] indicated in 2020, wine experts can describe wines in such a way that we can use their reviews to build NLP models. A future work presented

and not addressed in their study is the possibility to create an NLP model based solely on these reviews (Hendrickx & Lefever, 2023) [24]. Their study really represents for us a baseline from which we will only retrieve the useful character of reviews.

Finally, the potential influence of NLP models on the wine industry is enormous. They enable firms to efficiently meet their customers' demands by facilitating a comprehensive grasp of their preferences. Furthermore, NLP contributes to client satisfaction and loyalty, which eventually drives corporate success. As this technology advances, new domain-specific models, created for the wine business, are expected to emerge. This will increase the value and utility of NLP in this dynamic and changing sector.

This analysis has therefore enabled us to discover in greater depth the subtle marriage between NLP and wine. The creation of the aroma wheel in first instance and the exploitation of professional reviews in second instance really allowed us to challenge the most powerful NLP models in an interesting way. Therefore, a meaningful task would be to predict the variety of a wine based only on its reviews for example. However, this NLP challenge lacks opponents. At the moment, it is difficult to evaluate and compare the performance of a model without strong adversary. The return of empirical models is therefore underway. Such comparison remains very rare in literature and even more in the lexically rich field of wine. By assembling the pieces of the jigsaw, we can ask ourselves the question of a confrontation between NLP and quantitative ML in the context of a wine variety prediction task.

One day, while surfing on the internet, we came across a quantitative translation of the taste of a wine through a well established framework. It made perfect sense to establish a confrontation between these two disciplines that are similar in their roots, AI, but at the same time extremely different in their functioning. The wine tasting would take the form of human language in one model and empirical data in the other. All this in the specific context of a wine variety prediction task.

In addition, an interesting parallel could be drawn with ChatGPT. This second step also represents one of the first quantitative studies of the performance of ChatGPT. The comparison of a model like ChatGPT with the ones previously mentioned could also highlight the limits of this tool. The objective being then, within the NLP models, to compare the performances of two different models ChatGPT-4 and BERT, presenting very different characteristics.

To go further, the final stage of this work will concern an in-depth evaluation of our models. Some prediction mistakes will probably be made, like classifying a white wine in a red category. This means that it will be of interest to assess our different models according to a wine-specific metric. This will then rely on the resemblance between different varieties of wines.

In this regard, it could be interesting to check the existing literature. In 2017, Maan Hasan Salih [74] proposed a methodology to compute the genetic distances between grape varieties. His analysis was based on the DNA of different grapes. Thanks to some Random Amplification of Polymorphic DNA markers, he found a biological way to compute genetic distances between grape cultivars. Biological and technical details of his process can be found in his study. The final results can be summarized as follows:

|             | Kishmishi | Des-Alaniz | Aswad | Omeeri | Buhrizi | Zaitouni | Halwani | Kamali | Shada Bedha |
|-------------|-----------|------------|-------|--------|---------|----------|---------|--------|-------------|
| Kishmishi   | 0.000     |            |       |        |         |          |         |        |             |
| Des-Alaniz  | 0.205     | 0.000      |       |        |         |          |         |        |             |
| Aswad       | 0.227     | 0.115      | 0.000 |        |         |          |         |        |             |
| Omeeri      | 0.198     | 0.159      | 0.200 | 0.000  |         |          |         |        |             |
| Buhrizi     | 0.219     | 0.164      | 0.165 | 0.142  | 0.000   |          |         |        |             |
| Zaitouni    | 0.202     | 0.219      | 0.201 | 0.145  | 0.098   | 0.000    |         |        |             |
| Halwani     | 0.198     | 0.178      | 0.180 | 0.140  | 0.092   | 0.097    | 0.000   |        |             |
| Kamali      | 0.218     | 0.144      | 0.146 | 0.155  | 0.125   | 0.141    | 0.103   | 0.000  |             |
| Shada Bedha | 0.201     | 0.168      | 0.150 | 0.213  | 0.124   | 0.142    | 0.157   | 0.122  | 0.000       |

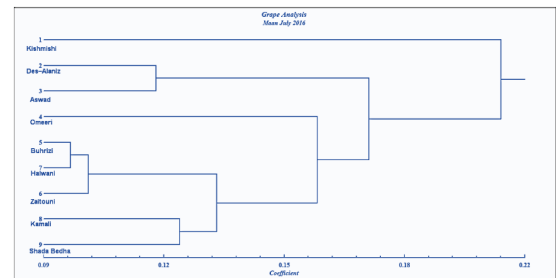


Figure 1: Genetic distances and dendrogram ([74])

Maan Hasan Salih [74] found that the DNA of grapevine influenced the expression of genes responsible for various chemical compounds contributing to the taste and other sensory properties (Cho, 2015) [21]. As the taste is the only criterion taken into account, other physical parameters being totally neglected, it perfectly connects with this thesis. However, the methodology seems difficult to be applied in this paper. Therefore, it can only give a good preamble for future work, biology-wise.

Based on our findings, this section has guided us through the literature from a general view to our research question, broken down in the next section.

## 2.2 Research Question

The lexical abundance of the wine domain combined with the technical advances of Natural Language Processing have aroused our curiosity. Does such an extensive vocabulary benefit to text mining models or does it, on the contrary jeopardize them ? The wine world is known to be highly demanding in terms of knowledge and expertise but relatively hermetic to the outside world. This industry was then particularly interesting to analyze.

Meanwhile, new AI tools are booming, thanks to their userfriendliness and attractive interfaces. The help in writing that these technologies offer makes them easier to adopt and appealing to the general public. Everything seems easier when we can interact with human language, describe, express our wishes and discuss.

However, machine learning models, based on empirical data, are both much less accessible to most people and even sometimes more time and resource consuming. Writing an opinion on a wine or explaining what you would like to drink seems much more accessible and quicker than having to provide a list of precise characteristics you wish to taste.

Moreover, comparisons between human language (unstructured data) and empirical data remain scarce in the literature. Only really case-specific papers exist (D. Hopkins, 2022) [43]) but none of them tries to quantify the loss when switching from empirical to textual data.

Based on these observations, we wanted to investigate the gap between human language and empirical data in a situation where the lexical field is extremely broad, a wine variety prediction task. Through a comparative analysis, this study aims to uncover potential benefits and challenges of language-based approaches, contributing to the development of NLP technologies. It could also drive professionals to use human language more effectively in their industry.

Therefore, our research question will be:

*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 ?*

## 2.3 Business Incentives

Notwithstanding its technicality, this work may be of interest to the business world. We hope to make our contribution to the wine industry, benefiting to customers and sellers alike.

First, recommender systems could be upgraded to human language. By providing feedback on a particular wine, a system could, without losing too much information as shown in our results, meet customer’s preferences and recommend the most suitable wines. Unlike the machine learning model, a non-expert would not be disadvantaged in the recommendations. Retailers would be able to understand their customers more easily. This level of personalization can contribute to higher customer retention and loyalty, as customers feel understood and followed up on an individual level. Satisfied customers are more likely to repeat purchases, engage in positive word-of-mouth recommendations, and develop long-term relationships with the business. Incorporating human language is already an important issue in the world of recommendations.

Second, in the wine sector, incorporating human language capabilities into a business plan may promote innovation and technical improvement, labelling the company as cutting-edge and attracting customers looking for modern wine selection techniques. We swiftly believe that many businesses from various industries, including the wine industry, will quickly adopt such technology as market opportunities gain pace. Additionally, the popularity and viability of language-based solutions are influenced by the general public’s access to AI tools, which may revolutionize this industry and provide competitors an advantage. For instance, a restaurant making use of human language recommender system may offer its guests an exclusive experience, which could make it stand out against its competitors.

Moreover, understanding the capabilities and limitations of various NLP models, such as ChatGPT and BERT, can assist businesses in making informed decisions regarding the adoption and use of these technologies. Our findings might help firms to choose the best model for their specific needs. Businesses may align their requirements with the capabilities of the models by recognizing their strengths and shortcomings. For example, if ChatGPT performs close to BERT and the empirical ML model, companies may prefer ChatGPT because of its affordability and ease of integration. This can be especially beneficial for small to medium-sized organizations with limited resources, allowing them to harness NLP technology without incurring large financial costs or skills.

An analysis of both an ML model based on continuous and binary data and an NLP model based on wine descriptions may provide organizations with a better knowledge of client preferences and behavior, AI thus becoming a powerful marketing tool. Our findings could therefore help businesses detect the relevant characteristics of some wine from a customer's point of view by investigating the most relevant characteristics that lead to accurate wine variety prediction. As a result, organizations will be able to identify different client segments depending on their preferences for those same characteristics. For example, our model could highlight that tannicity is an extremely important parameter in the differentiation of grape varieties. This information will therefore be valuable since knowing customer's tastes on this single characteristic will already allow to classify them effectively. Nowadays, wine marketing sometimes seems to be limited to separating men from women with more or less fruity and sweet wines. The insights from our analysis could change these segments to more effectively target the right customers for the right wine. Similarly, the NLP model may extract insights from wine descriptions, assisting businesses in understanding the language and vocabulary that customers use the most to define their wine preferences. This method of segmentation and targeting could enable firms to adjust their marketing operations, product offerings, and customer experiences to specific consumer segments, resulting in increased customer satisfaction and conversion rates. This knowledge can lead to the creation of new wine varieties, the refinement of existing products, and the development of marketing strategies that align with customer

preferences. By incorporating customer insights into product development, businesses can create offerings that are appealing to their target audience and gain competitive advantage in the market.

At the same time, an in-depth understanding of ChatGPT and chatBOTS in general could allow businesses to deploy that kind of technologies in different fields. This can be especially interesting across industries where the vocabulary used is really specific. Finance is a concrete example. People could, as a result, request precise information about investment strategies or just ask specific questions that would allow bankers to focus on high added value work. This can really improve the efficiency of hundreds of companies and make them evolve faster. The implementation of these ChatBOTS can then be perceived as a profitable tool more than a gadget.

### 3 Data and Methods

Having reviewed the literature, we can begin the experimental part of this thesis. The aim of this section is therefore to present the data exploited and to propose an first analysis of them. The final subsection, methods, will focus on our implementation choices and our approach. At the end of this chapter, everything will be then in place to produce our results, which will be analysed in the next section.

#### 3.1 Datasets

The first step is to create datasets. This needs to be well thought out since this step will condition the performance and validity of our models. Thus, we must compose two different datasets.

This first dataset is elementary. It is composed only of an explanatory variable, the reviews, and an explained variable, also called target variable, the grape variety. For this dataset, we have taken an existing dataset from 2017, including all the wines scanned from the “Wine Enthusiast” website [93] available in this Kaggle. Among all variables, we only kept the reviews and the grape varieties. As previously explained, this dataset will be used for the NLP model based only on the reviews. Here is how an overview of the dataset:

| Reviews                                                                                                                                                                                                                                                                                                              | Variety            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|
| Mildly pickled red-fruit aromas fail to convey freshness or ripeness, and so it's no surprise that the core flavors are a bit herbal and briney. Decent mouthfeel helps it along, while the finish is short and borders on medicinal.                                                                                | Cabernet Sauvignon |
| Smells herbaceous and weedy, with lots of green, stalky notes and just the barest trace of berry fruit. The flavors are pretty much the same, thin and austere, with a solid dose of tannins. It's well made, but there's not a whole lot to like here.                                                              | Cabernet Sauvignon |
| Like all of Reverie's 2007 reds, this one's super-tannic. In fact, it's one of the hardest, most astringent Napa Cabs on the market. Still, it has a rich core of blackberries and currants, with lots of smoky oak, and the spicy fruit lasts a long time on the finish. A good sign for cellaring. Try after 2013. | Cabernet Sauvignon |
| Straightforward plum and rubber aromas seem slightly baked or burnt, while the palate features hard peppery flavors at the center of a powerful yet mildly scouring red-fruit palate. Not offensive but basic and raw. Imported by South Valley Wines.                                                               | Cabernet Sauvignon |

Figure 2: NLP Dataset layout

The second dataset will only consist of empirical data related to taste. As there is currently no existing dataset with these specific characteristics available online, we decided to build it using web mining techniques. Web mining is considered as one of the most popular and effective techniques of information retrieval. Web mining is defined as an application of data mining to extract and uncover relevant and hidden information on the web (Srivastava, 2014) [80]. To accomplish this task, we will mainly use the package BeautifulSoup [7] in Python, which is probably the most used package for web data extraction in this language [1].

In order to avoid random data extraction, we first have to figure out which data is required and in which proportion. This essential point ensures consistency, as our datasets must be aligned, it implies having similar proportions of wine varieties and approximately the same number of observations in both datasets.

Firstly, the number of varieties in the first dataset shows the richness of wine as 650 varieties were represented. As it was not possible to keep all of them, we decided to remove all blended wines that are composed of more than one variety, as Rosé. Otherwise, our predictions would have been very difficult to make.

Secondly, we will exclude all region-specific varieties, meaning that we only keep varieties out of which more than 1000 different wines are available. Some of the classes are indeed special with only one or a few observations. This choice will allow training process to be more robust.

Finally, it is crucial to ensure that the varieties in the dataset match exactly the ones documented on the “Wine Enthusiast” website. Even though a variety is highly represented and pure, we sometimes needed to remove it because we could not map it perfectly with a variety on the website. Pinot, for example could be decomposed according to many subvarieties which led us to leave it out of our selection.

Ultimately, we retained only a subset of 8 major grape varieties: Pinot Noir, Cabernet Sauvignon, Chardonnay, Riesling, Merlot, Malbec, Sangiovese and Zinfandel. These 8 varieties represent, after all removal steps, more than 80 % of the initial dataset. Also, as we wanted this work to be as real as possible, we tried to respect the reality of vine farming. It implies that among these classes, we only have two varieties of white grapes (Chardonnay and Riesling) while the others are red grapes. This 25-75 proportion is a good representation of the reality that locates around 20-80. A perfect match can be established between these classes and the ones available on the “Winedexer” website [95]. Here are the final proportions of the dataset selected to train the NLP model:

| Variety            | Number of Observations | Total |
|--------------------|------------------------|-------|
| Pinot Noir         | 8378                   | 23.8% |
| Cabernet Sauvignon | 8127                   | 23.1% |
| Chardonnay         | 8108                   | 23%   |
| Riesling           | 3274                   | 9.29% |
| Merlot             | 3111                   | 8.83% |
| Malbec             | 1582                   | 4.49% |
| Sangiovese         | 1414                   | 4.01% |
| Zinfandel          | 1224                   | 3.48% |

Table 1: Varieties proportions

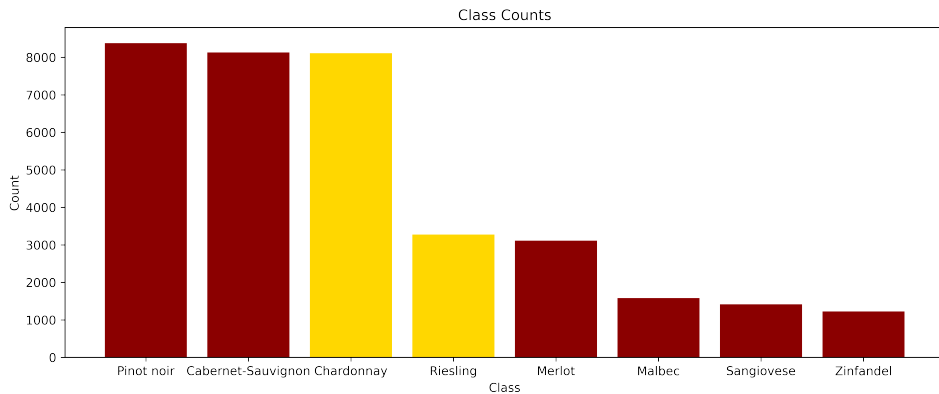


Figure 3: Wine varieties proportions

Since we have a better idea of what the first dataset consists of, we can start extracting data from Winedexer [95].

As an example, the layout of a detailed page of a random Cabernet-Sauvignon shows three panes that are especially interesting for this thesis:

- 1) The taste structure only composed of empirical data:

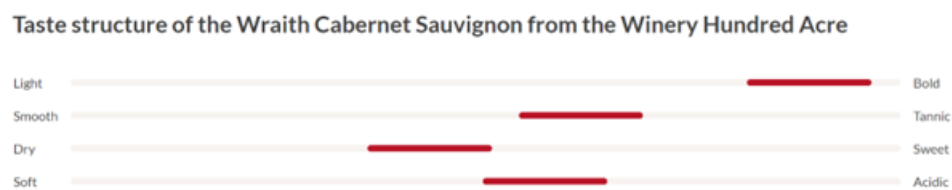


Figure 4: Quantitative taste structure of a wine

- 2) The wine flavors and olfactive analysis:

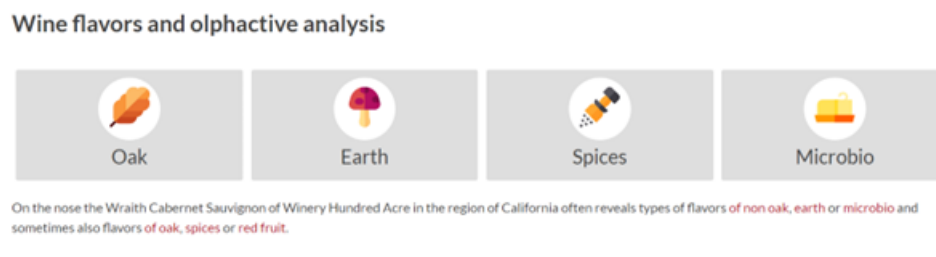


Figure 5: Aromas of the wine

- 3) The technical details :

|                          |                     |               |                   |
|--------------------------|---------------------|---------------|-------------------|
| Winery                   | Winery Hundred Acre | Country       | United States     |
| Grape varieties          | Cabernet-Sauvignon  | Style of wine | Red               |
| Region/Great wine region | Napa Valley         | Alcohol       | 15.5°             |
| Great wine region        | California          | Allergens     | Contains sulfites |

Figure 6: Technical Details of the wine

For each wine, we will extract all information available in Fig.4 and Fig.5. The Fig.6 will be used to check if it is single-varietal wine, yet to be synchronized with the first dataset. Accordingly, we will check the Grape varieties section to make sure that there is only one variety. If so, the wine can be uploaded in the dataset. The final goal is to extract as many wines as possible to be the closest as possible to the proportions shown in the first dataset. All Python codes are available in the appendix (I).

In the end, we succeeded to build up a solid dataset with 35 218 wines compared to the 38 368 of the first dataset based on the reviews. Prior to training models, certain adjustments will therefore be made to ensure complete alignment between both datasets.

In conclusion, this is what the dataset is composed of :

|    | Légèr à Puissant | Souple à Tannique | Sec à Moelleux | Doux à Acide | Aromes                                                                                          | Olfactives                                                                                   | Cépages            |
|----|------------------|-------------------|----------------|--------------|-------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|--------------------|
| 1  | 81.5062059;      | 54.0060706;       | 35.70447;      | 49.8511554;  | ///Boisé/Minéral & terre/Épices/Fermenté (levuré)                                               | d'évolution (vieillessement)/minéral & terre/fermenté (levuré)/de boisé/épices/fruits rouges | Cabernet-Sauvignon |
| 2  | 80.1895032;      | 62.0066381;       | 41.0377535;    | 54.3287955;  | ///Boisé/Sous-bois/Épices/Fermenté (levuré)                                                     | d'évolution (vieillessement)/sous-bois/fermenté (levuré)/de boisé/épices/fruits rouges       | Cabernet-Sauvignon |
| 3  | 83.372641;       | 50.110101;        | 31.002883;     | 54.3531582;  | caramel///Boisé/Sous-bois/Épices                                                                | d'évolution (vieillessement)/sous-bois/boisé/d'épices/fruits rouges/fruits noirs             | Cabernet-Sauvignon |
| 4  | 79.2265005;      | 53.1112656;       | 25.0184682;    | 65.731435;   | caramel///Boisé/Épices/Sous-bois                                                                | d'évolution (vieillessement)/sous-bois/boisé/d'épices/fruits rouges/fruits noirs             | Cabernet-Sauvignon |
| 5  | 79.3516191;      | 53.0021331;       | 23.5935234;    | 57.6366776;  | caramel/graphite///Boisé/Sous-bois/Épices/Floral/Fermenté (levuré)                              | d'évolution (vieillessement)/sous-bois/boisé/d'épices/fruits rouges/fruits noirs             | Cabernet-Sauvignon |
| 6  | 82.845441;       | 55.0033878;       | 23.107726;     | 56.5653475;  | caramel/graphite///Boisé/Sous-bois/Épices/Fermenté (levuré)                                     | d'évolution (vieillessement)/sous-bois/fermenté (levuré)/de boisé/épices/fruits rouges       | Cabernet-Sauvignon |
| 7  | 79.7632926;      | 57.1268169;       | 31.0374134;    | 57.4056986;  | ///Boisé/Minéral & terre/Épices/Fermenté (levuré)                                               | d'évolution (vieillessement)/minéral & terre/fermenté (levuré)/de boisé/épices/fruits rouges | Cabernet-Sauvignon |
| 8  | 82.697962;       | 57.8075565;       | 25.1649538;    | 58.3804004;  | /graphite///Boisé/Sous-bois/Épices/Floral                                                       | d'évolution (vieillessement)/sous-bois/boisé/d'épices/fruits rouges/fruits noirs             | Cabernet-Sauvignon |
| 9  | 81.0281659;      | 59.070682;        | 29.2457341;    | 56.9142898;  | boisé, tabac/terre, fumée/cannelle, réglisse/crème, fromage/violette///Boisé/Minéral & terre/Ép | de crème/minéral & terre/cassis/de caramel/menthe/cannelle                                   | Cabernet-Sauvignon |
| 10 | 76.497926;       | 54.6856204;       | 30.6966532;    | 56.528427;   | ///Boisé/Épices/Sous-bois/Fermenté (levuré)                                                     | d'évolution (vieillessement)/sous-bois/fermenté (levuré)/de boisé/épices/fruits rouges       | Cabernet-Sauvignon |
| 11 | 77.2656029;      | 59.4024005;       | 24.3360338;    | 58.763613;   | ///Boisé/Sous-bois/Épices                                                                       | d'évolution (vieillessement)/sous-bois/boisé/d'épices/fruits rouges/fruits noirs             | Cabernet-Sauvignon |
| 12 | 80.2704521;      | 59.3738549;       | 33.422119;     | 57.6827085;  | caramel/graphite/menthol, eucalyptus///Boisé/Sous-bois/Épices/Floral/Fermenté (levuré)/Vég      | d'évolution (vieillessement)/sous-bois/fermenté (levuré)/de boisé/épices/fruits rouges       | Cabernet-Sauvignon |
| 13 | 80.7773581;      | 56.6325862;       | 31.9900152;    | 54.0743446;  | ///Boisé/Sous-bois/Épices/Floral                                                                | d'évolution (vieillessement)/sous-bois/boisé/d'épices/fruits rouges/fruits noirs             | Cabernet-Sauvignon |
| 14 | 79.8159758;      | 59.5040409;       | 32.7769758;    | 58.1714245;  | caramel/graphite/menthol///Boisé/Minéral & terre/Épices/Floral/Fermenté (levuré)/Végétal        | d'évolution (vieillessement)/minéral & terre/fermenté (levuré)/de végétal/boisé/épices       | Cabernet-Sauvignon |
| 15 | 83.251771;       | 66.5549065;       | 25.6289719;    | 59.3597942;  | /graphite///Boisé/Sous-bois/Épices/Fermenté (levuré)                                            | d'évolution (vieillessement)/sous-bois/fermenté (levuré)/de boisé/épices/fruits rouges       | Cabernet-Sauvignon |

Figure 7: ML Dataset Layout

The first four columns are, as previously mentioned, our 4 empirical variables. They represent, on a scale from 0 to 85, the boldness, the tannins, the sweetness and the acidity of the white quantitatively.

After that, two different columns represent the gustative and olphactive flavors of the wine. Overall, around 635 different aromas have been detected throughout the entire dataset. To make this information usable in ML model, it should be turned into binary variables. In the end, there will be 635 binary columns representing aromas where “0” means that this aroma does not appear in the flavors of the wine and “1” means that it appears. Finally, the target variable, the variety is indicated in the last column.

As both datasets are ready to use in their respective models, the comparative analysis of these models can now be addressed. Despite their differences, they also share some similarities. Both are indeed based only on taste. However, the manner in which this sensory attribute is decomposed differs significantly. Because the quantitative model is projected to be more accurate, our primary goal will be therefore to determine the extent to which it outperforms the NLP model in terms of accuracy. We are also interested in analyzing if certain dataset features, like imbalanced classes, have different effects on each model.

However, before implementing the models, it appeared that out of the thirty-five thousand observations, nearly ten thousand had missing values for all four continuous variables. As a result, we decided to remove all these observations as they did not bring any new information. Replacing the missing data with class means would have distorted our results, as it would have reduced variance among observations and made predictions overly straightforward. To ensure alignment between datasets, we therefore had to remove an equal number of observations from each class in the NLP dataset too. By maintaining proportional representation across classes, we preserved the balance and consistency between the two datasets. Finally, the distribution of the latter can be described as follows.

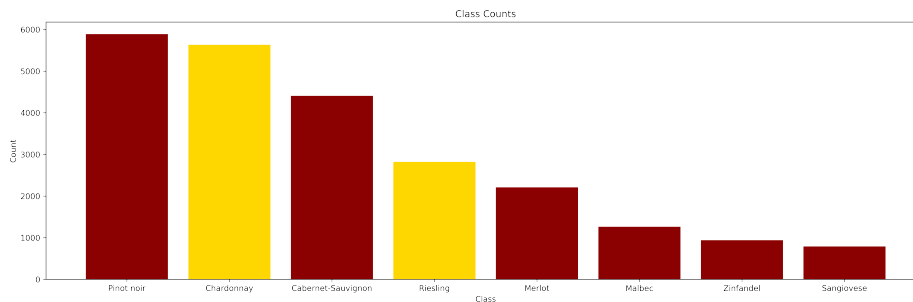


Figure 8: White & Red wines proportions

## 3.2 Datasets Analysis

Before entering our classification models, it is interesting to perform a first analysis of the datasets to extract first clues about the data which could be useful for future analysis.

Our NLP oriented dataset is composed only of two variables : varieties as target variable and descriptions. These data have been extracted from the “Wine Enthusiast” website in 2017 [93] and are available on the Kaggle [94]. We will therefore try to determine if it is possible to detect differences between the varieties based on the reviews. To do so, we will create wordclouds in order to identify the main flavors within the different classes. Differences between the wordclouds could therefore be indicative of proven differences between the classes.

The wordclouds are represented as follows:

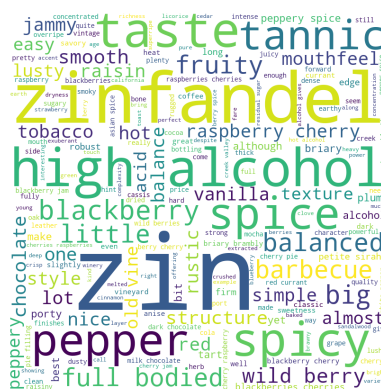
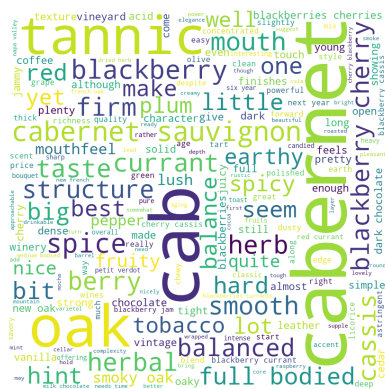


Figure 9: Wordclouds by variety

These wordclouds produce valuable insights as they differ considerably across the different varieties. This finding is already a great performance clue concerning a model like BERT since its performance relies significantly on lexical variations within different varieties. Therefore, noticeable discrepancies in these wordclouds would suggest the model performing quite well at first glance.

However, wordclouds also reveal challenging words to consider, such as “cherry”. It is the most common word for the variety Sangiovese, but it is important as well for varieties like Malbec or Merlot. Consequently, it raises the question whether this word is discriminant across these three varieties or not. This feature can be determined by BERT in the background, when vectorizing the tokens.

Another noteworthy observation within the wordclouds is the presence of variety names. We can often notice the name of the variety in the cloud, meaning that this word appears many times in descriptions. As the goal is to predict the variety, it would not make sense to keep these names as they directly provide the final answer. As a result, we decided to remove variety names from the descriptions. The cleaned wordclouds are available in Appendix (II.I).

As far as the other proper nouns are concerned, we decided to leave them as they were. After further research, we realized that the variety of the wines did not directly come from a specific country or region. We do not use protected appellation of origin wines such as Champagne in France or Chianti in Italy. Therefore, the identification of the region, country, or other pertinent proper nouns by a sommelier during wine tasting should not provide valuable information that could be leveraged by BERT in its analytical processes. On the other hand, removing them would automatically weaken the algorithm. As the length of the reviews are acceptable for BERT, it will not have to remove any words. Therefore the model will use proper nouns only if they are source of information.

The second analysis is dedicated to the machine learning dataset, containing only

quantitative data. The investigation was mostly centered on the continuous data, which included four metrics defining the key characteristics of wines: power, tannicity, acidity, and sweetness. The goal of this analysis was firstly to determine if all the wines within the same grape variety had similar characteristics. Secondly, it was essential to observe the distribution of the features across these varieties and get a sense of how well the model can predict these kinds.

We used a systematic strategy to get insights into these features within each grape type. Four different boxplots, one for each parameter, are available in the appendix (II.II, II.III, II.IV, II.V). Then, we investigated the characteristics using univariate histograms and pairplot to validate and deepen the original conclusions derived from the boxplots (II.VI, II.VII, II.VIII). In addition, we looked at the link between acidity and tannicity in all different grape types (II.IX), to better understand the proper characteristics of white wines and Pinot Noir.

Our observations showed significant differences in these properties amongst grape varieties, indicating distinct profiles within each category. Notably, we discovered that white wines may be easily identified based on tannicity. Indeed, white wines have no tannicity due to the winemaking process, which separates tannin-containing grape skins, seeds, and stems from the juice before fermentation. This separation prevents tannins from imparting astringent aspects to the wine. Furthermore, the high acidity of Riesling wines has the ability to effectively separate the two white wine varieties. This disparity is even more clear when looking at the appendix (II.IX), where the combination of acidity and tannicity underlines the differences between grape varieties.

It was also possible to underline that the density of observations is important with Zinfandel, Sangiovese and sometimes Malbec, showing especially clustered values in the boxplots. This result is particularly interesting since the dense representation of values around their median can suggest that the model will predict more easily these specific types. Pinot Noir, Cabernet-Sauvignon, and Merlot, on the other hand, show less compactness and dissimilarities in their boxplot representations.

Although there are variations among the boxplots representing different varieties, Pinot Noir stands out with its distinct values for certain characteristics combination like tannicity and acidity, as shown by the bivariate analysis from the pairplot (II.IX). Considering the abundance of wines belonging to this category (8378, 23.8%), the model is expected to possess the ability to make better predictions for this particular wine type.

Concerning the examination of binary data, which indicates the presence of aromas in different grape types, we discovered that certain fragrances, such as 'wood', 'spices,' and 'yeast,' are common across many, if not all kinds (II.X). Therefore, we used a chi-squared calculation to determine the importance of all aromas in the classification process (II.XI). This test can determine if there is a significant relationship between one variable and the target one. It compares the actual distribution of aromas across different grape varieties to the expected distribution if there was no association between aromas and wine varieties. If there is a substantial difference between the observed and expected distributions, it suggests that certain aromas play a significant role in classifying wines, indicating their importance in the classification process. According to our findings, the aroma of "fruitiness" emerged as a crucial element in appropriately identifying wines.

Furthermore, upon deeper inspection (II.XII), we determined that white wines have the most fruitiness, offering significant evidence for their classification. This analysis was not surprising as white wines are often very different from red wines. The question was raised whether it was possible to find aromas in red wines that provided as much information as fruit did for white wines. However, when the latter were omitted from the study, the importance of aromas in differentiating wines dropped (II.XIII). While certain scents such as 'sous-bois' remained relatively important, they were not as significant as fruitiness (II.XIV).

In addition, we investigated the presence of distinct aromas associated with individual grape types. The results, however, were not particularly instructive, with only a few examples of cranberry scent in Pinot Noir and petroleum perfume in Riesling (II.XV). Furthermore, we analyzed the fraction of wines with aroma present in a maximum of two kinds, but no much more noticeable results emerged, with the exception of Riesling which revealed some other distinct aromas still associated with petroleum (II.XVI).

In conclusion, while fragrances can help distinguish white wines and Pinot Noir from other red wines, their overall relevance is lower than in the study based on continuous variables. The continuous data, which includes qualities like power, tannicity, acidity, and sweetness, appears to have a higher predictive value for correctly classifying the wines.

### 3.3 Methods

Our research question will be addressed in two steps. First, we will try to predict the wine variety. Then, we will proceed with an in-depth comparison and evaluation of our models.

The first model, inspired by the ChatGPT wave, is about gathering preferences of knowledgeable customers. It aims to predict wine variety solely based on sensory descriptions, using advanced NLP models that analyze only textual data. ChatGPT and BERT are game changers in the world of Natural Language Processing. We believe that both have the potential to revolutionize the study of language, from its creation to its processing, which is why we have chosen to focus on them for the NLP analysis in the framework of this thesis.

As previously explained, BERT is a pre-trained model based on a huge amount of text data. After loading the pre-trained model, it is necessary to fine-tune it by training the model on the specific dataset to make it learn new patterns, specific to wine in this study. The fine-tuning will enable the comparison of performance of the model with ChatGPT, trained massively more than task-specifically. As the lexical field of the wine world is extremely specific, we expect a specifically trained model as BERT to perform better than the chatBOT. Otherwise, there would be no need to specifically train ML and NLP models.

An important model parameter to handle is its initial weight which is the starting point of the training phase that allows the model to leverage knowledge learned during pre-training. As we chose 8 classes, we will set the initial weight on  $1/8$  by default for each class.

Before tokenizing our reviews, we had to preprocess them by removing stop words. As previously explained, we added the grape variety to the list set up with the classical english stop\_words from the *nlTK* package as it was the target variable. Lower-casing has also been processed before tokenization. Padding and truncation were set to True as well, as the tokenization process cannot exceed 512 tokens by input sequence. Padding is the process of adding special tokens to ensure that shorter sequences have the same length as the maximal one. Truncation, on the other hand, aims to reduce the length of sequences that are too long. The main goal of these two parameters, set to True in this study, is to avoid any problem in the number of tokens created for each input sequence. Fortunately, our reviews do not exceed the maximum length of 512 tokens, so no information had to be lost. After these preliminary steps, our reviews were ready to be tokenized.

Tokenization is the process of splitting paragraphs and sentences into smaller units to which a meaning can be more easily assigned (A. Burchfiel, 2023) [12]. However, before tokenizing our descriptions, it is necessary to partition our dataset into three distinct sets, namely the training set, validation set, and test set. The first one serves to train the model, as the second one is used to validate the training at each epoch while the last one is used to test the model and evaluate its performance. It is common practice in Machine Learning to avoid the model training on the test set. To make the separation, we use the “train\_test\_split” function of the “sklearn” package twice[76]. A proportion of 80/10/10 commonly shared in ML was chosen.

After that, the choice of the tokenizer is essential as diverse variants exist. Overall, the choice of tokenizer can have an impact on the model performance on different NLP tasks, as well as its computational efficiency and memory usage. We started by testing the “bert-base-uncased” classical tokenizer available on “huggingface” as it is the common starting point for fine-tuning a BERT model. We also tried other types of tokenizers such as Roberta or DistilBert, to retain the tokenizer with the best performance as final NLP model.

After the tokenizing process, we could finally compile and fit the model specifically on our dataset. Since a huge amount of parameters are changeable, we decided to keep them by default to get a first global idea of its functioning. In order to facilitate training, we determined that conducting three epochs would be most advantageous as the model would train three times before testing. This choice of three epochs lead to a good balance between execution time and improvement of the results. After each underpass, the model was tested on the validation set.

Finally, predictions were made on our testset to evaluate its accuracy. The results are detailed in section 4.1.

The second model tested in this research follows a statistical machine learning approach, using the power machine learning techniques. Therefore, it uses empirical data only and represents a theoretical upper bound for the previous textual model.

In a concise manner, we followed the same approach as we did for the NLP model. Due to huge diversity in models, presenting different pros and cons, we had to determine which model to choose and why. According to the literature presented, SVM is a good suitor due to its ability to handle many features and the characteristics of the data. We noticed that the different classes presented various combinations in the values of the first four columns, leading to potential clear linear separations. Therefore, as previously detailed in the data analysis section, the representation of varieties in 4-dimensional space could clearly separate the 8 different varieties. The only issue could concern the 635 binary variables since this high amount could impact the linearity of the data. It demonstrates the importance to test models such as Random Forest and ANN, both described as well performing with a high number of features in the literature.

In this regard, Support Vector Machine, Artificial Neural Networks and Random Forest were tested. All of them have been implemented in Python with the “Sklearn” package. [76]. To push them to their limits, we boosted them with the pre-implemented “Grid-Search” function, allowing the best hyperparameters for our dataset to be found. This function runs different combinations of parameters and keeps the combination with the best results.

Concerning SVM, we found out that the combination of hyperparameters  $C = 100$ ,  $\text{gamma} = \text{auto}$  and  $\text{kernel} = \text{rbf}$  was the best for our study.  $C$  is a regularization parameter, controlling the trade-off between achieving a low training error and avoiding overfitting. The smaller the parameter will be, the higher authorized errors on the training set will be and inversely. Gamma represents the complexity of the hyperplane. The smaller the boundary is, the easier and smoother it will be. Kernel specifies the kernel function used to transform the data in a higher dimensional space. RBF, which stands for “Radial Basis Function”, was the best in this case, as we will develop in the experimental results section 4.1.

The hyperparameters have been optimized in the same way for ANN and Random Forest. It was particularly interesting to verify their ability to handle both categorical and numerical features. Their results will also be detailed in the next section 4.1.

Ultimately, the objective of this first step is to explore the impact of switching from structured to unstructured (textual) data on the accuracy of our predictions in the context of wine variety classification. Boosting our models as much as possible is not the goal as our research question concerns the gap between the two types of models rather than boosting them to the very last percentage we’re capable of fetching.

To quantify the impact of switching from quantitative to textual data, we use different different metrics, all available in what we call a classification report.

First, “Precision”. It is defined as the ratio of true positives to the sum of true and false positives. For example, the number of Chardonnay predicted correctly divided by the number of predicted Chardonnay.

Secondly, “Recall”. Recall is defined as the ratio of true positives to the sum of true positives and false negatives. For example, the number of Chardonnay predicted correctly divided by the total number of Chardonnay in our dataset.

Thirdly, “F1-score” is the weighted harmonic average of precision and recall.

Fourthly, “Accuracy” is the number of correctly predicted data out of all observations.

Finally, “Support” is the number of actual observations of the class in the specified testset.

The second part of this work concerns an in-depth evaluation of the models. Except for predicting right or wrong, we also want to know to what extent the recommendations can be false. This relies on the quantification of errors and on distances between classes of wine. This task does not change across models. As varieties are identical, this step will start from the common base of the different models known as, the confusion matrix. To explore these errors, we will focus on the color of the wines. Classifying a white wine as red is indeed worst than classifying it in another white wine class. This granularity within the predictions will be detailed in the evaluation section of this work 4.1.

## 4 Experimental Results

### 4.1 Comparison of the different models

In this section, the results of the different models will be presented and compared so as to answer our research question. As a reminder, the objective is to understand to what extent the current state-of-the-art NLP models are as accurate as empirical models, i.e. handling only quantitative data. The quantitative model may be expected to be better than the textual one. However, the interest lies in the quantification of the loss. This comparison has been then carried out within the framework of a task of predicting the variety of a wine with a common basis, the description of taste. The focus will then also be on the world’s most impressive chatBOT, ChatGPT, to understand to what extent the choice of a trade-off between userfriendliness and technicality influences the accuracy of the predictions. In order to structure this section, we will first compare the different models before adding ChatGPT and its capacity of adaptation and dialogue with humans. This analysis, based on classification reports and confusion matrix, will give insights about the performance of our models. It will also help determine models reactions, as they could respond differently to some characteristics such as the imbalanced dataset.

A model-by-model analysis and a specific SVM without aroma are available in the appendix if required (IV.I, IV.II, IV.III, IV.IV). The first overview of the performances of the empirical models can be described as follows:

|           | SVM  | ANN  | RF   |
|-----------|------|------|------|
| Precision | 0.91 | 0.88 | 0.9  |
| Recall    | 0.9  | 0.88 | 0.87 |
| F1-Score  | 0.91 | 0.88 | 0.88 |
| Accuracy  | 0.93 | 0.92 | 0.92 |

Table 2: Comparison of performance between the quantitative models

To compare the two best-performing models of NLP and ML, it is essential to select them. Results of empirical models were really close to each other but SVM seemed to perform slightly better than the two others when analyzing each metric. Based on this observation, we decided to choose SVM as best performer and representative of the numerical models to carry on with the analysis.

The next step of our comparison is to introduce NLP models. As Transformers proved to be the best ones, we tested some of their variants, BERT base and large, RoBERTa and DistilBERT. Here is the first overview of the results of these variants.

|           | BERT - base | RoBERTa | DistilBERT | BERT - large |
|-----------|-------------|---------|------------|--------------|
| Precision | 0.85        | 0.66    | 0.81       | 0.87         |
| Recall    | 0.79        | 0.62    | 0.74       | 0.80         |
| F1-Score  | 0.82        | 0.63    | 0.76       | 0.83         |
| Accuracy  | 0.87        | 0.76    | 0.82       | 0.87         |

Table 3: Comparison of performance between the NLP models

To be aligned with the quantitative models, we decided to keep BERT-base NLP model, as it seemed to be better than RoBERTa or DistilBERT. The larger version of BERT, BERT - large, presented the same accuracy as BERT-base. However, it was more demanding in terms of computational resources and differences in results were not significant enough to keep it as “best” model in our study.

As a result, our quest to answer our research question will rely on SVM and BERT-base. The comparative analysis leads to the following results:

|           | BERT - base | SVM  |
|-----------|-------------|------|
| Precision | 0.85        | 0.91 |
| Recall    | 0.79        | 0.9  |
| F1-Score  | 0.82        | 0.91 |
| Accuracy  | 0.87        | 0.93 |

Table 4: Comparison between the two best-performing models

At first glance, the latest NLP models seem to be getting closer and closer to the upper bound of structured data models in terms of efficiency, as the change from quantitative to textual data makes us lose only 6% of accuracy. Looking at the different metrics, SVM performs better on every metric. First, Precision. SVM performs slightly better than BERT (6% difference) meaning that the proportion of false positive is a bit higher for BERT. SVM also performs better on the Recall metric. That suggests that SVM is overall 10% better in identifying the true positives. Therefore, this combination of recall and precision produce a larger gap in the F1-score compared to accuracy. This suggests that the models differ in their ability to correctly classify positive and negative instances. It is worth mentioning that the same support was used as both models were tested on the same number of observations.

|              | Precision | Recall | F1-Score | Support |
|--------------|-----------|--------|----------|---------|
| Macro Avg    | 0.91      | 0.90   | 0.91     | 2395    |
| Weighted Avg | 0.93      | 0.93   | 0.93     | 2395    |

Table 5: SVM Classification Report

|              | Precision | Recall | F1-Score | Support |
|--------------|-----------|--------|----------|---------|
| Macro Avg    | 0.85      | 0.79   | 0.82     | 2395    |
| Weighted Avg | 0.87      | 0.87   | 0.86     | 2395    |

Table 6: BERT Classification Report

Overall, we can also conclude that BERT is more impacted by the imbalanced dataset than SVM. The weighted metric is always higher, showing good performance for well represented classes. The gap between macro average metrics and weighted average metrics is always bigger in BERT than in SVM demonstrating that BERT is more sensitive to the imbalanced characteristic of the dataset than SVM.

Upon comparing the performance of SVM and BERT, it is evident that SVM achieves better results overall. This raises the question of knowing whether BERT can reach or surpass the accuracy of SVM for specific wine varieties. To address this question, multiple analytical approaches are available. While a classification report can provide insights by comparing precision, recall, and F1-scores for each variety across both models, we found examining confusion matrices to be particularly enlightening. These matrices not only capture the recall rates but also offer visually accessible means to compare the two models.

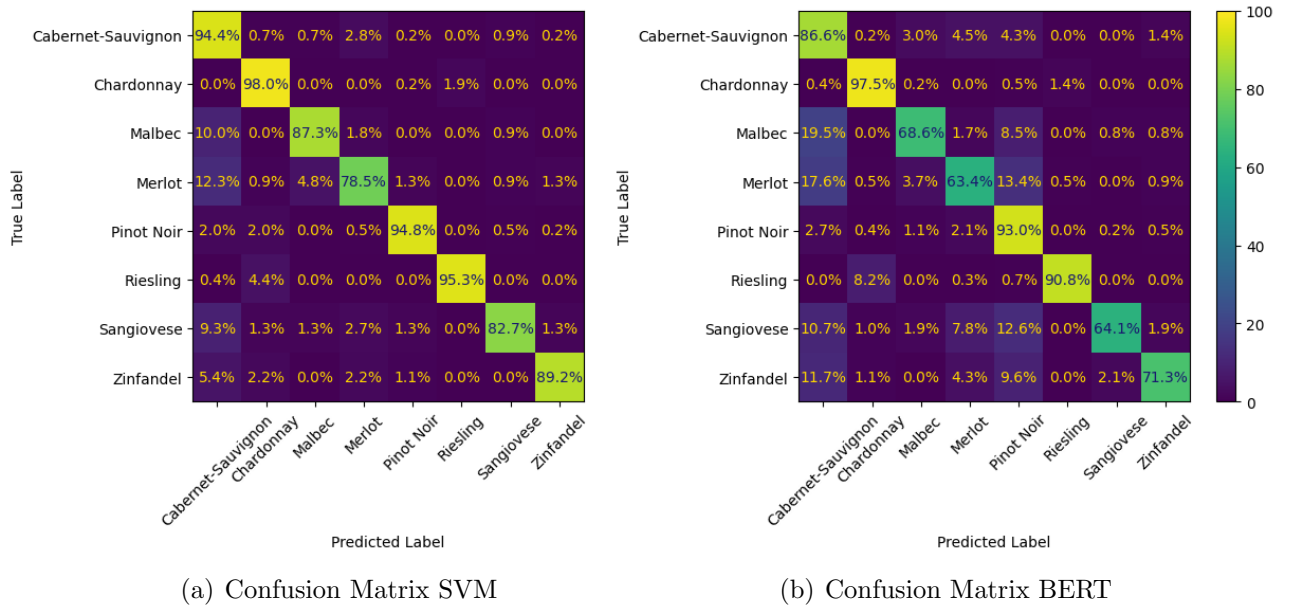


Figure 10: SVM & BERT Confusion Matrices

|                    | 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    |

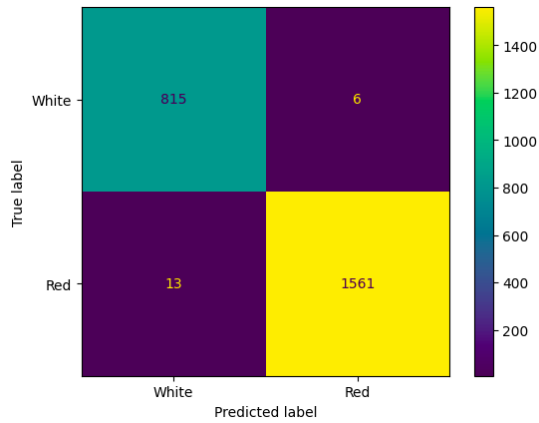
Table 7: BERT Classification Report detailed

|                    | 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    |

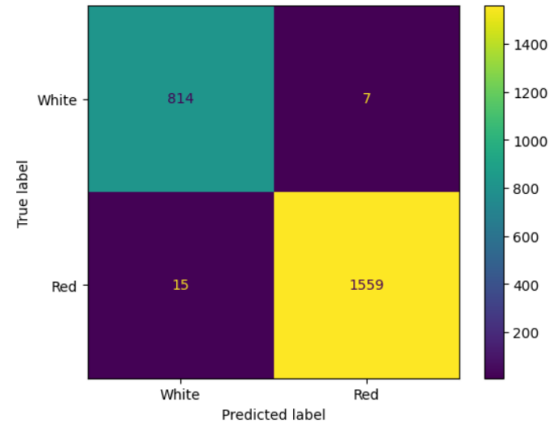
Table 8: SVM Classification Report detailed

An analysis of the confusion matrices reveals that BERT does not surpass SVM for any of the wine varieties. The four varieties with the most significant disparities between the models are Malbec, Merlot, Sangiovese, and Zinfandel, which happened to be the varieties with fewest observations. While BERT displayed the largest deviations for these four varieties, they also consistently performed poorly for both models. Notably, the majority of errors for these varieties were misclassified as Pinot Noir and Cabernet-Sauvignon, the two varieties presenting the largest amount of observations. Determining whether these subpar results stem from our imbalanced dataset or the inherent difficulty in predicting these varieties remains challenging. It is however possible to conclude that during the analysis of the SVM dataset, we had already noticed that wines like Merlot or Malbec did not really have characteristics that distinguished them from the others and moreover, their values were quite scattered. We can therefore imagine that these wines simply present a less extended vocabulary and are therefore more difficult to predict.

Nonetheless, it comes remarkably close concerning Chardonnay, with a difference of only 0.5% - equivalent to one additional error in every 200 predicted wines. This is particularly interesting since it means that under certain conditions NLP is indeed able to achieve results similar to those of the empirical model. A consistent trend across both models is the strong performance of white wines, which were rarely confused with red wines. This realization prompted us to conduct a third comparative analysis, exploring whether BERT can accurately predict the color of wines.



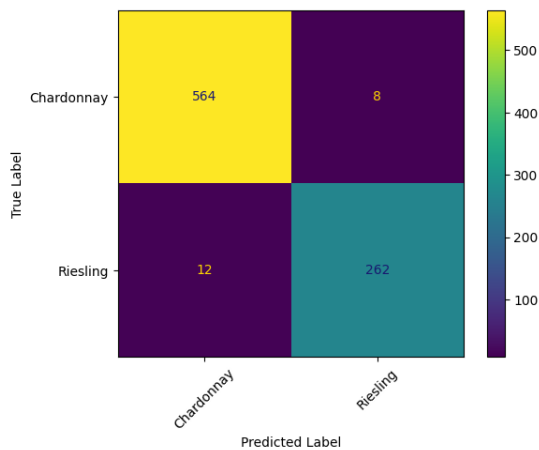
(a) Red & White Confusion Matrix SVM



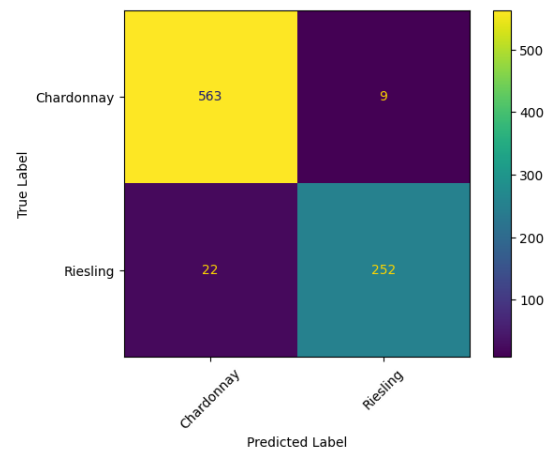
(b) Red & White Confusion Matrix BERT

Figure 11: Red & White Confusion Matrices

The answer becomes evident from the confusion matrices, indicating nearly perfect results for both BERT and SVM in predicting wine color with an accuracy of 99%. Once again, BERT proves its exceptional capability to align with the empirical model in certain situations. Moreover, this implies that the previous disparities observed between SVM and BERT predictions for Chardonnay and Riesling were primarily due to confusions within white wines. Consequently, we aimed to further analyze the predictions exclusively for these two grape varieties to determine if the models performance remains consistent when white and red wines are separated.



(a) White Confusion Matrix SVM



(b) White Confusion Matrix BERT

Figure 12: White wines confusion matrices

As previously observed, BERT approximated the SVM results for Chardonnay prediction very closely. However, a slight difference emerges with Riesling, where BERT exhibits more errors. This disparity may arise from the model being influenced by the imbalanced dataset with Chardonnay being a lot more represented or the model simply being slightly less efficient. Regardless, for white wines, the difference between the two models remains negligible. The advantage of simplicity and speed offered by the textual descriptions might definitely outweigh the difference of performance in these situations.

|                    | F1-Score      |      |             |             |
|--------------------|---------------|------|-------------|-------------|
|                    | All Varieties |      | Red Only    |             |
|                    | SVM           | BERT | SVM         | BERT        |
| Cabernet-Sauvignon | 0.90          | 0.83 | <b>0.92</b> | <b>0.83</b> |
| Malbec             | 0.87          | 0.71 | 0.86        | 0.61        |
| Merlot             | 0.84          | 0.68 | <b>0.88</b> | <b>0.69</b> |
| Pinot Noir         | 0.97          | 0.89 | <b>0.98</b> | <b>0.90</b> |
| Sangiovese         | 0.84          | 0.76 | <b>0.85</b> | <b>0.83</b> |
| Zinfandel          | 0.91          | 0.77 | <b>0.95</b> | <b>0.80</b> |
| Accuracy           | 0.93          | 0.87 | <b>0.93</b> | 0.82        |

Table 9: Comparaison between the models predicting all the varieties and the models predicting only the red ones

Finally, regarding red wines, performances vary quite differently. When white wines are excluded from the evaluation, the majority of varieties show improved results for both models. While SVM maintains consistent accuracy, BERT exhibits more pronounced variations. Some varieties, such as Sangiovese, demonstrate better performance and even get close to the levels achieved by SVM. On the other hand, one variety, Malbec, experiences a significant 10% drop in the F1-Score when evaluated by BERT. Consequently, the overall accuracy of BERT is adversely affected, while SVM remains stable.

In conclusion, our analysis of SVM and BERT models for wine prediction reveals intriguing insights. While SVM consistently outperforms BERT overall, BERT demonstrates its potential to approach SVM's accuracy, particularly for Chardonnay and in predicting wine color. The variations between the model performance for different varieties and the impact of imbalanced datasets require further investigation. Nonetheless, both models show remarkable proficiency, especially in distinguishing between white and red wines, with white wines consistently performing well. Ultimately, the effectiveness offered by BERT's textual description remains particularly good. Even though some deviations may occur, BERT offers accessibility over simplicity to consumers.

## 4.2 Comparison with ChatGPT

In this section, we will compare the performance of our task-specific trained models with ChatGPT, the currently most famous highly sophisticated chatbot (D. Lund Brady, 2023) [54]. Our goal is to understand how close this new tool implemented by OpenAI, based on its massive training on a huge amount of data, can be to a model specifically trained on this complex domain of wine. In this way, we will feed ChatGPT with our reviews in the same way as Zhong did in his study [99]. We will then put it in condition by telling it the eight possible varieties to predict. The ultimate goal will be to produce its classification report and confusion matrix.

Based on 100 reviews per variety each, the results can be presented as follows:

|                    | 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     |

Table 10: F1-Score by variety of ChatGPT

This table perfectly demonstrates unstability in predictions of ChatGPT. Even though some papers quantitatively assess ChatGPT (2023, Zhong)[99], the robustness of these analyses should be questioned. For one of his tasks, Zhong lost almost 5% when testing all data (2023) [99]. In other tasks like ours, these changes could nevertheless have a bigger impact. To sum up, a quantitative analysis of the performance is really risky in a sense that many people could have different results for the same data. In addition, these results depend on the technicity of the prompt as well. In the same paper, Zhong emphasized the importance of mastering the prompt to get better results.

With an accuracy of 61.75%, ChatGPT is indeed less efficient than all previous models. This shows the trade-off to choose between the intrinsic complexity of a model and its ease of use. The performances collected can be summarized as follows:

|          | BERT | SVM  | ANN  | RF   | ChatGPT |
|----------|------|------|------|------|---------|
| Accuracy | 0.87 | 0.93 | 0.92 | 0.92 | 0.62    |

Table 11: Comparison of Accuracy between ChatGPT and other models

ChatGPT therefore performed worse than the other models in the context of wine variety prediction task. However, such task must probably require more training and knowledge than the simple distinction between red or white wine, the first characteristic of the latter. The color of the wine as vague characteristic, could be more in the scope of ChatGPT than the grape variety. Thus, evaluating the open AI tool with a lower granularity of precision could be interesting. Here is its confusion matrix, enabling us to assess its ability to distinguish white from red wines.

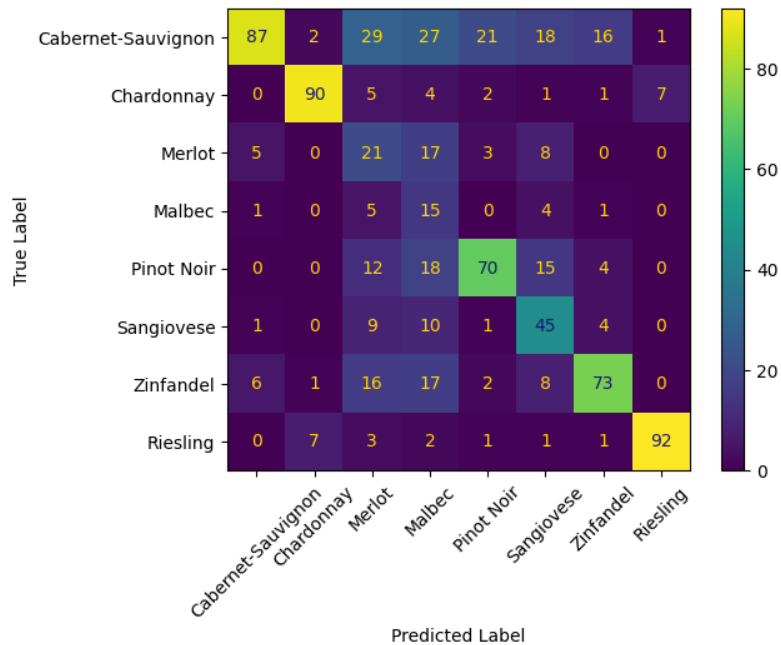


Figure 13: Confusion Matrix of ChatGPT

We can see that ChatGPT performs better at recognizing white wines. This might be due to its extensive training in recognizing and comprehending complex components of human language, such as small details and changes. As a result, determining the color of a wine from its description turned out to be an easier assignment for the model.

Therefore, we attempted to comprehend the reasons behind this pattern. In comparison to the aforementioned models, one advantage of ChatGPT is its ability to provide explanations for the prediction decisions it makes. The existence of certain aromas in the descriptions drove each response given by ChatGPT, which could explain its accuracy in recognizing the color of the wine.

Significant differences were discovered in fragrances associated with white and red wines after analyzing the machine learning model dataset. As previously explained, white wines present variety-specific fragrances compared to red wines. ChatGPT understands the distinct aroma qualities of these wine types, which contributes to its ability to identify them successfully. Furthermore, ChatGPT has inherent knowledge about the varieties of wines, being aware of their colour, which naturally improves its color identification accuracy. Other models, on the other hand, lack this intrinsic information and would have to acquire it on their own, which is necessarily more complex.

Given the restricted number of distinct white types in our dataset, ChatGPT was requested to identify one wine category out of two especially different ones. This may explain its higher performance in correctly identifying them as compared to red wines. The model's prior knowledge of wine varietals, as well as its grasp of the unique scents, can explain its particular advantage in making predictions.

Another surprising observation was made while analyzing the model’s results. A large gap in the quality of the predictions was noticed: While white wines were generally well predicted, some categories of red wines were significantly poorly predicted. For example, while Riesling achieved a fantastic 88% success rate, Malbec and Merlot both fell short of the 30% mark. It is worth mentioning that, unlike the other two models, ChatGPT was not trained on our specific dataset to predict varieties. Therefore, the inferior results cannot be attributed to imbalanced dataset. Although it was not trained on the same dataset, ChatGPT also had more difficulty predicting the varieties that had performed the worst in our two previously analysed models.

Different hypothesis were formulated to explain the difference in accuracy. Two possibilities can be highlighted: either there is an actual similarity between Malbec, Merlot, and the other wine types, which could explain the poor results; or the dataset used to train ChatGPT was also imbalanced. If the first hypothesis was correct, we would predict bad performance in all similar categories. However, in this specific situation, Merlot and Malbec were frequently confused not only with each other, but also with Cabernet Sauvignon or Zinfandel, both of which received highly accurate predictions with the ChatGPT model. It is also worth noting that the drop in F1-scores of Cabernet and Zinfandel was caused mostly by a fall in precision rather than recall. This means that confusion with Merlot and Malbec varieties was often in the same direction, these two varieties being often predicted as Cabernet or Zinfandel while the opposite is not true. This could indicate that the training dataset for ChatGPT may be imbalanced.

This assumption was strengthened by looking at how frequently Chat-GPT predicted each variety. While Cabernet Sauvignon was named 202 times out of 800 descriptions, Malbec was referenced only 27 times. This disparity might show that ChatGPT might have been trained on a dataset that included more Cabernet Sauvignon than Malbec or Merlot. If this hypothesis cannot be further substantiated, the similarity with our models remains surprising.

In conclusion, ChatGPT faced many difficulties while solving our wine prediction task. Too volatile, the tool encountered many limitations. The addition of new varieties and the inability to predict due to a lack of information quickly brought ChatGPT down to an unsufficiently accurate chatbot. Unlike the BERT model, ChatGPT is designed to think and respond like a human in well-defined contexts. In unknown tasks or contexts, the open AI tool seems to struggle. From our observations, it is clear that ChatGPT is not able to perform at the level of new powerful NLP models, especially in Natural Language Understanding tasks as shown by Zhong & al. (2023) [99]. Open AI is, also working on the development of Codex, a new tool that should be able to translate natural language to code [64] which can suggest that these flaws are known from OpenAI. In the future, it might develop to play the role of interface between the human being and more complex models such as BERT. The springboard from ChatGPT to BERT is as a result extremely promising.

Therefore, BERT and ChatGPT could be considered as complementary. On the one hand, BERT is limited in its context understanding and text generation capabilities. On the other hand, ChatGPT masters generating human responses as well as handling long sequences. together they can undoubtedly revolutionize how content is created and processed (P. Paresh, 2023) [67].

In conclusion, we can say that new state-of-the-art AI tools are incredibly powerful for global tasks that do not require deep knowledge, but they still struggle to compete with specific trained models performing specific tasks.

## 5 Evaluation of our models

In this section, we will address the evaluation of our models. The goal is to deepen and quantify the prediction errors of the analyzed models. The objective is therefore to understand whether errors made by our models are major or minor mistakes. In order to do so, it is essential to imagine a way to classify the errors.

First of all, a deep evaluation based on distances as Salih [74] experimented in 2017 is biologically too complex and could be developed in a master’s thesis of its own.

It could also be possible to evaluate our models by looking at how accurately they classify white and red wines. This approach has nevertheless limitations as the support of each variety differs across the two models. This means that a model could be advantaged by easier predictions or, conversely, disadvantaged. However, these differences remain negligible in our specific case. In SVM model, only 0.23% of white wines were mistakenly classified as red while 1.31% of red wines were classified as white. On the other hand, in BERT model, 1.04% of white wines were confused with red ones, but only 0.46% of red wines were mistakenly classified as white. These results are intriguing since it appears that SVM makes more significant errors in classifying red wines, while BERT makes more major mistakes in classifying white wines. Since our dataset mainly consists of red wines, it could explain why SVM has a total of 22 major mistakes compared to BERT’s 16 major mistakes. Furthermore, since SVM makes fewer mistakes overall, it means that the proportion of major mistakes in SVM is much higher compared to BERT. The results can be summarized as follows:

|                              | BERT   | SVM     |
|------------------------------|--------|---------|
| Number of major mistakes     | 16     | 22      |
| Total number of mistakes     | 319    | 164     |
| Proportion of major mistakes | 5.02 % | 13.41 % |

Table 12: Major mistakes comparison

As we can see in Table 12, SVM's total number of errors remains much lower than BERT's. This means that, even if BERT makes proportionally less major mistakes than SVM, the latter remains undoubtedly the best to guess the right class.

After the analysis of what we consider to be major errors, we should also look at what can be considered as minor mistakes. To achieve this, we created dendrograms to show the hierarchical relationship between varieties. To do so, we computed the distances between the different centers of our classes in our quantitative dataset. The results of the deduced dendrograms are the following:

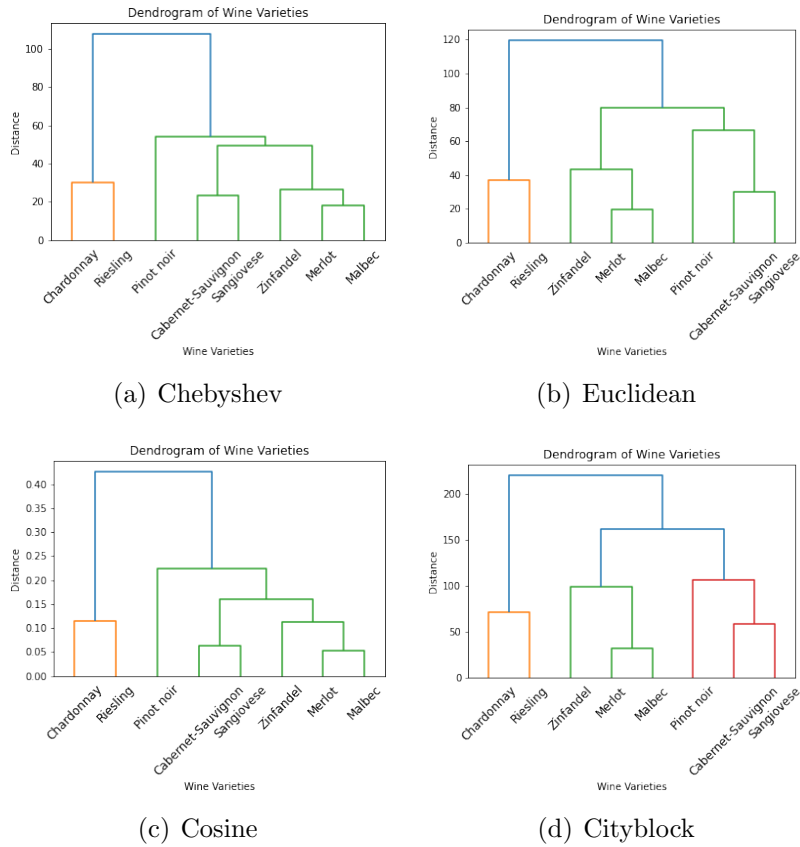


Figure 14: Dendrograms

Pragmatically, the analysis of mistakes could not be based on these dendrograms since they are deduced from our quantitative dataset. It implies that the distances would automatically promote the empirical model compared to the textual one. However, dendrograms could definitely help us to better understand the varieties that are considered as close. As we can observe, three pairs of classes seem to emerge whatever the distance metric used. Among them, the duo of white wines (Chardonnay - Riesling) clearly appears, but we also discover a close similarity between Cabernet-Sauvignon and Sangiovese as well as Merlot and Malbec. These findings have actually been confirmed on the internet [36] [30] by their similarities in taste. As a result, we can easily examine the minor mistakes of our models.

When looking at the confusion matrices in appendix IV.I and IV.II, we can compute that BERT actually made fifty-three “minor” mistakes compared to forty-seven for SVM. There is then only six mistakes of difference. However, we know that BERT actually made more mistakes overall than SVM. Therefore, it was enlightening to compute the ratio of “small” mistakes out of total mistakes. The results can be summarized as follows:

|                            | BERT  | SVM  |
|----------------------------|-------|------|
| Number of “minor” mistakes | 53    | 47   |
| Total number of mistakes   | 319   | 164  |
| % of “minor” mistakes      | 16.61 | 28.6 |

Table 13: Small mistakes comparison

In conclusion, we can say that proportionally SVM made more small mistakes compared to BERT, which is better. This conclusion is therefore opposed to the previous one related to major mistakes. As a result, it is difficult to say which model is the best in terms of errors. Additionally, it is important to note that only a small proportion of the mistakes have been analyzed and taken into account in this analysis (21.63 % for BERT and 42.01% for SVM).

Another method of evaluation could be based on sommeliers' opinions. These experts are indeed the most capable of distinguishing and evaluating similarities and differences between wine varieties. To do so, we surveyed two sommeliers and called upon their expertise in order to create the distance matrix between the different varieties. We asked them to evaluate on a scale of 1 to 10 the similarity between two grape varieties. This data was then used to evaluate our models more accurately.

Here is the average distance matrix built on the basis of the survey:

|                    | Cabernet-Sauvignon | Chardonnay | Merlot | Malbec | Pinot Noir | Riesling | Sangiovese | Zinfandel |
|--------------------|--------------------|------------|--------|--------|------------|----------|------------|-----------|
| Cabernet-Sauvignon | 0                  |            |        |        |            |          |            |           |
| Chardonnay         | 10                 | 0          |        |        |            |          |            |           |
| Merlot             | 10                 | 10         | 0      |        |            |          |            |           |
| Malbec             | 8.5                | 10         | 9.5    | 0      |            |          |            |           |
| Pinot Noir         | 10                 | 10         | 7      | 10     | 0          |          |            |           |
| Riesling           | 10                 | 9.5        | 10     | 10     | 10         | 0        |            |           |
| Sangiovese         | 10                 | 10         | 8      | 10     | 10         | 10       | 0          |           |
| Zinfandel          | 8.5                | 10         | 10     | 8.5    | 10         | 10       | 10         | 0         |

Table 14: Distances matrix based on experts

These expert opinions clearly confirm that these grape varieties are actually extremely different. Although, amateurs might think of some degree of similarity, the diversity of the estate makes these varieties a class of their own, with their own characteristics. In a context where, according to the professionals, these varieties are considerably unique, a calculation of distances would not make sense. At the same time, sommeliers stated that it was almost impossible to create such matrix as two wines from two different classes could be really different or extremely similar depending on the wines. The DNA analysis presented in the literature review was carried out on hybrids derived from the same variety and in the same region. Experts claim that, this can not be extended to our specific varieties. Accordingly, we would limit ourselves to the red and white classification within our models, which makes sense.

## 6 Conclusion, limitations and future work

### 6.1 Conclusion

This section will propose an answer to our research question: 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?

Firstly, building empirical datasets based on a defined framework is often challenging and time-consuming. To quantify data is usually a demanding task that requires lots of investments, time and knowledge. On the contrary, textual data generation offers a more accessible and user-friendly approach. Its ability to streamline the data collection process offers companies increased efficiency, productivity and cost savings. Therefore, embracing textual data offers a practical solution to the challenges associated with building empirical datasets.

Secondly, power of NLP algorithms does not need to envy the power of empirical models anymore. Our key findings show that only 6% of accuracy was lost when switching to textual data. This shows the ingenuity of NLP algorithms, which are now able to analyze human language in depth. At the end of the day, emphasis should then be placed on the gain of practicality that NLP models offer instead of on the gain of accuracy that quantitative models provide. The wine domain being particularly rich in lexical terms, it would be interesting to reproduce this kind of task in a domain with a more limited vocabulary.

Thirdly, ChatGPT has proven its limitations. Although the openAI tool is very impressive, our research has shown the importance of a trade-off between performance and user-friendliness. However, as we have shown, no tool can currently mix the two aspects. The introduction of the famous chatBOT in our research, clearly confirms that, in the future, it will be able to play the role of interface between user and specifically trained NLP models. This comparison also shows the importance of training models on data in a specific manner. Even though BERT is also a pre-trained model, it is definitely its second training process that makes it learn specific patterns linked to this precise data.

Our study also demonstrated complexity in the particular field of wine. A distance between two distinct varieties is actually almost impossible as each variety has its own characteristics. Conversely, we would have to tackle a matrix of distances between particular wines rather than classes.

In conclusion, it seems that in the framework of this study, the choice between an empirical model or an NLP model can no longer be made simply on the basis of accuracy. It would rather be induced by the ability to create quality dataset, whether textual or quantified. Access to one or the other will be the determining factor in the choice of the final preferred model.

## 6.2 Limitations

First, the main constraint of this work concerns datasets. In this study, we decided not to remove observations with missing values, represented by 0's. It is actually impossible to determine whether these are actually missing values or characteristics that a wine simply does not have. For example, the zero tannicity of white wines is actually correct since tannicity is the darkness that a red grape gives to the wine colour. However, all champagnes lack the sweetness scale, that was never mentioned. The number of observations with missing values (0s) in our dataset, assuming that the nul tannicity is correct for white wines can be summarized as follows:

|                    | Cabernet-Sauvignon | Chardonnay | Pinot Noir | Malbec | Merlot | Zinfandel | Sangiovese | Riesling |
|--------------------|--------------------|------------|------------|--------|--------|-----------|------------|----------|
| Number of Outliers | 21                 | 242        | 127        | 4      | 12     | 2         | 7          | 15       |

Table 15: Number of outliers per variety

The higher values for the Chardonnay and Pinot Noir varieties are explained by champagnes as they are the two most used varieties for this type of sparkling wine. The limitation lies in the fact that it is not possible to determine whether these 0 values actually represent reality, or whether this is simply a choice made by the website from which data was retrieved. In the latter case, our dataset and therefore our predictions would not exactly represent reality.

We also made an important assumption about the correlation of our variables with the variety being predicted. Statistical tests could have been performed to keep only the variables useful in this wine prediction task.

Second, for the NLP model, we only used one review per wine, the algorithm being demanding in terms of resources. This may be a limitation to this work since these reviews can be provided by one single sommelier. Therefore, another sommelier could provide another potential review, classifying the wine in another category. The final accuracy of the model could therefore be biased by the sommelier’s knowledge or ability to describe his olfactory and gustatory sensations. This is related to the fairness of our algorithm.

Third, we also decided to keep eight varieties (for the reasons explained in the dataset section). This can be a limit in the scalability of our model. A higher complexity and diversity of wines and wine varieties could significantly ruin the performance of our algorithm. The removal of blended wines as well as rosé may also have been a turning point for BERT. Otherwise, our predictions would definitely have become less accurate as these wines are composed of more than one variety.

Fourth, our models could have been rerun several times to produce a distribution of the accuracy and all metrics. As we are not certain that 80% of training of one model is the same as 80% of another model, the final accuracy could slightly vary. However, as the number of observations was quite high and a 5-folds method was used, we did not expect high variations.

Fifth, another important limitation was the computational resources. Throughout our models and runs, we were limited by the GPU of Google Collab. The time allowed to use a GPU is indeed restricted and we were sometimes blocked. This was especially true in the vectorization process done by BERT. Its running time was about 40h in local compared to 30 min with GPU. Moreover, tuning different parameters was then definitely challenging for us especially due to this running time. Optimizing our algorithms was therefore particularly difficult.

Sixth, another limit could be the website used to extract data. Has it been validated by experts? Is the framework perfectly adapted to define the taste of a wine? Would it be the same data if it came from another sommelier? We cannot confirm these elements.

Last but not least, the accuracy of our models also has to be noted. We could have used different techniques to gain 1 or 2% of accuracy. For example, we could have over-sampled or downsampled our observations to improve accuracy for example. We also noticed that there were some duplicated aromas in olfactive and gustative savors, which may have brought confusion into the ML model. Furthermore, we could have tried some techniques to make sure that the proportions of varieties within the train sets were aligned across the two different models. Finally we could have, we could instructed our model to maximize the weighted accuracy instead of the accuracy to decrease the impact of our imbalanced dataset.

As far as ChatGPT is concerned, our comparisons remains feverish. Compared to BERT, this tool is more volatile and the predictions are less accurate. Our methodology followed the procedures of previous studies from the literature. However, we are of opinion that this methodology is not sufficiently replicable or scientific and leaves room for improvement.

### 6.3 Future work

This study has the potential to open many doors in various fields. It would indeed be interesting to use our work in different applications. With the rise of artificial intelligence and the current performance of NLP models, exploitation of unstructured data could intensify. Among these applications, the food domain or the world of sports, finance or marketing could benefit from a more intensive use of unstructured data and human language. We could imagine bots such as ChatGPT, trained on specific data, to help these businesses to focus on high added value work.

Our comparison may be the first of many as generalization of our findings may be worth investigating. It would be of interest to know if our results remain valid across datasets and real-life applications. The viticural vocabulary being rich, it may not be easily extensible to other domains.

It would be interesting to exploit possible hybrid models. In this work, we have clearly separated NLP and ML. However, it would be equally interesting to know if a hybrid model, using both structured and unstructured data, would provide better results. It is indeed possible that information carried by descriptions would bring something new within the numerical data, which would make the model even more robust. The tricky part remains in how to combine the results provided by the NLP model with the ML models. Finally, it could also be interesting to rate the quality of the textual reviews. In our case, they were provided by domain experts and could therefore be considered of high quality. It would be interesting to observe the behavior of our model with varying data quality.

Last but not least, the deep analysis of our results could be further performed. As we mentioned some techniques already used in the literature to compute genetic distances between grapes, a better analysis of these results could help scientists to get around the drawbacks of the different models. In the particular framework of wines, it was difficult to establish proximity metrics between different varieties. This is why we restrained to the red and white classification. This difficulty actually originated from the significant differences in our wine varieties. According to expert opinions collected through Google Form (III), wines have very limited similarities between two classes. The latter are, according to professionals, immensely different. Their flavours are particularly remarkable which implies that it is impossible for them to evaluate certain proximity, except between two particular wines. As proof, almost every distance was set to 10 on a 1 to 10 scale. However, there might certainly be room for improvement in computing these proximities between subvarieties for example.

In conclusion, we believe that there are still interesting investigations to be carried out within the computerization of taste, Wineinformatics, as well as in many other fields.

## 7 References

- [1] 7 Python Libraries For Web Scraping To Master Data Extraction. Apr. 2023. URL: <https://www.projectpro.io/article/python-libraries-for-web-scraping/625>.
- [2] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. “Transformer models for text-based emotion detection: a review of BERT-based approaches”. In: *Artificial Intelligence Review* 54.8 (Feb. 2021), pp. 5789–5829. DOI: 10.1007/s10462-021-09958-2.
- [3] Tosin P. Adewumi, Foteini Liwicki, and Marcus Liwicki. “Word2Vec: Optimal Hyper-Parameters and Their Impact on NLP Downstream Tasks”. English. In: (2020).
- [4] Ethem Alpaydin. *Introduction to Machine Learning*. World Scientific, Feb. 2019, pp. 1–22. DOI: 10.1142/9789811201967\_0001.
- [5] Veronika Arefeva and Roman Egger. “When BERT Started Traveling: TourBERT—A Natural Language Processing Model for the Travel Industry”. In: *Digital* (2022).
- [6] “automation”. In: May 2023. URL: <https://dictionary.cambridge.org/dictionary/english/automation>.
- [7] *Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation*. URL: <https://beautiful-soup-4.readthedocs.io/en/latest/>.
- [8] Sudip Bhattacharjee et al. “Business and government applications of text mining & Natural Language Processing (NLP) for societal benefit: Introduction to the special issue on “text mining & NLP””. English. In: *Decision Support Systems* 162 (2022), p. 113867.
- [9] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. “A training algorithm for optimal margin classifiers”. In: July 1992. DOI: 10.1145/130385.130401.
- [10] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. DOI: 10.1023/a:1010933404324.

- [11] David S. Broomhead and David J. Lowe. “Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks”. In: *Complex Systems* 2 (Mar. 1988), pp. 321–355. URL: <https://apps.dtic.mil/dtic/tr/fulltext/u2/a196234.pdf>.
- [12] Anni Burchfiel. “What is NLP (Natural Language Processing) Tokenization?” In: *tokenex* (Feb. 2023). URL: [https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/#:~:text=Tokenization%20is%20used%20in%20natural,into%20understandable%20parts%20\(words\)](https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/#:~:text=Tokenization%20is%20used%20in%20natural,into%20understandable%20parts%20(words)).
- [13] Rosario Caballero. “Manner-of-motion verbs in wine description”. In: *Journal of Pragmatics* 39.12 (Dec. 2007), pp. 2095–2114. DOI: 10.1016/j.pragma.2007.07.005. URL: <https://doi.org/10.1016/j.pragma.2007.07.005>.
- [14] Qing Cao, Wenjing Duan, and Qiwei Gan. “Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach”. In: *Decision Support Systems* 50.2 (Jan. 2011), pp. 511–521. DOI: 10.1016/j.dss.2010.11.009.
- [15] Jair Cervantes et al. “A comprehensive survey on support vector machine classification: Applications, challenges and trends”. In: *Neurocomputing* 408 (Sept. 2020), pp. 189–215. DOI: 10.1016/j.neucom.2019.10.118.
- [16] Zuhara Chavez, Jannicke Baalsrud Hauge, and Monica Bellgran. “Industry 4.0, transition or addition in SMEs? A systematic literature review on digitalization for deviation management”. In: *The International Journal of Advanced Manufacturing Technology* 119.1-2 (Nov. 2021), pp. 57–76. DOI: 10.1007/s00170-021-08253-2.
- [17] Bernard Chen. *Wineinformatics*. Springer Nature, Nov. 2022.
- [18] Bernard Chen et al. “Wineinformatics: A Quantitative Analysis of Wine Reviewers”. In: *Fermentation* 4.4 (Sept. 2018), p. 82. DOI: 10.3390/fermentation4040082.
- [19] Anton Chernyavskiy, Dmitry I. Ilvovsky, and Preslav Nakov. *Transformers: “The End of History” for Natural Language Processing?* Springer Science+Business Media, Sept. 2021, pp. 677–693. DOI: 10.1007/978-3-030-86523-8\_{\_}41.

- [20] Alexey Ya. Chervonenkis. *Early History of Support Vector Machines*. Jan. 2013, pp. 13–20. DOI: 10.1007/978-3-642-41136-6\_3.
- [21] Kang Hee Cho et al. “Development of Sequence-characterized Amplified Region Markers for the Identification of Grapevine Cultivars”. In: *HortScience horts* 50.12 (2015), pp. 1744–1750. DOI: 10.21273/HORTSCI.50.12.1744.
- [22] Translated by Content Engine LLC. *Digital transformation, an imperative need in the wine industry*. English. 2023.
- [23] China Info & Comm Tech Grp Corp. *The Development of Natural Language Processing*. Springer, June 2021.
- [24] Ilja Croijmans et al. “Uncovering the language of wine experts”. In: *Natural Language Engineering* 26.5 (Sept. 2020), pp. 511–530. DOI: 10.1017/s1351324919000500.
- [25] Christophe Cruz, Cyril Nguyen Van, and Laurent Gautier. “Word Embeddings for Wine Recommender Systems Using Vocabularies of Experts and Consumers”. In: *Open Journal of Web Technologies* 5.1 (Sept. 2018), pp. 23–30. URL: [http://www.ronpub.com/OJWT\\_2018v5i1n04\\_Cruz.pdf](http://www.ronpub.com/OJWT_2018v5i1n04_Cruz.pdf).
- [26] Sue Curry Jansen. *What Was Artificial Intelligence?* English. Bethlehem, PA: mediastudies.press, 2022. ISBN: 9781951399061;1951399056;9781951399054;1951399064;
- [27] G. De Mets et al. “Sensory quality of wine: quality assessment by merging ranks of an expert-consumer panel”. In: *Australian Journal of Grape and Wine Research* 23.3 (2017), pp. 318–328. DOI: <https://doi.org/10.1111/ajgw.12287>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajgw.12287>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajgw.12287>.
- [28] Giovanni Luigi Di Gennaro, Amedeo Buonanno, and Francesco Palmieri. “Considerations about learning Word2Vec”. In: *The Journal of Supercomputing* 77.11 (Apr. 2021), pp. 12320–12335. DOI: 10.1007/s11227-021-03743-2.

- [29] Stefanie Dunlap. *10 Benefits of Process Automation — Impact*. May 2023. URL: <https://www.impactmybiz.com/blog/advantages-of-process-automation/#:~:text=The%5C%20benefits%5C%20of%5C%20automation%5C%20include,well%5C%20as%5C%20improved%5C%20employee%5C%20morale..>
- [30] Simone Felici. *Sangiovese vs Cabernet Sauvignon — Two Great Tuscan Wines*. June 2017. URL: <https://archivio.letitwine.com/en/sangiovese-vs-cabernet-sauvignon-two-great-tuscan-wines/>.
- [31] Frederic B. Fitch. “Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133.” In: *Journal of Symbolic Logic* 9.2 (June 1944), pp. 49–50. DOI: 10.2307/2268029.
- [32] Alexander L. Fradkov. “Early History of Machine Learning”. In: *IFAC-PapersOnLine* 53.2 (Jan. 2020), pp. 1385–1390. DOI: 10.1016/j.ifacol.2020.12.1888.
- [33] Wei Fu, Eun-Kyong Choi, and Hak-Seon Kim. “Text Mining with Network Analysis of Online Reviews and Consumers’ Satisfaction: A Case Study in Busan Wine Bars”. In: *Information* 13.3 (2022). ISSN: 2078-2489. DOI: 10.3390/info13030127. URL: <https://www.mdpi.com/2078-2489/13/3/127>.
- [34] Prakhar Ganesh et al. “Compressing Large-Scale Transformer-Based Models: A Case Study on BERT”. In: *Transactions of the Association for Computational Linguistics* 9 (Sept. 2021), pp. 1061–1080. DOI: 10.1162/tac1\_a\_00413.
- [35] Thomas Krendl Gilbert, Megan Welle Brozek, and Andrew Brozek. “Beyond Bias and Compliance: Towards Individual Agency and Plurality of Ethics in AI”. In: *arXiv (Cornell University)* (Feb. 2023). DOI: 10.48550/arxiv.2302.12149.
- [36] royal glass. *Malbec vs. Merlot: What’s the Difference? - Royal Glass*. June 2021. URL: <https://royal-glass.com/malbec-vs-merlot/#:~:text=Final%5C%20Notes,%20rustic%5C%20version%5C%20of%5C%20Merlot..>

- [37] Thilo Hagendorff. “The Ethics of AI Ethics: An Evaluation of Guidelines”. In: *Minds and Machines* 30.1 (Feb. 2019), pp. 99–120. DOI: 10.1007/s11023-020-09517-8.
- [38] Jeff Heaton. “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning”. In: *Genetic Programming and Evolvable Machines* 19.1-2 (Oct. 2017), pp. 305–307. DOI: 10.1007/s10710-017-9314-z.
- [39] Iris Hendrickx et al. “Very quaffable and great fun: Applying NLP to wine reviews”. In: Jan. 2016. DOI: 10.18653/v1/p16-2050.
- [40] Elisa L. Hill-Yardin et al. “A Chat(GPT) about the future of scientific publishing”. In: *Brain, Behavior, and Immunity* 110 (2023), pp. 152–154. ISSN: 0889-1591. DOI: <https://doi.org/10.1016/j.bbi.2023.02.022>.
- [41] Mihaela Hnatiuc, Mirel Paun, and Daniil Kapsamun. “IoT Sensors System for Vineyard Monitoring”. In: (2022), pp. 46–51. DOI: 10.1109/ICINT55083.2022.00015.
- [42] Tin Kam Ho. “Random decision forests”. In: Aug. 1995. DOI: 10.1109/icdar.1995.598994.
- [43] Danielle Hopkins et al. “Structured data vs. unstructured data in machine learning prediction models for suicidal behaviors: A systematic review and meta-analysis”. In: *International journal of environmental research and public health* 4 (Aug. 2022). DOI: 10.3389/ijerph.2022.945006.
- [44] Bernardo A. Huberman and Lada A. Adamic. “Growth dynamics of the World-Wide Web”. In: *Nature* 401.6749 (Sept. 1999), p. 131. DOI: 10.1038/43604.
- [45] David Inglis. “Towards a Sociology of Wine: The Power of Processes”. In: *Journal of Cultural Analysis and Social Change* 6 (Jan. 2022), p. 09. DOI: 10.20897/jcasc/11447.
- [46] Kevin Jolly. *Machine Learning with Scikit-Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python*. English. Birmingham: Packt Publishing, Limited, 2018. ISBN: 1789343704;9781789343700;

- [47] Seyed Amir Kasnavi et al. “The effect of kernel selection on genome wide prediction of discrete traits by Support Vector Machine”. In: *Gene Reports* 11 (June 2018), pp. 279–282. DOI: 10.1016/j.genrep.2018.04.006.
- [48] Keehang Kwon. “What is an Algorithm?: a Modern View”. English. In: (2021).
- [49] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. “Deep learning”. In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: 10.1038/nature14539.
- [50] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (Jan. 1998), pp. 2278–2324. DOI: 10.1109/5.726791.
- [51] Els Lefever et al. “Discovering the Language of Wine Reviews: A Text Mining Account”. In: Springer Science+Business Media, May 2018, pp. 3297–3302. URL: <https://biblio.ugent.be/publication/8562718/file/8562720.pdf>.
- [52] Stephen C. Levinson and Asifa Majid. “Differential Ineffability and the Senses”. In: *Mind Language* 29.4 (Sept. 2014), pp. 407–427. DOI: 10.1111/mila.12057. URL: <https://doi.org/10.1111/mila.12057>.
- [53] IAN LOWRIE. “ALGORITHMS AND AUTOMATION: An Introduction”. English. In: *Cultural anthropology* 33.3 (2018), pp. 349–359.
- [54] Brady D. Lund and Ting Wang. “Chatting about ChatGPT: how may AI and GPT impact academia and libraries?” English. In: *Library hi tech news* 40.3 (2023), pp. 26–29.
- [55] Diksha Manaware. “Artificial Intelligence: A New Way to Improve Indian Agriculture”. In: *International Journal of Current Microbiology and Applied Sciences* 9.3 (Mar. 2020), pp. 1095–1102. DOI: 10.20546/ijcmas.2020.903.128.
- [56] Warren S. McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The Bulletin of mathematical biophysics* 5.4 (Jan. 1990), pp. 115–133. DOI: 10.1007/bf02478259.

- [57] Athar Mohammed and Ali Ben Ali. “Survey of BERT (Bidirectional Encoder Representation Transformer) types”. In: *Journal of physics* 1963.1 (July 2021), p. 012173. DOI: 10.1088/1742-6596/1963/1/012173.
- [58] Weisong Mu et al. “Wine recommendation algorithm based on partitioning and stacking integration strategy for Chinese wine consumers”. In: *Italian Journal of Food Science* 34.2 (June 2022), pp. 98–109. DOI: 10.15586/ijfs.v34i2.2209.
- [59] Kevin Murphy. *Machine Learning : A Probabilistic Perspective*. Aug. 2012. URL: <http://cds.cern.ch/record/1981503>.
- [60] E-Commerce Nation. *Vin et digitalisation, l’industrie viticole passe au E-Commerce*. Apr. 2023. URL: <https://www.ecommerce-nation.fr/digitalisation-industrie-viticole/#:~:text=La%5C%20digitalisation%5C%20du%5C%20secteur%5C%20viticole%5C%20gr%5C%C3%5CA2ce%5C%20%5C%C3%5CA0%5C%20vin.co,-Ainsi%5C%2C%5C%20l’association&text=Parmi%5C%20celles%5C%2Dci%5C%2C%5C%20on%5C%20peut,ou%5C%20inclure%5C%20dans%5C%20leur%5C%20colis>.
- [61] Andrew Y. Ng and Michael I. Jordan. “On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes”. In: vol. 14. Jan. 2001, pp. 841–848. URL: <http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf>.
- [62] Luís Oliveira, Rodrigo Rocha Silva, and Jorge Bernardino. “Wine Ontology Influence in a Recommendation System”. English. In: *Big data and cognitive computing* 5.2 (2021), p. 16.
- [63] Stéphane Olivesi. “Le sommelier: Un médiateur, artisan du goût”. English. In: *Politiques de communication* N° 5.2 (2015), pp. 83–134.
- [64] *OpenAI Codex*. URL: <https://openai.com/blog/openai-codex>.
- [65] Walter R. Paczkowski. *Business Analytics*. Springer, Jan. 2023.
- [66] Prudhvi Parne. “Artificial Intelligence & Machine Learning Role in Financial Services”. In: *Advances in Machine Learning* (2021).

- [67] Paresh Patel. “BERT or ChatGPT? The Pros and Cons”. In: *Paresh M Patel* (May 2023). URL: <https://pareshmpatel.com/bert-or-chatgpt-the-pros-and-cons/#:~:text=on%5C%20specific%5C%20tasks.-,While%5C%20BERT%5C%20is%5C%20known%5C%20for%5C%20its%5C%20speed%5C%20and%5C%20ability%5C%20to,and%5C%20has%5C%20higher%5C%20memory%5C%20requirements..>
- [68] John Platt. “Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines”. In: *Microsoft Research Technical Report* (Apr. 1998), pp. 21–. URL: <http://recognition.mccme.ru/pub/papers/SVM/smoTR.pdf>.
- [69] Tomas Pranckevicius and Virginijus Marcinkevičius. “Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification”. In: *Baltic Journal of Modern Computing* 5.2 (Jan. 2017). DOI: 10.22364/bjmc.2017.5.2.05.
- [70] “Predictions of wine ratings using natural language processing”. English. In: *Issues in information systems* (2022).
- [71] Daniel D. Raj and Mr. Karthiban. “Artificial Intelligence in Agriculture: A Literature Survey”. In: 2022.
- [72] David E. Rumelhart, Geoffrey E. Hinton, and Ronald B Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (Jan. 1988), pp. 533–536. DOI: 10.1038/323533a0.
- [73] Yukio Sadahiro. “A method for analyzing the daily variation in the spatial pattern of market area”. In: *Journal of Retailing and Consumer Services* 58 (Jan. 2021), p. 102336. DOI: 10.1016/j.jretconser.2020.102336.
- [74] Maan Salih. “Molecular characterization of nine grapevine varieties cultivated in Salahaldin, Iraq by using RAPD-PCR marker”. In: 22 (Jan. 2017), pp. 1813–1662.
- [75] Iqbal H. Sarker. “Machine Learning: Algorithms, Real-World Applications and Research Directions”. In: *SN computer science* 2.3 (Mar. 2021). DOI: 10.1007/s42979-021-00592-x.

- [76] *scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation*. URL: <https://scikit-learn.org/stable/>.
- [77] Anouch Seydtaghia. “La croissance des données sera exponentielle”. In: (Jan. 2020). URL: <https://www.letemps.ch/economie/cyber/croissance-donnees-sera-exponentielle>.
- [78] Steven Shapin. “The Tastes of Wine: Towards a Cultural History”. English. In: *Rivista di estetica* 51.51 (2012), pp. 49–94.
- [79] David S. Siroky. “Navigating Random Forests and related advances in algorithmic modeling”. In: *Statistics Surveys* 3.none (Jan. 2009). DOI: 10.1214/07-ss033.
- [80] Mitali Srivastava, Rakhi Garg, and P. C. Mishra. “Preprocessing Techniques in Web Usage Mining: A Survey”. In: *International journal of computer applications* 97.18 (July 2014), pp. 1–9. DOI: 10.5120/17104-7737.
- [81] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. June 2008, pp. 151–206. DOI: 10.4324/9780080878096-12.
- [82] Ian Tenney, Dipanjan Das, and Ellie Pavlick. “BERT Rediscovered the Classical NLP Pipeline”. In: May 2019. DOI: 10.18653/v1/p19-1452.
- [83] Patrice This, Thierry Lacombe, and Mark G. Thomas. “Historical origins and genetic diversity of wine grapes”. In: *Trends in Genetics* 22.9 (Sept. 2006), pp. 511–519. DOI: 10.1016/j.tig.2006.07.008.
- [84] Qiuyun Tian, Brittany Whiting, and Bernard Chen. “Wineinformatics: Comparing and Combining SVM Models Built by Wine Reviews from Robert Parker and Wine Spectator for 95 + Point Wine Prediction”. In: *Fermentation* 8.4 (Apr. 2022), p. 164. DOI: 10.3390/fermentation8040164.
- [85] Vapnik V. “Pattern recognition using generalized portrait method”. In: *CiNii Research* (1963). URL: <https://ci.nii.ac.jp/naid/10020952249/>.
- [86] Ashish Vaswani et al. “Attention Is All You Need”. English. In: (2017).

- [87] Richard Vidgen, Sarah Shaw, and David M. Grant. “Management challenges in creating value from business analytics”. In: *European Journal of Operational Research* 261.2 (Sept. 2017), pp. 626–639. DOI: 10.1016/j.ejor.2017.02.023.
- [88] Vin.Co. *PIM vins - référentiel vins - wines data*. URL: <https://vin.co/about/>.
- [89] Alex Wang and Kyunghyun Cho. “BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model”. English. In: (2019).
- [90] Defeng Wang and Lin Shi. “Selecting valuable training samples for SVMs via data structure analysis”. In: *Neurocomputing* 71.13-15 (Aug. 2008), pp. 2772–2781. DOI: 10.1016/j.neucom.2007.09.008.
- [91] Paul J. Werbos. “Beyond Regression : ”New Tools for Prediction and Analysis in the Behavioral Sciences”. In: *Harvard University* (Jan. 1974). URL: <https://ci.nii.ac.jp/naid/10004070196/>.
- [92] *What is Intelligent Automation? — IBM*. URL: <https://www.ibm.com/topics/intelligent-automation>.
- [93] *Wine Enthusiast Magazine — Wine Ratings, Wine News, Recipe Pairings*. May 2023. URL: <https://www.winemag.com/>.
- [94] *Wine Reviews*. Nov. 2017. URL: <https://www.kaggle.com/datasets/zynicide/wine-reviews/code>.
- [95] Winedexer. *Trouvez et achetez les meilleurs vins du marché - Winedexer*. May 2023. URL: <https://www.winedexer.com/>.
- [96] Shuzhe Xu, Chuanlong Zhang, and Don Hong. “BERT-based NLP techniques for classification and severity modeling in basic warranty data study”. In: *Insurance Mathematics & Economics* 107 (July 2022), pp. 57–67. DOI: 10.1016/j.insmatheco.2022.07.013.
- [97] Fengqi Yan, Qiaoqing Fan, and Mingming Lu. “Improving semantic similarity retrieval with word embeddings”. In: *Concurrency and Computation: Practice and Experience* 30.23 (Dec. 2018), e4489. DOI: 10.1002/cpe.4489.

- [98] Suoxiang Zhang. “Sentiment analysis based on food e-commerce reviews”. In: *IOP conference series* 792.1 (June 2021), p. 012023. DOI: 10.1088/1755-1315/792/1/012023.
- [99] Qihuang Zhong et al. “Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT”. English. In: (2023).
- [100] Julian Zubek et al. “Performance of Language-Coordinated Collective Systems: A Study of Wine Recognition and Description”. In: *Frontiers in Psychology* 7 (Sept. 2016). DOI: 10.3389/fpsyg.2016.01321.

## 8 Appendix

All additional annexes, figures and tables related to this thesis are available in the attached document on the DIAL platform.