Wine Quality Prediction based on Chemical Components and Customer Reviews

MSc Research Project
Data Analytics

Swapnil Prakash Chinchwalkar
Student ID: x21106681

School of Computing
National College of Ireland

Supervisor: **Mohammed Hasanuzzaman**

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Swapnil Prakash Chinchwalkar
Student ID:	x21106681
Programme:	Data Analytics
Year:	2022/23
Module:	MSc Research Project
Supervisor:	**Mohammed Hasanuzzaman**
Submission Due Date:	01/02//2023
Project Title:	**Wine Quality Prediction based on Chemical Components and Customer Reviews**
Word Count:	7582
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use another authors' written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Wine Quality Prediction based on Chemical Components and Customer Reviews

Swapnil Prakash Chinchwalkar
x21106681

Abstract

Certification of food product quality is the primary priority of any nation. Each country's inhabitants are advised to utilise products that guarantee quality assurance. The same approach is applicable to the wine segment. By examining the data set, it is possible to extract relevant information about the number of chemical elements in each wine and the opinions of each customer on those wines. In this study, multiple data mining classification methods, including SVC, Naive Bayes, Random Forest, Cat Boost, Gradient Boost, and Multi-layer Perceptron, are applied to the samples of different wines with their qualities necessary for quality certification. Similarly, the customer's response after wine consumption along with their details are composed and with the help of VADER (Valence Aware Dictionary and Sentiment Reasoner) and BERT (Bidirectional Encoder Representations Transformers), the sentiments of the customer are determined. The wine with the best quality is recommended and the algorithms' accuracy is compared. Furthermore, it is determined that the key elements impacting wine quality are volatile acidity, citric acid, alcohol, and sulphate. With varying feature sets, the accuracy is observed in the range of 78% to 88%. This research may be utilised not only as a guide for consumers but also as a resource for wine producers seeking to enhance wine quality.

1 Introduction

Due to wine's potentially negative effects on human health, its quality is of vital importance. Since it was determined that drinking wine has a positive association with the rate volatility in the heart, there has been a little rise in the amount of wine consumed in recent years. This increase is profoundly related to the research Janszky et al. (2005). The wine organisation is a massive market where consumers are susceptible to being deceived about the product's quality. The manual method of determining the grade of wine and assigning labels to indicate various levels of quality takes a significant amount of time and does not provide precise results. Although the majority of the chemicals are the same across the board for the various varieties of wine based on the findings of the chemical tests, the amount of each chemical has a varied degree of concentration for each of the many types of wine. In today's world as stated by Rodriguez Mendez et al. (2016), it is essential to categorise the various wines produced for the purpose of quality control. A system that takes into account characteristics including free sulphur dioxide, chlorides, volatile acidity, citric acid, density, fixed acidity, pH, free sulphur dioxide, total sulphur dioxide, quality, residual sugar, and alcohol, which are indeed required while manufacturing. The quantity of these characteristics that are found in the wine may be used to categorise it into a number of quality sections. On the contrary hand, characteristics such as variety price, description, winery province, points, country, designation, and taster name, may assist to analyse the mindsets of consumers and offering a specific wine to customers based on their comments and suggestions.

The categorisation of response is achieved through the compound score(CS) as:

1. Positive Sentiment - if CS is greater than or equal to 0.5
2. Negative Sentiment - if CS is less than or equal to -0.5
3. Neutral Sentiment - if CS ranges between 0.5 and -0.5

On the other hand, the quality is categorised as Bad, Average, and Excellent based on the score of quality ranging from 0 to 10. The categorisation of quality is accomplished with the help of Quality Score(QS) as:

1. Bad Quality - if QS is greater than 0 and less than 4
2. Average Quality - if QS is greater than 3 and less than 8
3. Good Quality - if QS is greater than 7 and less than 11

Several different classifiers are used in order to bring about the best quality of the wine based on chemical components. These include K-Neighbors Classifier, Gradient Boosting Classifier, Cat Boost Classifier, Random Boost Classifier, Multi-Layer Perceptron, and so on. Similarly, 'VADER' is used to analysis the sentiments of the customers, and 'BERT' is used to recommend the best wine to the customers.

1.1 Overview

- Wine Recommendation using BERT – In the competitive and advanced wine industry of today, the success of a company is wholly reliant on its customers. The level of appreciation a consumer has for a product is solely regarded as the product's level of success. Before making a purchase, every consumer reads the reviews left by previous customers who have experience with the product in question Su et al. (2020). Researchers have spent decades analysing the effect of these evaluations using various ideas. According to this investigation, consumer evaluations have various negative effects. By constructing a fine-tuned BERT technique, it is possible to rapidly identify reviews as it is more useful and least time-consuming. As a consequence, a new thread in emotion analysis studies has been developed. Positive, neutral, and negative sentiments and information will be reviewed to produce sentiment analysis.

- Sentiment Analysis using VADER – Sentiment Recognition is basically the method of 'computationally' determining whether writing is happy, sad, or unbiased. Determining a viewer's perspective or mindset is called opinion mining. VADER is a vocabulary and principle-based sentiment analysis tool suited to digital networking. A sentiment lexicon VADER contains a set of lexical properties (words) categorised as positive or negative that are labeled with respect to emotional orientation. It's open-source and accessible in package NLTK, which can be used directly for unstructured text data. It detects emotion polarity along with intensity and provides sentiment's happiness and unpleasantness score for an individual. It contains a compound rating system that helps to add normalized lexical evaluations between -1 (most negative) and +1. (most extreme positive).

- Wine Quality using various Algorithms – The comprehensive analytical evaluation that was carried out by Tang et al. (2020) looks at a number of models, some of which include Support Vector Machine, CNN, and Naive Bayes. When compared to CNN-based models, the dictionary learning approaches that are based on Deep Learning perform much better than the traditional dictionary learning methods. Constructing a variety of models that surpass CNN in every way except for deep learning approaches would be a fascinating novelty for any individual to try. As a result, several classifiers, including K-Neighbors, Cat Boost, Gradient Boost, and Random Forests, as well as Multi Level Perceptron and Sequential Model, are used in order to identify the grade of wine that is considered to be the finest.

1.2 Motivation

Sales of wine are at record heights, therefore studying consumer habits is more important than ever. Consequently, the purpose of this research is to investigate both the internal motivators(such as personal attributes) that push the consumer into wine goods and the extrinsic motivators(such as situational attributes) that pull the customer toward wine products. Motivated by this goal, one can work to lessen the likelihood that their customers would have negative health consequences as a result of drinking bad wine. When deciding on the most cost-effective option among many different kinds of wine, a customer's reason for drinking wine and checking out the finest wine quality is crucial. This study provides reliable data that assists customers in the modern world in identifying various sentiment analysis methods, which may be enhanced by using feeds from BERT.

1.3 Problem Statement

The increased presence of bogus reviews on many social networking platforms makes it more difficult to trust the platform's users. The utilisation of the BERT approach comes with a number of drawbacks, each one of which has the potential to affect the sentiment classification of bogus comments. According to Zhong et al. (2019) research, considering the qualities of natural language processing might lead to pessimistic perspectives on the lives of people all over the world, which can be improved. The majority of customers in today's world have come to feel that it is important to thoroughly research a product before making a purchase of that thing. The effectiveness of reviews as a technique for improving a company's sales is growing to the point where it is becoming more valuable. Customers are sometimes prevented from purchasing a product because of bogus reviews that offer an inaccurate impression of the product's quality. As a result, it is necessary to identify and eliminate false reviews.

1.4 Research Aim

The purpose of this study project is to establish the value of using VADER to comprehend the sentiments of consumers and the role that BERT plays in assisting customers in selecting the appropriate bottle of wine for their tastes and budgets. In addition, the study intends to make use of a variety of different classifiers in order to examine the appropriate characteristics that, if improved, may lead to an increase in the flavor of wine as well as that either consumers or producers need to pay attention to. Finally, providing the consumer with an accurate assessment as to which method is most effective when selecting and consuming wine of the appropriate quality. This also assists businesses, stakeholders, and manufacturers to realize in time how their budgets are going to be affected by profit or loss depending on the quality of wine being upgraded or degraded as a result of the inclusion of appropriate or inappropriate chemical components.

1.5 Research Objective

1. To create a BERT based on customer feedback to advise wine buyers.

2. Implement VADER based on word and intensity analyzer to discover client emotions.

3. Using BERT's forward and reverse feeds to find the best accurate sentiment categorization algorithm.

4. To investigate the importance of various classifiers on chemical components to identify wine quality attributes.

5. Based on quality, classify wine as excellent, terrible, or neutral.

1.6 Research Questions

1. How can BERT recommend a wine based on consumer feedback?

2. How will VADER analyze consumer review sentiments?

3. How effectively do classifiers determine wine quality's primary factors?

1.7 Outline of Research

As a result of the widespread implementation of digitization, the market share of the wine sector is consistently expanding along with the reliance on customers. By using social networking sites such as Facebook, Amazon, Flipkart, Twitter, and many others, individuals are able to have access to a comprehensive range of choices Su et al. (2020). People, in the course of their hectic lives, are continually acquiring different sorts of manufacturing services from online retailers and providing their opinions on these transactions. The response given by an individual after the consumption of wine is of high importance to all stakeholders, businesses, and sellers. Feedback is provided in the form of both good and bad ways in which individuals have expressed their views in relation to the use of manufacturing services. It is essential to investigate the reviews in order to provide better services and to steer clear of the reviews that are given as feedback in order to improve the company's brand image. The purpose of the study effort will be to outline the numerous ways that may be used to enhance sentiment analysis via the use of fine-tuned BERT in order to investigate the feelings of customers. This will outline various techniques for constructing a fine-tuned BERT that will employ the classifier to do sentiment classification on reviews written by customers. It will assist in making better selections about purchases.

1.8 Paper Plan

The structure of the study is centered on separating the research into chapters. Based on the outline of the report, additional research will be conducted. 1 Introduction of the first chapter pertains to the research effort that includes identifying the study objective, goals, motivation, paper plan, and concerns. 2 The literature review entails the second section of the study that defines the literature review on the particular research in accordance with the generated research questions. Methodology, which is discussed in Chapter 3, identifies the techniques for continuing the investigation following the CRISP-DM process. The 4 chapter views the design definition, followed by implementation 5 and evaluation 6. While the final chapter 7 focuses on establishing the conclusion and scope of Future Work.

1.9 Summary

It may be inferred from the preceding discussion that the introductory chapter 1.1 gives an overview of the general study purpose, goals, and questions. In accordance with the study subject of sentiment evaluation of customer reviews as well as wine grade prediction, the usefulness of the BERT model, VADER, and a number of classifiers are evaluated. The chapter has presented a summary of the significance of the BERT model and the role of wine quality in preventing health risks during consumption, in addition to discussing the impetus for the study and formulating the problem statement 1.3. In the sentiment study of consumer reviews, the BERT model and several classifiers play an important role. The preceding portion of the introduction also gave an overview of the dissertation's structure and objective 1.5, detailing the main study project.

2 Related Work

2.1 Introduction

The review of the literature is a major component of the entire report that provides an evaluation of the relevant literature and research publications. Numerous academic publications discussing the relevance of sentiment classification in the sphere of customer reviews are consulted for the literature study. In addition, the literature study examines methodologies such as deep learning, combining technological approaches, and machine learning capabilities. Significant applications of deep learning and machine learning include language processing, picture classification, and text categorization, which are all real-world problems. The literature review section examines how deep learning and machine learning techniques assist in review detection by employing trained frameworks. The section on literature also discusses the substantial advantages of deep learning algorithms in natural language processing. Many types of research had shown that the quality of alcohol may have an effect on a person's business and health. Following a similar investigation, this publication assists:

1. Using several kinds of classifiers focuses on the most significant chemical components that impact wine quality.

2. With the assistance of a VADER sentiment Analyser, enlighten the wine industry as to whether to raise or lower the price of the wine by understanding why certain

responses for a specific wine are negative.

3. With the support of BERT, propose the finest wine to the consumer based on categorized description and quality.

Using a variety of academic sources, the researcher intends to examine how sentiment classification has been extensively used in to avoid negative evaluations, particularly on major wine industry platforms.

2.2 Comparison of LSTM and BERT for Wine Sentiments

The wine industry is seeing the growth of negative evaluations, which impede effective consumer contact and involvement. The need of avoiding and responding to unfavorable reviews in the commercial sector via the use of sentiment analysis with multiple classifiers and the BERT method is a substantial effort to reduce the negative impacts of customers ' evaluation. The wine industry sells the best quality of wine to the customers but are the customers really happy with the product they are getting? This should be the main aim of all businesses that are trying to boost their gains. The taste of every individual varies with quality. So, one thing should be taken into consideration even if the quality of wine is good, some individuals may not like it as their taste varies as compared to the generic taste. Hence, all the sellers should take into account that if an individual is giving negative feedback for a particular wine then first they should check - if the same feedback they are receiving for the same wine from different customers. If yes, then there are high chances that there is some issue with the taste of the wine and the manufacturers should focus on improving the quality. If not, then the same taste can be carried out with all the other customers with the best quality wine. Thus, the negative feedback of single or two people on the same wine should not affect the wine seller to a large extent.

According to Medhat et al. (2014), a text segment is analysed using algorithms, processing of natural language, or text analytics in order to identify sentiment schools that recommend entity categories and sentiment courses. LSTM networks are intertwined with RNN extensions for the purpose of learning from sequential data specifically developed for conventional RNNs. As a solution to the issue that word embeddings convey more semantic information than sentiment information, the authors of study Fu et al. (2018), suggest a lexicon-enhanced LSTM model. Each word's sentiment embedding may be obtained by training a word classification model with the sentiment lexicon as input. The LSTM model has been used extensively for emotion classification that differentiates between negative and positive client evaluations in the wine industry. The inclusion of CNN layers into the LSTM model aims to develop and extract relevant characteristics from the input data in order to provide sequential predictions. This also leads to improved results and analysis in video identification activity identification and picture labeling. This accelerates the process of anticipating an emotion across a text and improves the use of customer feedback in the wine industry. Both models in the LSTM method of sentiment analysis would aid in categorizing a message in text analytics, as well as analyzing and establishing its relevant significance.

- Is BERT superior to (pre-trained or built from start) LSTM Language approach for Text Classification training?

The solution to this question is reliant on the amount of data being used to fine-tune the model. There are many possible variants of the issue statement being addressed:

1. There are five to ten annotated papers in each lesson. A job of the few-shot zero variety.

2. Each class has between 100 and 200 annotated documents.

3. A modestly sized dataset. One has over a thousand annotated papers in a train set. A huge dataset.

Taking into account that although there are basic LSTMs that can be trained from start for text categorization, there are also BERT-like Language Model pre-trained LSTM variations available. BERTs employ transformers to teach word vectors, while LSTMs are utilized to train and fine-tune language models for classification. There is no way that an LSTM can be trained from start for classification with 1–10 samples per group. Fewer data is available, hence only models with less context may be useful. As BERT beats LSTM on big datasets, it is reasonable to believe that it will do better in this situation (130000 records in our case).

2.3 Research Outline of VADER for Customer Review Sentiment Analysis

In recent decades, it has been evident that opinion-based reactions are helping to transform corporate and public feelings, and that emotions influence the loss and profit of the systems. Opinions are crucial to the vast majority of human actions since they are the primary determinants of behavior. Modern technology adopted from other sectors of food processing, such as milk processing, makes it possible to freely combine fractions derived from wine. This may be advantageous for large-scale winemaking so that mass-market wine types may be produced Fischer (2007). If one has to take a decision, one often seeks the input of others. Every organization and company always seeks consumer and public feedback on its goods and services. Therefore, it is vital to collect and analyse the feedback of consumers. Determining the direction of every response and summarising the thoughts included in them is challenging for the ordinary human reader since each product often comprises a large volume of text expressing viewpoints. Therefore, automated sentiment analysis is required to determine the neutrality score and categorise the evaluations as positive, neutral, or negative. According to the data presented in Bonta and Janardhan (2019), it is evident that VADER significantly outperforms Text Bolb. VADER is a set of lexical characteristics that are optimized for discovering semantics in microblog material. If only sentiment analysis was intended for micro-blog material, and if it requires to be processed quickly, the 0.05 threshold for VADER is a preferable option. VADER also adheres to grammatical and syntactic norms for expressing and enhancing the depth of emotion.

2.4 Research Outline of BERT for a wine recommendation to customers

The implementation of the BERT approaches in sentiment analysis denotes unsupervised customer evaluations based on the computational application of the process of fine-tuning. The creation of the BERT method in sentiment analysis allows the classification of both negative and positive categories of reviews, which are then subjected to hyper-parameter tuning to guarantee the increase of model specificity.

Recent developments in neural network language representation have made it possible to transmit the learned emotional experiences of a huge pre-trained language approach to cascade the processing of natural language (NLA) applications. This technique to transfer learning enhances overall performance on a variety of tasks and is especially advantageous when labeled data are few, making pre-trained language modeling attractive resources for languages with a handful of annotated training instances. The models are assessed on 3 downstream NLP tasks: textual sentence similarity, textual entailment recognition, and named entity acknowledgment. According to the research by Souza et al. (2020), recent advances in language representation using neural networks have made it viable to transfer the learned internal states of large pre-trained language models (LMs) to downstream natural language processing (NLP) tasks. This transfer learning approach improves the overall performance on many tasks and is highly beneficial when labeled data is scarce, making pre-trained LMs valuable resources, especially for languages with few annotated training examples.

According to Tang et al. (2020), the BERT Model's fine-tuning procedure creates more algorithmic input and forecast on the reviews while offering better results to prevent customer reviews in the field of the wine industry. This improves the contact and connection between consumers and the firm, while also assisting investigators with the use of major sentiment classification to eliminate negative sentiment. Customer evaluations include the categorisation of conventional neural networks and diverse convolutional network strategies. The pre-trained algorithms are beneficial for providing a viewpoint on the categorization of texts in real-time, which may be powered by both CNN and RNN models. Text categorization in sentiment analysis is driven by neural networks that aid in enhancing pre-training techniques and a variety of executable algorithms. This sentiment analysis employs the BERT model, which aids in the extraction of semantic representation, thus facilitating the extraction of different underlining elements. As per the research done by Cho et al. (2014), context interaction retrieved data using Bi-LSTM. BERT's pre-trained language model aids in exceeding the relevance of negative evaluations that significantly standardize success and customer happiness in the wine industry. Utilizing the logit model as a standard model for predicting and avoiding public opinion would facilitate the production of more estimates and forecast reports through the application of machine learning and deep learning techniques. Zhong et al. (2019) emphasizes the importance of analysing how the different classifiers and BERT's fine-tuning model assure the identification of customer reviews while creating a robust approach to sentiment analysis.

2.5 Fine Tuning of BERT model with Multi-Model approach

The majority of the text classification papers that can be found on the internet are Bag of Words models paired with some form of Machine Learning model that is often used to solve binary text classification problems. The development of natural language processing (NLP), and especially BERT as well as other multilingual transformer-based models, has made it possible to address an increasing number of text categorization issues. Tensorflow Keras, Hugging Face Transformers, and BERT must be used, however, in order to successfully solve a cross, inter text classification issue.

- Why would one want to utilise Hugging face Transformers rather than Google's very own BERT solution?

Because it is very simple to transit between various models using Transformers, such as ALBERT, GPT-2, BERT, XLnet, and so on. This indicates that you may "just about" replace one model with the other in any code by using a different one.

Li et al. (2020) came to the conclusion that the unique way of predicting customer reviews can be effectively managed by the fine-tuned BERT model. This is also made possible by the multimodal integration ideas, which are responsible for updating the inner nonverbal data as well as the reviews left by clients on the many different wine hubs. This strategy, which makes use of a multi-model framework, aims to improve the efficiency of the operational practices of sentiment classification while also drawing attention to the need to deal with negative reviews on a variety of winery locations Wu et al. (2019). Taking this technique allows for more accurate categorization of the numerous sentences into their respective polarity feelings. The fine-tuned BERT model produces a more accurate and unique inside perspective of the efforts and effects of computational data sets. This also concentrates on many difficult data sets, such as visual representations, emoticons, remarks including sarcasm, and word embeddings. As shown in figure Figure 13, the multi-model BERT consists of Twitter as a sample consisting of a Language Model, Caption Transformer, Sentiment Target, and Multi-Model Tweet. The overvaluation of the BERT model aids in insuring the development of classification approaches for customer reviews. This is the case regardless of the size of the datasets. This assurance covers a variety of major techniques for controlling the unprecedented shift in the text content across a variety of negative reviews from the general public on numerous varieties of wines.

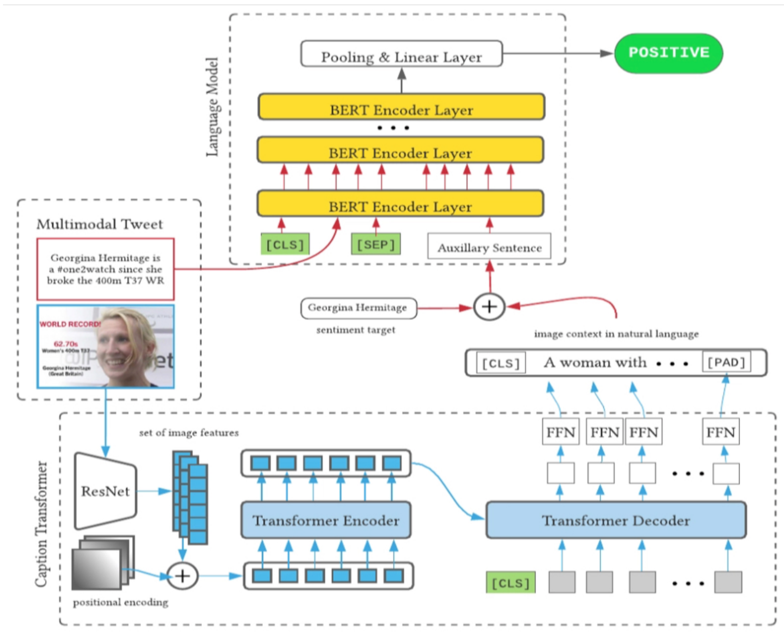


Figure 1: Multi Model BERT Khan and Fu (2021)

2.6 Influencing Factors on the Performance of BERT

The uses of the BERT model, which is significantly altering how consumer reviews are avoided and handled in relation to wine quality, rely on a number of variables. This model's design employs a neural network, which itself is computationally more expensive than any other approach. The usage of several management frameworks and other training contextual frameworks demonstrates how the bidirectional BERT model connects with supervised pattern classification, ELMo, generative contextual representation, and ULMFit. The numerous elements rely heavily on the efficient algorithmic intervention of their properties. Commentary including sarcasm, image representation, and word embeddings are examples of variations that may result from the BERT model's many influencing elements. Using learning approaches, the BERT model assesses the complexity of data sets while conducting many investigations on their effects. According to Sun et al. (2019), the classification of text has little effect on the BERT model's effectiveness in fine-tuning. However, it will be significantly affected by the shift in the text's content that is then subjected to sentiment analysis. Adoption of different algorithmic interventions, such as the optimization of ADA-Boost and ADAM, is believed to facilitate algorithmic diversity as a whole [1]. This sheds light on the important aspect that significantly influences BERT performance while simplifying the available management capabilities with the use of multiple fine-tuning algorithms and location in the sentiment evaluation of customer reviews. Jeewantha et al. (2021) observed that the critical method to establishing the numerous components and their impact on the effectiveness of BERT analysis, decreases the quality of customer review detection. Moreover, Huang et al. (2019) challenges the assertion that these parameters have a direct relationship with market methods and other operational approaches. This model is derived from tiny data sets that assure the categorisation of different deep learning methods in order to apply a major free-trained model for addressing sentiment analysis procedures on customer evaluations.

2.7 Gaps in Literature Review

The researcher has read several academic articles and journals to get data for this study on the importance of the BERT Model in carrying out sentiment analysis. Recently, sentiment research has become more important for any company concerned about providing a satisfying client experience. However, a substantial void within the literature has indeed been discovered, which may indicate that the experimental research papers published concerning the same issue are not very relevant. The following are the gaps:

1. According to the research Shruthi (2019), the future cost of wine may also be predicted using data mining methods to evaluate its quality. The addition of price to various wines may be subdivided into amount categories ranging from 1 to 10. Consequently, each amount range may be classified as poor, medium, or great based on its particular rating and price.

2. The Simulated Annealing (SA) set fared better than the Genetic Algorithm (GA) set, and the SVM performed well compared to other classifiers, according to a study conducted by Kumar et al. (2020). Including more performance metrics and machine learning approaches would improve the comparability of findings.

3. The quantity of samples to be included affects the wine's quality. In the research released by Johar et al. (n.d.), a larger number of samples from various places may assist in determining which extra elements might affect wine quality.

[1] Adopting Algorithmic Interventions: `http://www.eebda.org`.

2.8 Summary

This chapter covers the significance of adopting the BERT approach to assist in doing a sentiment evaluation of customer feedback and recommending the best wine to consumers. Utilizing diverse classifiers such as Cat Boost, Gradient boost, Random Forest, etc. aids in identifying the optimal chemical characteristics to improve wine quality. The need of addressing the bad feelings of consumers in the social attitudes necessitates improved instances of categorizing the public evaluations of wine's diverse remarks. The implementation of this approach further caters to the diverse demands of the procedure as a whole. This chapter illuminates the different aspects that affect BERT performance and examines the multi-modal strategy of the BERT framework as a means of achieving the optimal approach. Diverse algorithmic methods, such as ADAM and Cat-boost, classify the contents of diverse texts while analysing unfavorable evaluations in great detail. This also assures the quality of BERT idea identification while making more attempts to convey the fundamental meaning of customer reviews.

@articleafzaal2019predictive, title=Predictive aspect-based sentiment classification of online tourist reviews, author=Afzaal, Muhammad and Usman, Muhammad and Fong, Alvis, journal=Journal of Information Science, volume=45, number=3, pages=341–363, year=2019, publisher=SAGE Publications Sage UK: London, England

3 Methodology

This chapter covers the methodology used in the implementation of research. Figure 2 and Figure 3 demonstrate the methodological stages that were followed.

Figure 2: Methodology Process for Chemical Components

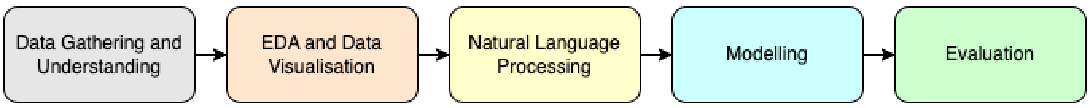


Figure 3: Methodology Process for Customer Reviews

3.1 Data Gathering and Understanding

This is a crucial section since the quality and amount of data collected will directly impact the model's accuracy. The acquired information is subsequently compiled and designated as Training Data. As data, the attribute values that define the quality and kind of wine are provided. Following the collection of training data, the next phase is data preparation, in which the data is imported into an appropriate location and afterward prepared for usage. Additionally, the data is partitioned into two sections. The first component utilised to train our model which will constitute the bulk of the data set, whereas the latter will be utilised for evaluating the performance of the trained model. This section also describes the significance of the research methodologies, centered on the BERT method and its design processes. CRISP-DM is the essential methodology for evaluating the emotions of customer reviews in this investigation. The data sets [2] [3] are taken from the Open Source Kaggle website.

The number of attributes and its value included in both datasets is shown in Figure 4. This paper makes use of two different data sets, each of which is independent of one other. While the 'Customer Reviews' data collection has over 130,000 records and 14 attributes, the 'Chemical Components' data set contains 1,600 records and 12 attributes. Description and quality are, respectively, the target variables that are present in the data. With the use of the Sentiment Intensity Analyzer tool, the customer responses were analysed to determine if the description should be labeled as positive, negative, or neutral.

Sr.No	Attribute Names	
	Customer Reviews Attributes	Chemical Components Attributes
1	*Rank*	*Fixed Acidity*
2	*Country*	*Volatile Acidity*
3	*Description*	*Citric Acid*
4	*Designation*	*Residual Sugar*
5	*Points*	*Chlorides*
6	*Price*	*Free Sulfur Dioxide*
7	*Province*	*Total Sulfur Dioxide*
8	*Region 1*	*Density*
9	*Region 2*	*pH*
10	*Taster Name*	*Sulphates*
11	*Taster Twitter Handle*	*Alcohol*
12	*Title*	*Quality*
13	*Variety*	
14	*Winery*	

Figure 4: Data Set Attributes ()

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is the practice of visualizing information in order to detect and eliminate unnecessary characteristics. It is possible to do exploratory data analysis using both visual and mathematical approaches. It is an essential tool for a data analyst to begin processing the data. It assists analysts in gaining insights from the available dataset. With the use of EDA, an investigator is able to decide which model to employ in an analysis. It also aids analysts in selecting the appropriate approach to use during data analysis. It helps determine the connection between dataset characteristics. EDA may thus be regarded among the most essential techniques in data analysis research.

[2]Customer Review Data Set: `https://www.kaggle.com/datasets/zynicide/wine-reviews`

[3]Chemical Component Data Set: `https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009`

3.3 Natural Language Processing

It is evident from the dataset that the information this research intends to analyse is present in natural language. However, machines do not grasp the context of text on their own. Therefore, a text must be handled such that irrelevant portions are eliminated. The remainder is utilised for subsequent processing. This is referred to as part-of-speech tagging. The processing of natural language consists of the following steps: Loading Packages, Special Characters Removal, Removing stop words, Stemming, Lemmatization and Tokenization.

3.4 Dimension Reduction

Dimensionality minimization is the approach for lowering the under-examined random variables. It is a crucial component of data analysis. Using dimensionality reduction, higher-dimensional data is decreased to a lower dimension, making it simpler for a computer to process. It minimizes the machine's processing needs to process the data. This is accomplished by identifying strongly connected traits. The strongly associated characteristics may then be reduced to a lesser number of features. Dimensionality reduction may be accomplished by a variety of approaches, including

1. The covariance matrix of the data is calculated
2. Eigenvectors of the matrix are calculated
3. The largest eigenvalues given by the eigenvectors are used to reconstruct the large variance feature set
4. Principal Components are used for further analysis

3.5 Modelling and Evaluation

The subsequent stage in the procedure is selecting a model. Choosing the most applicable model for wine analysis. The model employs many ways to give a very effective framework for laying out alternatives and investigating the potential implications of selecting those options. It also helps you establish a balanced view of the related risks and benefit of every potential line of action. The accuracy of the acquired categorization is used to assess the models used in the research. This is accomplished by separating the data set into two groups, viz. Training Set and Testing Set. The data from the training set is utilized to train the models, while the data from the testing set, which serves as un-labeled data for the models, are used for classification. Following data classification, the projected class of the sample data is compared to the real samples to determine the average model accuracy. The result obtained for all the approaches includes metrics such as accuracy, precision, support, and F1 score.

4 Design Specification

This portion of the research examines the usage of the study's models. The research is performed in Python with Jupyter notebook Google Collaboratory(in the case of TensorFlow). Below are the Python libraries used in this study –

1. sklearn - A library for building machine learning algorithm such as Nave Bayes and SVM, as well as associated procedures like as PCA.
2. Keras - A vital library for implementing Neural Networks
3. Transformers - Used to put BERT models into operation
4. NLTK – Textual data processed with Natural Language
5. Numpy - Python package for numerical computations
6. Pandas - Reads the data from dataset into a data framing object
7. Seaborn - Interface for creating intriguing and statistically detailed graphics
8. Lazy Predict - Comparing the effectiveness of several ML models
9. Collections – A container for storing data collections
10. Os - Contains capabilities for communicating well with the operating platform

The architecture of wine quality prediction and customer sentiment analysis is shown in Figure 5 . The process of model architecture is as follows:

1. The dataset is acquired and comprehended in preparation for pre-processing and visualisation.
2. After that, the Wine data sets are separated into train and test parts.
3. The data is then analysed using a designed method to model training.
4. Following that, the training models are assessed by acquiring the outcomes of the categorised data by the produced model.

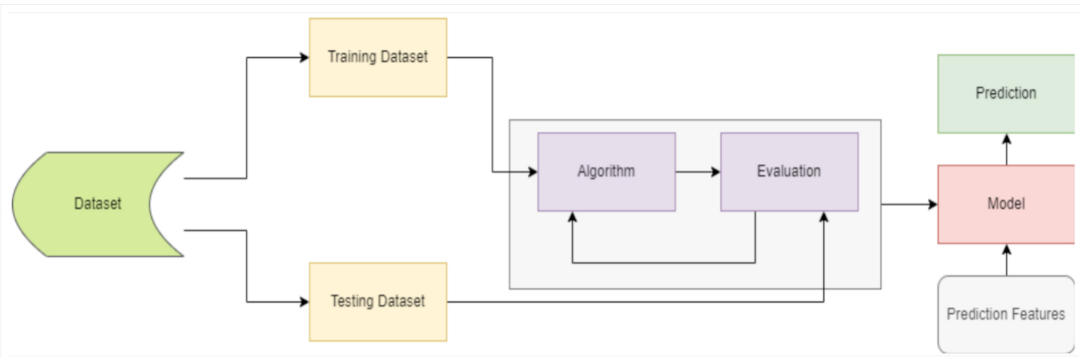


Figure 5: Model Architecture

5 Implementation

To determine whether consumer wine evaluations vary from expert ones, the paper Yang et al. (2022) compares text reviews and quantitative variables for wine rating categorization. It demonstrated that expert wine evaluations are more valuable than age and price in wine data analysis by comparing logistic regression model classification accuracy. Before utilising the dataset for the research, it must be pre-processed. Because the machine learning approach utilised in the research is supervised, the models must have target classes connected with the training data. The sentiment connected with a review is not explicitly included in the dataset. The dataset's "Overall" characteristic, on the other hand, reveals the underlying emotions linked with each review. As a result, this data is used to indicate the feelings of the reviews. The "Overall" characteristic accepts values scale from one to ten. Values 4–7 are considered neutral, values 8–10 are positive sentiments, and the remaining values represent negative emotions. Each sample in the dataset has been given a class in this manner. To determine the correctness of the models, first, the terms in positive and negative reviews are evaluated. The BERT tokenizer is then used to tokenize these words. The tokenized word embeddings are utilised to train the models constructed in the research. For vectoring the words, the BERT tokenizer employs a pre-trained model. Following vectorization, all vectors are normalised using "sklearn's" StandardScaler() method. Due to the large size of these vectors, the research used PCA to determine 13 vectors' properties. Figure 5 displays a flow chart of the processes used in the study to determine wine quality.

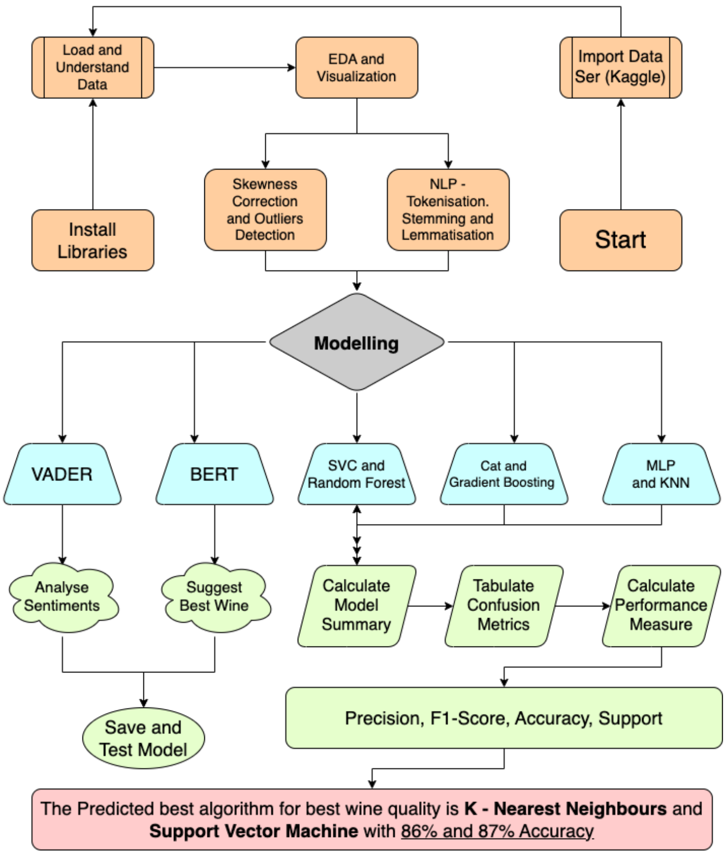


Figure 6: Process Flowchart to determine wine quality

Classifiers were utilized in supervised machine learning for tourist sentiment analysis Waghmare and Bhala (2020).

- Some of the errors that were faced while implementation are as follows:

FileNotFoundError: [Errno 2] No such file or directory: 'WineQualityIsRed.csv' - The 'CSV' file of data was not present in the mentioned directory. With the change in the location of the file to the correct directory the error was eliminated.

ModuleNotFoundError: No module named 'lazypredict' - There was no library installed with name 'lazypredict'. The syntax 'pip install lazypredict' helped to eliminate this error.

ModuleNotFoundError: No module named 'catboost' - There was no library installed with the name 'catboost'. The syntax 'pip install catboost' helped to eliminate this error.

TypeError: init() missing 1 required positional argument: 'units' - The argument init() didn't have any value to pass when instantiating the class. The error is solved by setting a default value and hard-coded value to pass through the argument.

6 Evaluation

This chapter discusses the findings obtained for evaluating the different models used in the research. This research used SVC, Naive Bayes, Cat boost, Gradient Boosting, Random Forest, MLP, KNN and BERT. The results of the modeling are explained below. Research on hotel review sentiment analysis employed Multinomial Naive Bayes (NBM) classifiers Farisi et al. (2019). In that work, the authors used an NBM classifier and a lexicon of terms to retrieve information after data preprocessing to categorize customer evaluations as positive or negative, achieving an average F1 score of almost 85%.

6.1 K-Neighbours Classifiers

This is a technique that makes classifications or predictions on the grouping of individual data points based on their closeness to one another. The data aims in performing multi-task issues, only a single attribute is required to be selected. Hence to adapt spontaneous changes with null assumption, KNN algorithm is best suited for classification regression tasks. Although it may be used for issues involving either regression or classification, it is more often employed as a classification technique since it is based on the concept that points with similar characteristics can be discovered in close proximity to one another. The accuracy achieved is 86.87% which seems to be the highest among all the others. The Figure 7 shows the classification report for KNN.

6.2 Gradient Boosting Classifier

The weak hierarchy of wine data can be enhanced sequentially by combining them with the strong ones in order to increase accuracy. This can be achieved with the help of the Gradient Boosting Classifier, which also helps in providing predictive changes that can never be trumped. These models are a collection of ML algorithms that construct a robust prediction model by fusing together a large number of less capable learning models.

KNN Classification Report

	precision	recall	f1-score	support
0	0.87	1.00	0.93	329
1	1.00	0.04	0.07	52
accuracy			0.87	381
macro avg	0.93	0.52	0.50	381
weighted avg	0.89	0.87	0.81	381

Figure 7: Classification Report for KNN

Because it provides a prediction modeling in the nature of an ensemble model, which is often decision tree, it is used whenever the target item being analysed is binary. This model achieves an accuracy of 62.46% which is the lowest among all the other models used, the Figure 8 shows the classification report.

GradientBoosting Classifier Classification Report

	precision	recall	f1-score	support
0	0.87	0.67	0.75	329
1	0.15	0.37	0.21	52
accuracy			0.62	381
macro avg	0.51	0.52	0.48	381
weighted avg	0.77	0.62	0.68	381

Figure 8: Classification Report for Gradient Boosting

6.3 Support Vector Classifier

In order for SVC to function, data points are first mapped to a high-dimensional space, and then the best hyperplane that separates the data into two classes is determined. SVM, another well-known approach, was employed in a study on hotel review sentiment analysis using supervised training Shi and Li (2011). That study uses SVM with TF-IDF and a collection of terms to analyze sentiment. TF-IDF performed better in experiments. TF-IDF scored 87.2% and bag of words 86.4% on F1. In order to determine the hyper-plane by segregating the data into classes and to map the data into high pointer space, SVC plays a vital role. Data provided by one is properly fitted by SVC and returns the hyper-plane that categorizes the data. It is a C-support vector classification, and the implementation of this classification is based on the library known as 'libsvm'. The name of the module that scikit-learn employs is "sklearn.svm" and the level of accuracy that it achieves is high in comparison to other models: 86.35%. The report on SVC's classification can be seen in Figure 9.

SVC Classification Report

	precision	recall	f1-score	support
0	0.86	1.00	0.93	329
1	0.00	0.00	0.00	52
accuracy			0.86	381
macro avg	0.43	0.50	0.46	381
weighted avg	0.75	0.86	0.80	381

Figure 9: Classification Report for SVC

6.4 Naive Bayes

The Gaussian variant of a Naive Bayes approach is implemented in a grid that determines the classifier setup in a manner that improves the classifier's performance. Paper by Afzaal et al. (2019) showed that NBM properly identified 88.08% of the restaurants' reviews dataset and 90.53% of the hotels' dataset. As the data deals with the classification of texts and includes multiple classes, Naïve Bayes is used with less model training time and it doesn't over fit as the behavior is not complex. It uses cross-validation to determine the optimal hyper-parameters of a classification model in order to achieve optimal performance. This practice is called hyper-parameter optimization. The GridSearchCV() method of the sklearn package is used. The accuracy achieved is 83.46%, Figure 10 shows the classification report for this model.

Naive Bayes Classifier Classification Report

	precision	recall	f1-score	support
0	0.86	0.96	0.91	329
1	0.13	0.04	0.06	52
accuracy			0.83	381
macro avg	0.50	0.50	0.48	381
weighted avg	0.76	0.83	0.79	381

Figure 10: Classification Report for Naive Bayes

6.5 Cat Boost Classifier

Yandex created the open-source CatBoost library. Without pre-processing, this classifier handles categorical variables. This saves time and effort with categorical datasets. Other advantages include Fast training. It translates categorical data into numbers utilizing statistics on categorical and numerical properties. To be more competitive with all the other machine learning models, Cat Boost provides the result in the right form art state. It works very well with categorical attributes and is efficient and ease to use. The accuracy obtained is 83.98% and Figure 11 shows the Cat Boost classification report.

CatBoostClassifier Classification Report	precision	recall	f1-score	support
0	0.90	0.91	0.91	329
1	0.41	0.38	0.40	52
accuracy			0.84	381
macro avg	0.66	0.65	0.65	381
weighted avg	0.84	0.84	0.84	381

Figure 11: Classification Report for Cat Boost

6.6 Random Forest Classifier

This classifier is useful for finding solutions to situations involving either regression or classification. It is composed of a group of decision trees, with each individual tree in the group being built up of a data sample selected from a training dataset with replacement and referred to as the bootstrap sample. The accuracy obtained is 75.06% and Figure 12 shows the classification report for this model.

RandomForest Classifier Classification Report	precision	recall	f1-score	support
0	0.91	0.79	0.85	329
1	0.28	0.52	0.36	52
accuracy			0.75	381
macro avg	0.60	0.65	0.60	381
weighted avg	0.83	0.75	0.78	381

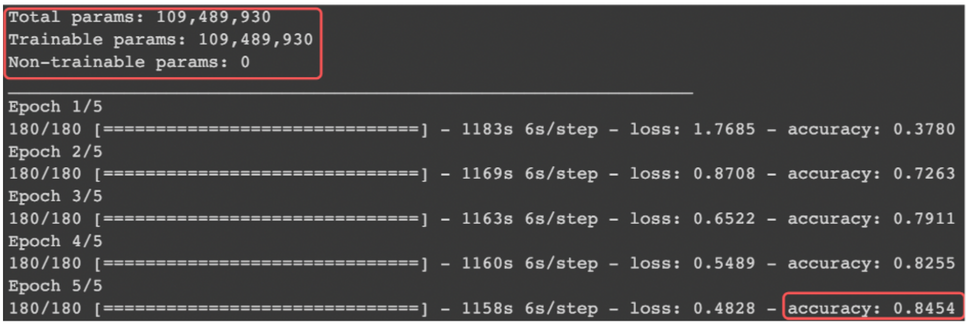
Figure 12: Classification Report for Random Forest

6.7 BERT

An accuracy of 84.54% has been reached with the BERT model as seen in Figure 13. This is shown by the fact that the accuracy improves as the number of epochs grows. In a similar vein, higher validation precision often results in lower validation loss. The research shows that the overall number of parameters analysed is more than 109 million, which is the same as the number of trainable parameters. The process may be characterised in terms of a certain number of epochs, and the amount of non-trainable parameters is zero.

6.8 Discussion

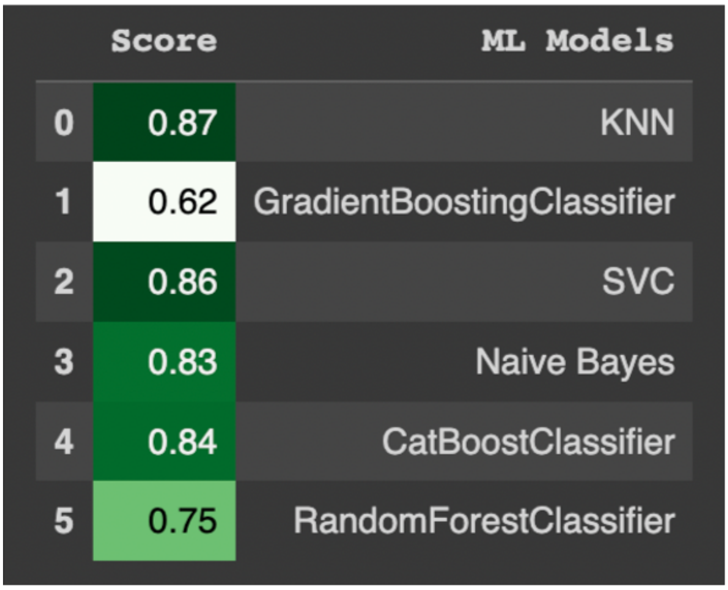
This section discusses the findings achieved from testing the different models used in the research. Figure 14 compares the different algorithms used in this research. KNN and SVC have the best accuracy, but gradient boosting does have the least. With the correct addition of chemical components, the wine's quality is improved, and the buyer will have a clear understanding of which characteristics/samples to concentrate on while drinking



```
Total params: 109,489,930
Trainable params: 109,489,930
Non-trainable params: 0
_________________________________________________________________
Epoch 1/5
180/180 [==============================] - 1183s 6s/step - loss: 1.7685 - accuracy: 0.3780
Epoch 2/5
180/180 [==============================] - 1169s 6s/step - loss: 0.8708 - accuracy: 0.7263
Epoch 3/5
180/180 [==============================] - 1163s 6s/step - loss: 0.6522 - accuracy: 0.7911
Epoch 4/5
180/180 [==============================] - 1160s 6s/step - loss: 0.5489 - accuracy: 0.8255
Epoch 5/5
180/180 [==============================] - 1158s 6s/step - loss: 0.4828 - accuracy: 0.8454
```

Figure 13: Accuracy of BERT with 5 Epochs

wine. Puh and Bagić Babac (2022) also employed TF-IDF and SVM for hotel review sentiment analysis and reached 78% accuracy.



	Score	ML Models
0	0.87	KNN
1	0.62	GradientBoostingClassifier
2	0.86	SVC
3	0.83	Naive Bayes
4	0.84	CatBoostClassifier
5	0.75	RandomForestClassifier

Figure 14: Model Comparison

Figure 15 displays the VADER analysis of customer feedback attitudes. The outcome of the sentiment analysis for the comment "Aromas includes tropical fruits, brooms, milestones, and dry herbs" is positive. Similar analyses were performed on the remaining consumer comments, and based on these analyses, the firm has a clear idea of where they need to take action to enhance the wine's quality.

Another Figure 16 shows the wine suggestions made to consumers by BERT. As seen in the graph, a Pinot Noir wine was suggested based on a customer's remark of "Grapy, plumy, and juicily fruity flavor." Subsequently, the same analysis was performed on the remaining comments; hence, the finest wine would be recommended to each consumer depending on their responses.

```
Aromas include tropical fruit, broom, brimstone and dried herb
sentiment: POSITIVE
Fresh apple, lemon and pear flavors are accented by a hint of
sentiment: POSITIVE
```

Figure 15: Sentiment Analysis using VADER

```
Input text: Strong wine made of red grapes
Recommended variety: Red Blend

Input text: Grapy plummy and juicy taste
Recommended variety: Pinot Noir
```

Figure 16: Wine Recommendation using BERT

7 Conclusion and Future Work

This research concludes that the quality of wine can be increased with the proper addition of chemical components while manufacturing. Traditional methods of feature selection are time-consuming and costly; the feature selection algorithm, on the other hand, gave a clear concept about the value of the characteristics for the prediction of quality. The correlation between quality metrics and other parameters is examined using various visualization tools (e.g., box-plot). The wine samples like citric acid, sulphate level, and volatile acidity are shown to be the most influential aspects of wine quality. The findings show that the SVC is the best method, with an accuracy of 87% following the KNN method, since the dataset splits into train and test data. Additionally, VADER is being used to do consumer sentiment analysis as part of this study. Nonetheless, the count plot reveals an imbalance across classes when using the VADER language to analyse consumers' attitudes. There are more positive comments in this collection than any other kind of sentiment. This is due to the fact that the information does not correspond to the lexical model, since the text has fewer references to feelings and more facts about the wine. Using the quality score obtained from the data, BERT has been effectively applied in this study to propose the appropriate wine to the consumer. It is intended that this study can be utilised by businesses to better anticipate consumer satisfaction and produce higher-quality goods in the future by predicting wine quality based on certain qualities. Categorise discounts to customers if the sales go off in case of bad wine quality. Adding a categorical column of social media(Facebook, Instagram, Twitter) with a username to determine if the wines can be sold online to the relevant customers instead of visiting the location.

References

Afzaal, M., Usman, M., Fong, A. C. and Fong, S. (2019). Multiaspect-based opinion classification model for tourist reviews, *Expert Systems* **36**(2): e12371.

Bonta, V. and Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis, *Asian Journal of Computer Science and Technology* **8**(S2): 1–6.

Cho, H., Kim, S., Lee, J. and Lee, J.-S. (2014). Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews, *Knowledge-Based Systems* **71**: 61–71.

Farisi, A. A., Sibaroni, Y. and Al Faraby, S. (2019). Sentiment analysis on hotel reviews using multinomial naïve bayes classifier, *Journal of Physics: Conference Series*, Vol. 1192, IOP Publishing, p. 012024.

Fischer, U. (2007). Wine aroma, *Flavours and Fragrances*, Springer, pp. 241–267.

Fu, X., Yang, J., Li, J., Fang, M. and Wang, H. (2018). Lexicon-enhanced lstm with attention for general sentiment analysis, *IEEE Access* **6**: 71884–71891.

Huang, Z., Epps, J., Joachim, D. and Sethu, V. (2019). Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection, *IEEE Journal of Selected Topics in Signal Processing* **14**(2): 435–448.

Janszky, I., Ericson, M., Blom, M., Georgiades, A., Magnusson, J., Alinagizadeh, H. and Ahnve, S. (2005). Wine drinking is associated with increased heart rate variability in women with coronary heart disease, *Heart* **91**(3): 314–318.

Jeewantha, H., Gajasinghe, A., Naidabadu, N., Rajapaksha, T., Kasthurirathna, D. and Karunasena, A. (2021). English language trainer for non-native speakers using audio signal processing, reinforcement learning, and deep learning, *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*, IEEE, pp. 117–122.

Johar, S. D. S., JNNCE, S., Ganavi, M. and Nayak, S. N. (n.d.). Analyzing wine types and quality using machine learning techniques.

Khan, Z. and Fu, Y. (2021). Exploiting bert for multimodal target sentiment classification through input space translation, *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, Association for Computing Machinery, New York, NY, USA, p. 3034–3042.
URL: *https://doi.org/10.1145/3474085.3475692*

Kumar, S., Agrawal, K. and Mandan, N. (2020). Red wine quality prediction using machine learning techniques, *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6.

Li, Z., Li, R. and Jin, G. (2020). Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary, *Ieee Access* **8**: 75073–75084.

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, *Ain Shams engineering journal* **5**(4): 1093–1113.

Puh, K. and Bagić Babac, M. (2022). Predicting sentiment and rating of tourist reviews using machine learning, *Journal of Hospitality and Tourism Insights* .

Rodriguez Mendez, M., Preedy, V. et al. (2016). Electronic noses and tongues in food science., *Electronic noses and tongues in food science.* .

Shi, H.-X. and Li, X.-J. (2011). A sentiment analysis model for hotel reviews based on supervised learning, *2011 International Conference on Machine Learning and Cybernetics*, Vol. 3, IEEE, pp. 950–954.

Shruthi, P. (2019). Wine quality prediction using data mining, *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing Communication Engineering (ICATIECE)*, pp. 23–26.

Souza, F., Nogueira, R. and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese, *in* R. Cerri and R. C. Prati (eds), *Intelligent Systems*, Springer International Publishing, Cham, pp. 403–417.

Su, J., Yu, S. and Luo, D. (2020). Enhancing aspect-based sentiment analysis with capsule network, *IEEE Access* **8**: 100551–100561.

Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019). How to fine-tune bert for text classification?, *China national conference on Chinese computational linguistics*, Springer, pp. 194–206.

Tang, H., Liu, H., Xiao, W. and Sebe, N. (2020). When dictionary learning meets deep learning: Deep dictionary learning and coding network for image recognition with limited data, *IEEE transactions on neural networks and learning systems* **32**(5): 2129–2141.

Waghmare, K. A. and Bhala, S. K. (2020). An approach to effective co-referential aspect and entity based sentiment classification of tourist reviews using neural networks and machine learning, *Advanced Innovations in Computer Programming Languages* **2**(1).

Wu, J., Lu, K., Su, S. and Wang, S. (2019). Chinese micro-blog sentiment analysis based on multiple sentiment dictionaries and semantic rule sets, *IEEE Access* **7**: 183924–183939.

Yang, C., Barth, J., Katumullage, D. and Cao, J. (2022). Wine review descriptors as quality predictors: Evidence from language processing techniques, *Journal of Wine Economics* **17**(1): 64–80.

Zhong, G., Zhang, K., Wei, H., Zheng, Y. and Dong, J. (2019). Marginal deep architecture: Stacking feature learning modules to build deep learning models, *IEEE Access* **7**: 30220–30233.