
# DATA MINING REPORT

Introduction to Data Mining

Made by: Sarwesh Balsingh 631859 & Dario Pijman 1658424

Teacher: Mr. W. Ten Hove

Code: DATDRD06

Hand in date: 29-10-2023

# Executive Summary

In the world of Marketing and Sales, customer reviews have become a critical factor that can significantly impact a business's image and future sales. To address this, as Marketing and Sales students, our goal was to explore how Data Mining could be leveraged to analyse customer reviews and solve business-related problems. We used a dataset containing wine reviews, including points, origin country, price, province, region, variety, winery, designation, and customer-written descriptions. Our aim was to create multiple solution designs to compare their performance outcomes.

Our model, grounded in Natural Language Processing, aimed to identify words associated with positive or negative customer reviews. This tool can help companies assess product performance and make informed decisions. For instance, if many customers complain that a wine is too sweet, a company could adjust the product accordingly. Furthermore, automated triggers could alert the marketing department to address negative reviews promptly, potentially enhancing customer satisfaction and loyalty.

Beyond improving product quality, our model can also help companies address biases. We can test hypotheses, such as whether the wine's origin affects its point score. This valuable data can inform decisions about where to produce wine and what consumers prioritize in their choices.

Our research aligns with the CRISP model, encompassing Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. We developed three models:

1. Naïve Bayes Classifier: Categorized reviews into different definitions (e.g., bad, good, very good, excellent) based on point scores. Achieved an accuracy rate of 71.22%.

2. Random Forest Classifier with TF-IDF vectorization: Encountered challenges with accuracy calculation and prediction errors, which led us to downsize the dataset. Initially, it showed a suspicious 100.00% accuracy rate.

3. Naïve Bayes Classifier Multinomial: Simplified the model by classifying reviews as either positive or negative, based on a threshold of 90 points. Achieved an accuracy rate of 86.40%.

In the evaluation phase, Model 3 demonstrated the highest accuracy and versatility, making it our preferred choice for further development. This model can be deployed in real-world scenarios, linking it to review systems or websites, where it can predict the sentiment of consumer-generated reviews.

In conclusion, our journey in developing this model was a valuable learning experience, despite not matching the precision of existing AI models like Chat GPT. We discovered the potential for data-driven Marketing and Sales roles, opening up new career possibilities in the field. In a rapidly evolving digital world, businesses must harness the power of customer reviews to thrive, adapt, and ensure their success.

# Table of Contents

# Introduction

As we are both Marketing & Sales students our goal is to see how we can combine Data Mining and Marketing & Sales to solve a business-related problem. Customer reviews has become a volatile aspect in businesses, it can have an impact on business's images and future sales. This perspective gave us a different view on reviews and interest us, as it is essential for companies.

We are using a dataset with wine reviews. The reviews are based on points: with 80 being the lowest and 100 the highest. It also includes the origin country, price, province, region, variety, winery, designation, and description written by the customer. As we want to challenge ourselves, we are opting for creating multiple solution designs, where it is possible to compare the performance outcomes.

With the use of our model, we can link certain words to a positive or negative customer review. This can be helpful for companies to assess product performance. For instance, if a lot of customers have negative reviews about a product, let's say: the wine is too sweet. It might be an option to change the wine and make it less sweet. It can go even further than this, it is possible for companies to implement a trigger when a certain event takes place. So, for instance when someone places a negative review this can cause a trigger in the system, which the marketing department will have to look for solutions to see if it is possible to satisfy the customer with a different product or even make suggestions to improve the product. This model can save companies a lot of time, money, and effort because it could be fully automated. It improves customer interaction, and it can have an impact on customer loyalty.

The model can also be used to check certain biases. For instance: does the origin of the wine matter? Does wine that come from France or Italy have a higher point score than from Argentina or America? With the model it is possible to test these hypotheses. Companies can use this data to assess where they want to produce their wine and what people favour more. This information can be very valuable, especially to have a better understanding of its consumers and what they prioritize in selecting their wine.

Our purpose and objective are thereby to see if we can make this work for our dataset and to assess if this is applicable in practice. Also, understanding code, knowing how to code and being able to make decisions would be a big benefit for both of us. We want to create the best possible machine that can make such a sort of prediction and hypothesis testing. The machines will be assessed based on CRISP model. In this report we will also conduct a literature review to see the importance of reviews on consumer decisions and how companies must adapt to these changes in the business.

## Literature review

In recent years reviews generated by users have become an important information source for consumer purchase decision making. Thereby it has become valuable for marketers to invest in creating strategies on how to cope with user reviews which eventually leads to word-of-mouth advertising.

There is also a distinction between positive and negative reviews. As negative reviews are often more impactful than positive reviews. “According to accessibility/diagnostic theory, negative information is usually more diagnostic than positive one. Because the product rated to be positive can be in high, average, and low quality, prone to be more ambiguous; On the contrary, negative information strongly suggests inferior performance”. (Cialdini, 1984). When a customer doesn’t know much about a product or service the trust in reviews will increase. On the other hand, if a customer has a high expertise, the reviews might be irrelevant which creates a group that are affected by customer reviews.

Buying decisions are often impacted by physiological factors. One of the theories we can apply is the power of social proof. This means that people tend to follow the actions and opinions of others. Especially when they are not certain about something. This means that businesses with a high number of positive reviews are perceived as more reliable, valid, and trustworthy by their customers. (Cialdini, 1984).

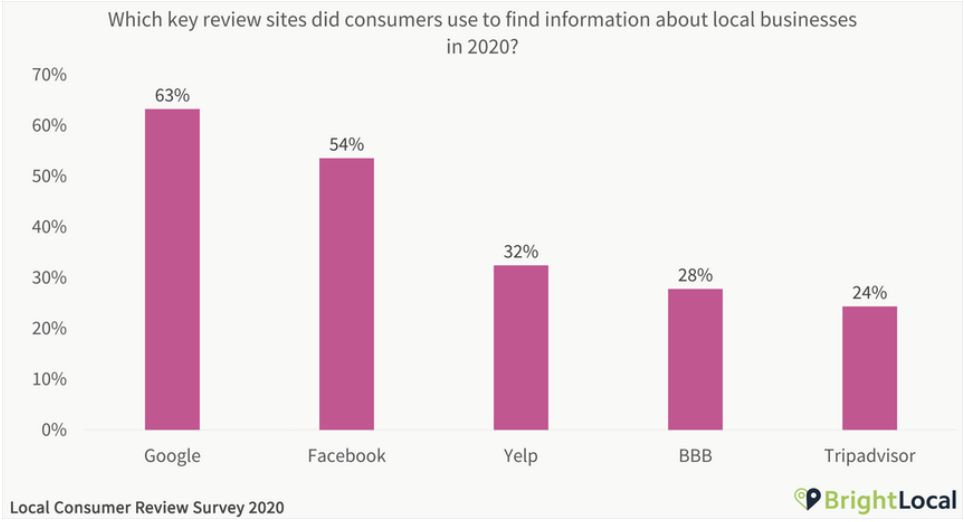
But how reliable are reviews and are some reviews more credible than others? And the simple answer is yes. Customers place more trust in reviews from experts compared to regular users. Also, the sense of sharing the same beliefs and values as the author of the review play a crucial part in how reliable these reviews are. However, perfect scores aren't always a good thing either. It turns out that customers distrust products with no negative reviews and see a product with little negative reviews as more authentic. Little negative reviews are thereby boosting sales, as long as it's a small amount. (Dellarocas, 2003).

The number of reviews also has an impact on the reliability. Customers tend to trust a product with worse scores but a higher amount instead of better scores but a low number of reviews. To further understand this concept ask yourself this: would you rather buy a pair of shoes with a 4.4 out of 5 score with more than 5000 reviews, or a pair of shoes with a 5 out of 5 score but with 20 reviews. I think the choice is obvious if you just look at this. (Dellarocas, 2003).

The time of the review is also crucial for customers to base their decision on. If a review is 5-years old, it is less reliable than one from last month. Businesses should thereby encourage people to leave reviews so that their reviews don't get outdated and therefor unreliable. Because a lot can change over a certain time period. (Chevalier and Mayzlin, 2006)

All these factors play a crucial role in the purchasing decision of customers, and this is just the part of online reviews. It is important for businesses the mechanisms behind customer reviews and how they impact the choice of the customer. The next step would be to decide on how to deal with these reviews effectively to boost sales. As the digital platform continuous to grow customers will leave more reviews and base their decisions more often on existing reviews. Companies that want to survive in this digital world need to adapt and invest into creating strategies to perform better in this environment.

Companies should use multiple platforms for their reviews. While most people look and post on Google, studies show that it is important to use other platforms as well since this increases customer interaction and more people will hear about your company. In the bar chart down below, you can see the review sites consumers use to find information about businesses.




Do note that it says about" local businesses" and it was from 2020 the year of the pandemic. You can clearly see that if a company only has Google reviews it misses out on a lot of other consumers because there are simply no reviews elsewhere. This means that companies should create accounts on as many platforms as possible so that consumers can leave reviews which results in more interactions. More (positive) interactions mean more sales and profit for the company. So, it is important for companies to investigate into this. (Coleman Kate, n.d).

User-generated reviews as an important information source when it comes to purchasing decisions. Companies should manage these reviews to improve sales and encourage word-of-mouth advertising. Companies that want to grow in this digital world need to adapt and invest in strategies that that result in great performances. (Coleman Kate, n.d).

In conclusion, the era of online reviews has opened a new era of consumer empowerment and decision-making. Businesses must harness this phenomenon to their advantage by listening, engaging, and evolving, ensuring that they survive in this digital world that is continuously evolving.

# CRISP MODEL

## Business understanding

The main goal is to being able to implement this model in a business setting. In the world of marketing and sales customer reviews have become a volatile aspect. For companies it is vital to know how to deal with these reviews and how to process this data to make further improvements in their companies and eventually increase their sales output. With the use of the model that we are going to create companies can assess product performance. When a specific result is concluded from the assessment companies can decide to adjust their product or service to improve it. It's also interesting to see if companies can use this model in practise and if it would be helpful to improve their sales.

## Data understanding

We figured out that the reviews are based on a point system with scores from 80 till 100. The score eighty being equal to very bad, and the score hundred being equal to excellent. The data itself has also different other variables, which we can compute more analysis on. The variables are origin country, price, province, region, variety, winery, designation, and description written by the customer.

## Data preparation

### Model 1 Naïve Bayes Classifier

During the data preparation phase, we looked at the complete data set. Here we discovered that one column was meaningless. It had an unnamed column with values from 1 till 150 thousand. Giving each review a numerical order, which is unnecessary because it does not add anything to the data. For the data cleaning we decided to discard this column, in the hope that it won't give us any hiccups during the analysis. After this was sorted, we briefly looked at the data and the kind of reviews. This preparation was done for the first model that was initially build. This would help us decide which machines we want to create and what sort of machines understand sentiment.

### Model 2 Random Forest Classifier TF-IDF vectorization

Because in our last model everything worked perfectly, we assumed that using the same dataset and its variables would have the same effect. Unfortunately, this was not the case. This will be further explained in the modelling section. After finding out this error, we downsized the dataset with the help of Python. The input was the dataset and the Python code made it possible to downsize the dataset from 150 thousand reviews to 75 thousand reviews. In the Python code we expressed that every point, so from 80 till 100, will have an equal distribution of reviews. There are a total of 20 different points, each point has 3,750 reviews in total. This ensured that the data will not be biased towards a type of point score and will make more fair and accurate predictions.

### Model 3 Naïve Bayes Classifier Multinominial

For this code there was not a lot of data preparation. This is because we used the same dataset as model 2 and model 1. The code was tested with both the 150 thousand and 75 thousand reviews, to see if the model would react differently.

## Modelling

The modelling phase was quite an interesting phase. There are several models made to discover which one fits the best for our objective and also what model fits the best with the outcome. What we also wanted to do is hypothesis testing to look at different correlation and see if there is a causation between different variables. There are three hypotheses that we have tested. This will later be explained in more detail. Because of the lack of the coding skills, we consulted Chat GPT for improving and writing the codes. For issues that were not able to be fixed with GPT, we consulted our teachers Witek and Tijmen, as a last resort option.

### Model 1 Naïve Bayes Classifier

The first code we wrote was a Naïve Bayes code. The dataset that is used as input, is the original dataset with 150 thousand reviews. For this code we categorized the points into different definitions to make the machine learn what the different points mean. We categorized greater than 80 equals bad, greater than 85 equals good, greater than 90 equals very good, greater than 95 equals excellent. Based on these categories the machine would learn the most common words in each category. It would then predict if a review would be in one of these five categories. This machine has an accuracy rate of 71.22 percent.

### Model 2 Random Forest Classifier and TF-IDF vectorization

The second code we wrote was a Random Forest Classifier code and TF-IDF vectorization. This is a complex code to predict text classification based on sentiment. The first hiccup we crossed with this code was that the code was calculating constantly and got stuck in a loop. Because it got stuck in a loop the code was not able to calculate its accuracy rate and predict the samples. The reason that this was happening was because the dataset that was initially used was too big for the code to run it (too many variables needed to be calculated). To fix this issue we had to downsize the dataset. As explained in the data preparation we used a Python code to do this. After downsizing the dataset and running the code, the output that got calculated was a 100.00 percent accuracy rate. Our initial reaction was very enthusiastic. Purely, because it took us some time to figure out what the problem was with the code. But we then realised that a 100.00 percent accuracy rate is suspicious. In the evaluation we will discuss this topic further.

### Model 3 Naïve Bayes Classifier Multinominial

The final code that was written is the Naïve Bayes Classifier Multinominial. Because we chose to use a simpler model than the one before, we consulted GPT to write us a code for this. We also decided to change the code format compared to model 1. In model 1 we classified the points to a certain text related feature, for example greater than 80 equals bad and etc. In this code we decided to instead of categorizing it in 4 different categories only use 2. So, a review can either be positive or negative. We learned the machine that in the dataset all the reviews that are equal to and greater than 90 are positive, smaller than 90 is seen as negative. Theoretically this would be easier for the machine to predict because it only has to distinguish the review in either positive or negative. In model 1 the machine had to predict in more categories which makes predicting more difficult. Also, because words can be common in more categories. After the code was finished, we ran the code, and it gave us an accuracy rate of 86.40 percent. This was a big increase compared to the first code.

## Evaluation

As explained in the modelling part, several codes were used and every single one of them we separately tested to see which code fits our objective perfectly.

### Model 1 Naïve Bayes Classifier

After finishing the first code, we added new codes to test the machine and look at the outcome of the predictions. Initially we tested only three reviews:

Review 1: “This wine is absolutely fantastic. It has a rich and complex flavour with a smooth finish.”
Review 2: " This wine is flat and has quite a strong taste. The colour is very dark and has a bitter after taste.”
Review 3: " the best wine in the world.”

The outcome was quite surprising. The machine predicted that two of the three reviews was negative. Review one, two were predicted correctly. Unfortunately review 3 was wrongly predicted, this was a

very surprising outcome. Due to seeing that the review has stated the word “the best”. After this outcome we decided to use a bigger sample. We asked GTP to generate us 15 positive and 15 negative reviews and view the result. Here the machine predicted out of the 30 reviews, 28 correctly. The machine even predicted if it was a good or excellent review. The wrong predictions were only made in the negative sample, the positive sample was fully correct. This outcome was very impressive because this means that from the sample it predicted 94% accurately instead of the initial 71.22%. But of course, in the three test reviews before it predicted 2/3 correctly which equals than to 66.666%. Based on the accuracy that it calculated, we decided to build a different machine. We chose to do this to see if the outcome will be the same or different.

### Model 2 Random Forest Classifier and TF-IDF vectorization

When the code finally worked, and we received an accuracy score of 100.00 percent. The best way to approach this was to test the code with our sample reviews and look at the outcome. We used the same sample reviews as before, provided by GPT, 15 positive and 15 negative reviews. First the negative reviews were loaded in, and the machine predicted all the negative reviews as negative, which sounds excellent. Afterwards we loaded in the sample ‘positive reviews’, the machine predicted all the positive reviews as negative, which was of course incorrect. It was unclear why the machine was making such predictions and still giving the output as 100.00 percent accurate, which was totally not the case. After hours of trying to figure it out, I asked my teacher Tijmen Weber to have a look at the code. Tijmen saw that the variables in the code were two of the same things and that this is causing the problem. So, I am now comparing <95 equals excellent and etc. these variables are exactly equal, and the machine is than predicting based on points instead of based on words. After finding out what the problem with the code was, I conducted GPT on how I could fix this issue. GPT advised me to either change the training dataset, but this could give issues or build a new code. The code we currently were using was quite a complex code for the task it needs to perform. We decided to go chose for creating a simpler machine.

### Model 3 Naïve Bayes Classifier Multinominial

As used in the models before, we tested this machine the exact same way. First, we gave the machine an input of 15 negative reviews, the machine predicted all the reviews correctly. Then we inserted the positive reviews, here the machine only predicted one mistake. This was incredible. We wanted to even go a step further and insert reviews that are not based on wine, but totally other products. So, we gave an input of 5 positive and 5 negative reviews based on shoes. The machine predicted 4 out 5 negative reviews correctly and 4 out of 5 positive reviews correctly. This showed us that the reviews that can be used as input do not have to be wine reviews only.

## Deployment

After the evaluation phase we have decided to continue further with model 3 Naïve Bayes Classifier Multinominial. The reason is that this model has proven to be the most accurate out of all of the models and can even be used for different products and purposes. This machine can be used to link to a review system or website. Were consumers can generate reviews, based on wording the machine will then try to predict if the review is either positive or negative. Ass a cross reference we tried to see how Chat GPT can predict certain reviews. We used the same sample and GPT predicted all the reviews perfect. It even gave a deeper understanding of the review and was even able to predict if a review was moderate or not. Unfortunately, with this type of testing we can see that the current AI is better than our model.

## Conclusion / Learning Journey

The model we build was essentially no match against the current AI models that already exist, like Chat GPT. The current AI models are so smart and accurate. It is nice to see that out model was not awful, and the results were actually great! But compared to what is available at the moment it doesn't stand a chance. The learning journey was something special, a lot of trial and error. Eventually we have made an end product to be proud on especially because this is our first-time coding. The learning goals have been achieved. Making the connection to Marketing was very nice, especially that most coding is focussed on Supply Chain Management or on Finances. This has opened up possibilities for us to pursue a career in the Marketing, but the more data driven aspect. Maybe a Data Scientist specialized in Marketing.

## Personal Reflection

Being a Marketing and Sales student with minimal prior coding experience, this journey has been nothing short of remarkable. The opportunity to create a functional AI model capable of discerning whether a review is positive or negative based on its content has been both fascinating and enlightening. Combining Marketing concepts with advanced Modelling techniques has been a captivating experience, offering a glimpse into the potential of developing similar tools for marketing departments in the future.

Collaborating with my partner, Dario, has been an enriching experience. We effectively divided our responsibilities, with Dario focusing on data preparation, literature review, and the business understanding of our models, while I delved into the intricacies of modelling, evaluation, and deployment. Our teamwork during the creation of the introduction and executive summary was particularly gratifying.

Throughout the modelling phase, I encountered numerous challenges, but with the invaluable assistance of GPT, most of these hurdles were surmounted. For the more complex issues, the guidance of our mentors, Witek ten Hoven and Tijmen Weber, proved instrumental in finding solutions.

What we have achieved is a robust machine capable of versatile applications across various business practices. However, it is evident that GPT's capabilities far surpass our model's. Nonetheless, this experience has unveiled exciting possibilities, motivating me to explore a career path in the data-driven aspects of Marketing, possibly as a specialized Data Scientist in the field. This journey has opened doors to a future rich with opportunities in the realm of data-driven Marketing.

# Bibliography

Chevalier and Mayzlin (2006). The effect of word of mouth on sales. Retrieved September 17, 2023.

Cialdini, R. B. (1984). Influence: The Psychology of Persuasion. Retrieved on September 17, 2023.

Coleman Kate. (n.d). 9 ways online reviews affect your business' reputation. Retrieved on October 1st, 2023, from: https://statuslabs.com/blog/online-reviews

Dellarocas, C. (2003). The Digital 'Word of Mouth' Phenomenon. Retrieved September 17, 2023.

# Appendices

## Appendix 1. GitHub links to codes

In this link you will find the codes created for data preparation, for the modelling, and for the evaluation. The downsized dataset is also uploaded to the GitHub repository.

https://github.com/SRRBalsingh/Data_Mining_Project_2023.git

## Appendix 2. Kaggle link

This is the link where we retrieved the dataset from:

Wine Reviews (kaggle.com)