Big Data in Media Technology

Examining Preferences of Wine Consumers on Social Media

Authors: Mihaela Bakšić, Aina Mas Tena and Alejandro Campayo Fernández

12th of October, 2022

Abstract

Big data has opened many new opportunities for getting insight into customer preferences. This is also true for the wine industry, with people leaving numerous reviews online. This paper examines useful insights into wine consumer's preferences regarding wine taste, country of origin and wine style. This allows producers to align their offer with trends in the wine market, create a better profit margin and expand their customer base. The data was collected from Twitter and Vivino web page. After it was preprocessed and a sentiment label was assigned to it. Adequate sentiment classification of review text allowed for analyzing which properties of wines are liked and whether sentiment towards wine can be determined using information about its taste. The research showed that preferences towards more sweet wines exist, but no reliable predictions about general sentiment towards wine can be made solely on its taste components. In addition to taste, customer preferences also depend on wine origin and wine style.

Index

1 Introduction and related work

Knowing your customer's preference is the key to good business. Be it clothes, cars or video games, producers have been benefiting from implicit and explicit customer feedback for a long time. The age of big data has allowed this type of feedback to enter most markets, including the wine industry. With people leaving reviews of wines on various social networks, such as Twitter, or more dedicated wine-lover web pages, such as Vivino, producers can get insight into the preferences of their consumers.

Since these insights have proven to be essential for the success of many companies, employing data science has become a trend in many industries. Most of them rely on recent advances in sentiment analysis and classification using machine learning techniques (Srinivasarao and Sharaff 2021) after mapping the texts to different vector representations (Ghosal et al. 2020; H. Mishra et al. 2017).
When talking about the analysis of the data, unsupervised techniques (Ghosal et al. 2020; S. Mishra et al. 2017) are vital for obtaining useful results, as annotated data remains somewhat scarce and expensive.Apart from unsupervised techniques, objective testing methods and development of different parameter optimization methods are highlighted as key contributors to inflow of data science into many domains (Syakur et al. 2018).
All this process depends on quality of the data. Scraping the data from the internet is the most common way to obtain large amount of information. Its quality however, may vary depending on the source (Zhao 2017).

The aim of this project is to help Ciù Ciù understand better the preferences of todays consumers. Do they prefer wine from a specific country? What is the optimal sweetness, acidity and intensity of a good wine? What are the main differences between the best-rated wines and the ones that people dislike the most?
Insight into such information can open new opportunities for Ciù Ciù Tenimenti Bartolomei when launching new products or redistributing their wine offer globally. This will allow the company to create a better profit margin and expand its customer base.

2 Data extraction

In this project three different datasets have been used. While one of them (Amazon reviews dataset) has been downloaded from a Kaggle repository, the other two have been scraped from web and social media. Amazon reviews dataset consists of instances of text and a label indicating whether the sentiment of the sentence is positive or negative.
Regarding the two scraped datasets, the first one contains tweets related to the wine industry. These tweets not only consist of text but also all the information acquired with twitter's API. Besides these, features from the text have been extracted and sentiment has been assigned to them.
The second scraped dataset has been acquired from Vivino's webpage. From it, it has been possible to obtain wine reviews as well as much meaningful information related to each bottle. This second option has been considered after the realisation of lack of fit of the Twitter data, which will be explained in the following subsections.

2.1 Amazon reviews

For training the sentiment classification model, the Amazon reviews dataset was used. The dataset is publicly available on Kaggle under the same name. It consists of 3,600,000 training samples and 400,000 testing samples. Each review sample has assigned a positive or negative sentiment, title and review text. For this purpose, the title column was omitted, while text and sentiment columns were used for classifier training.

The dataset preprocessing consisted of the following steps:

1. emoji removal,
2. lowercasing,
3. punctuation removal,
4. stopwords removal,
5. stemming.

Character and emoji removal was performed using regular expressions. The stopword list used is provided by Natural Language Toolkit (NLTK) and PorterStemmer is utilised to perform stemming.

This dataset was chosen for training sentiment classification models due to its size and content heterogeneity, considering that the reviews were collected from more than six million users reviewing over two million products.

2.2 Twitter dataset

In order to obtain reviews from different users on social networks, wine-related tweets were extracted. The Twitter API was used for extraction, which required creating an application and requesting the required permissions. The scraping was carried out by filtering tweets in English, to facilitate their pre-processing and subsequent analysis. Furthermore, only tweets that contain a specific hashtag related to wine were extracted.

The choice of hashtags was made taking into account interesting features for the study and the hashtags recommended by the Vivino website elaborating how to find conversations about wine on Twitter [1].

The following hashtags were selected: #badwine, #bestwine, #goodwine, #sonomachat, #wine, #winechat, #winelover, #winereview, #winewednesday, #worstwine.

The Twitter API only allows extraction of tweets from the last week. For this reason, it was only possible to obtain 18,842 tweets (see Table 1).

Hashtag	Number of Scraped Tweets
#badwine	0
#bestwine	28
#goodwine	6
#sonomachat	3
#wine	16,937
#winechat	0
#winelover	1326
#winereview	132
#winewednesday	410
#worstwine	0

Table 1: Number of scraped tweets for each hashtag

Twitter has a retweet option. For this reason, many of the extracted texts were the same. Duplicates were removed, leaving 6,524 different examples. The texts were homogenised, converted to lowercase and removing of stopwords, punctuation marks, numbers and emoticons was performed.

After that, some features (*country*, *wine type* and *color*) were extracted from the tweets text. This was done using a list of keywords for each feature. If a word contained some of that keywords, it was classified accordingly. The keywords list was defined after exhaustive research, using the most common types of wine and some combinations of words that are frequently used in Twitter.

[1] Vivino: 5 Hashtags To Help You Navigate Wine Conversations on Twitter

brand: 'champagne', 'chardonnay', 'pinot', 'cabernet', 'noir', 'merlot', 'airen', 'sauvignon', 'tempranillo', 'syrah', 'garnacha', 'trebbiano'
color: 'red', 'white', 'rose', 'sparkling', 'redwine', 'whitewine', 'rosewine', 'sparklingwine'
country: 'italy', 'italian', 'spain', 'spanish', 'france', 'french', 'chile', 'chilean', 'australia', 'australian'

2.3 Vivino dataset

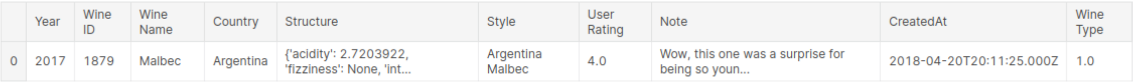
In order to extract wine-related information from a reliable source, more than 150.000 reviews have been scraped from Vivino's webpage. Vivino is an online wine marketplace that collects precise information about millions of different wines and reviews from wine enthusiasts.

The data was scraped using a python scraper publicly available at **github** repository and the code was adjusted to satisfy our means. 100 pages of wines were downloaded and for each different bottle, 4 pages of reviews were considered which constitutes a maximum of 200 reviews per wine. By doing so we obtained more than 150,000 reviews of 1581 different wines from 17 countries. 12 features have been scraped for each bottle. Besides this, one last feature (the sentiment of the review) has been subsequently assigned by a pretrained classifier and added to the dataset.

The extracted wine features contained, among others:

1. wine ID - unique identifier of wine,
2. country - origin country of wine,
3. structure - sweetness, acidity and intensity measurements,
4. note - text of the review

An example entry from the dataset is displayed in Figure 1



	Year	Wine ID	Wine Name	Country	Structure	Style	User Rating	Note	CreatedAt	Wine Type
0	2017	1879	Malbec	Argentina	{'acidity': 2.7203922, 'fizziness': None, 'int...	Argentina Malbec	4.0	Wow, this one was a surprise for being so youn...	2018-04-20T20:11:25.000Z	1.0

Figure 1: Example entry from Vivino dataset

Preprocessing of this dataset was fairly simple as it consisted only of removal of emoticons.

3 Research question and methodology

The main focus of the research is on the following research topic:
"Examining preferences of wine consumers about sort, sweetness, taste and similar properties of wine"

In order to obtain sentiment labels for wine reviews and extract meaningful insights from Vivino and Twitter datasets several methods were used.

The sentiment labels were assigned using supervised learning methods. The data was subsequently processed using unsupervised learning methods and standard practice exploratory analysis methods.

3.1 Supervised learning

In order to produce sentiment labels for reviews scraped from Vivino's webpage, we resorted to using several supervised learning models.

All supervised machine learning models were trained and tested on the preprocessed Amazon reviews dataset.

3.1.1 Naive Bayes Classifier

Naive Bayes classifier (Webb et al. 2010) applies Bayes' theorem with addition of a ”naive” assumption. The ”naive” assumption is that every pair of features is conditionally independent of each other, given a label. That assumption is generally untrue but despite that, the Naive Bayes classifier performs well. The estimations of probabilities are made by a Maximum A Posteriori (MAP) estimator. The assigned label is the label that maximises the MAP estimator.

Naive Bayes is known to be a fairly decent classifier with a short and simple training and prediction process.

3.1.2 Support Vector Machine Classifier

The Support Vector Machine (Suthaharan 2016) is an algorithm focused on finding a hyperplane in an n-dimensional space that distincts datapoints to two classes and has a maximum margin.

For this model, multiple kernels, such as linear, polynomial and RBF (radial basis function), can be used. Depending on the kernel, several other hyperparameters have to be optimized.

3.1.3 Gradient Boosting Classifier

Gradient boosting (Natekin and Knoll 2013) is an ensemble machine learning algorithm. It makes predictions based on predictions of several weak learners, that are typically shallow decision trees. Weak learners are added iteratively, with goal of correcting the residual between the expected output and output that the current iteration ensemble produced. In most of the cases, gradient boosting outperforms random forest classifier.

We used an implementation of gradient boosting classifier named XGBoost (eXtreme Gradient Boosting) (Chen and Guestrin 2016), that is an optimized, distributed gradient boosting library.

3.2 Unsupervised learning

To analyse Vivino and Twitter reviews and find correlations between wine properties and sentiment analysis of the reviews, an exploratory analysis was conducted using unsupervised techniques.

3.2.1 K-means

K-means (Al-Daoud et al. 1995) is an unsupervised machine learning algorithm that clusters similar data points to discover underlying patterns. The number of clusters it finds is a hyperparameter that must be defined before running the algorithm. Unfortunately, there is no strong analytical method for tuning number of clusters, and one must resort to visualization methods such as elbow and silhouette method.

The algorithm starts by selecting k random points that are the initial centroids of each cluster ($k = numberofclusters$). Then it performs iterative (repetitive) computations to optimise the centroid positions until the centroids stabilise or a number of iterations is reached. The remaining observations are assigned to the cluster with the closest mean, minimising the within-cluster variances.

3.2.2 PCA

Principal component analysis (PCA) (S. Mishra et al. 2017) is a statistical technique used in exploratory data analysis and for making predictions with large data sets. It projects the data into the first few principal components to obtain lower-dimensional data while preserving as much of the data's variance as possible. The first principal component is the direction that maximizes the variance of the projected data.

As it reduces the dimensionality of the data by transforming it into a new coordinate system, it is widely used for the visualisation of multidimensional data.

3.3 Exploratory data analysis methods

In order to gain valuable insight about wine preferences with regards to wine taste, structural components, country of origin and similar, we resorted to using standard exploratory analysis statistical methods. This is comprised of:

- descriptive analysis of data - analysis of means, standard deviations, unique values (Kemp et al. 2018; Lawless and Heymann 2010)
- Kolmogorov-Smirnov test for data normality (Massey 1951)
- Lilliefors test for affiliation with a particular distribution (Williams 1984)
- Nonparametric Mann-Whitney U test for equality of measures of the central tendencies (MacFarland and Yates 2016)
- visualization and plotting techniques
- Principal component analysis (PCA) (Section 3.2.2)
- Spearman's rank correlation coefficient and testing Gauthier 2001).

4 Tools and environment

The project implementation was carried out using Python, Jupyter notebooks and Conda environment. Code was executed on Kaggle, an online community platform for data science and machine learning.

We used Python scikit-learn and XGBoost libraries for machine learning models, evaluation metrics and train/test dataset generation.

5 Assigning sentiment

The objective of this project was to analyse whether customers were happy or not with the purchase of a bottle of wine. In order to do that, sentiment was assigned to all Tweet texts and Vivino review texts from the aforementioned datasets.

We tried different representations of the sentences from the Amazon review dataset and trained different models on them. Out of those, we kept the best performing model and used it to predict the sentiment of Twitter and Vivino reviews.

5.1 Text encoding schemes

In order to train our classifiers, an adequate representation of the sentences was needed first. The classifier models that have been used in this project needed a vector as their input. This vector has been acquired using two different techniques, both of them are available in sklearn.feature_extraction module.

1. TF-IDF: Takes into account the frequency of each word (term frequency) in one sentence compared to the frequency of the word in the whole dataset (inverse document-frequency). After, it computes a weight for each term that together with the rest of the words constitutes a vector representation for each sentence. With regards to the parameters, we chose 5000 features to be extracted for each text.
2. Count vectorizer: Tokenizes all the words in our corpus and then it counts the word occurrences returning a vector of length same as the training vocabulary size. In sum, each position reffers to a specific word in the vocabulary and its value indicates the amount of time each word appears in the sentence.

Both methods were fitted and performed on the Amazon reviews dataset and subsequently applied to Twitter and Vivino reviews.

5.2 Classifier training

Considering that the Amazon reviews dataset is fully balanced, the primary metric chosen for determining the best-performing model is accuracy, while for the same accuracy models, the distinction was made using the F1 metric. The Amazon reviews dataset is already split into the training and testing datasets. However, the order of examples has been shuffled within the train and test sets.

Models used are Naive Bayes, SVM (linear kernel) and XGBoost classifiers.

For Naive Bayes and SVM models, texts were encoded using the TF-IDF scheme. For XGBoost, texts were encoded using CountVectorizer.

Classifiers' performance is reported in Table 2.

Model	**Test accuracy**	**Test F1**
Naive Bayes	0.81423	0.81127
SVM (linear kernel)	0.72119	0.72134
SVM (polynomial kernel)	0.61758	0.69497
SVM (sigmoid kernel)	0.72219	0.723910
XGBoost	**0.94210**	0.94063

Table 2: Test metrics for sentiment classifiers trained and tested on Amazon reviews dataset

Aligning with our expectations, the XGBoost classifier performed the best, with a classification accuracy of 0.94210 on the Amazon test dataset. Therefore, this fitted XGBoost model has been used to generate sentiment labels for Twitter and Vivino reviews.

5.3 Sentiment of Twitter and Vivino reviews

After classifier training and evaluation, it was used to generate sentiment labels to Twitter and Vivino reviews, based only on the text of the review. Said label was added to the datasets and is from then on considered true for exploratory data analysis purposes. Label 0 encodes negative sentiment, while label 1 encodes positive sentiment.

Opposite of what was expected, the reviews were not very unbalanced in terms of sentiment. 90.000 entries have been classified as positive and 60.000 as negative. This allowed us to proceed with further analysis without having to resort to synthetic dataset alteration methods.

6 Exploratory data analysis

The central part of the project is the exploratory data analysis of the Vivino and Twitter datasets. This procedure aims to extract insightful information about the preferences of wine consumers. Extracted information should provide wine producers about what features of wines consumers prefer and allow them to adjust their offer, research and production process.

6.1 Twitter dataset analysis

The Twitter dataset was made by scraping tweets from the Twitter API. Due to the API's policy, it was only possible to extract one week of tweets. After eliminating the tweets duplicates, it had 6,524 different wine-related tweets (see 2.2).

In social networks, it is common to use colloquial vocabulary, contractions, specific jargon and it is easy to find many spelling mistakes. In addition, the extracted reviews are free text, so there is no identifier to identify the wine characteristics.

The combination of having very few examples and unstructured text makes it difficult to extract useful wine features and to analyse their correlation with sentiment analysis. For this reason, it was not expected to achieve meaningful results.

To find if there is any correlation between the features and the sentiment analysis, adequate testing was carried out in the exploratory analysis. The selected features were *wine type*, *color* and *location country.*

First, the data was transformed to the vector space using one-hot and label encoding. For the one-hot encoding, the obtained matrix was too sparse.

Once the data was numerically represented, it was projected into 2D and 3D spaces (Figures 2 and 3), using PCA (see 3.2.2). Unfortunately the plots were not insightful, as the examples with positive and negative sentiment label were mixed.

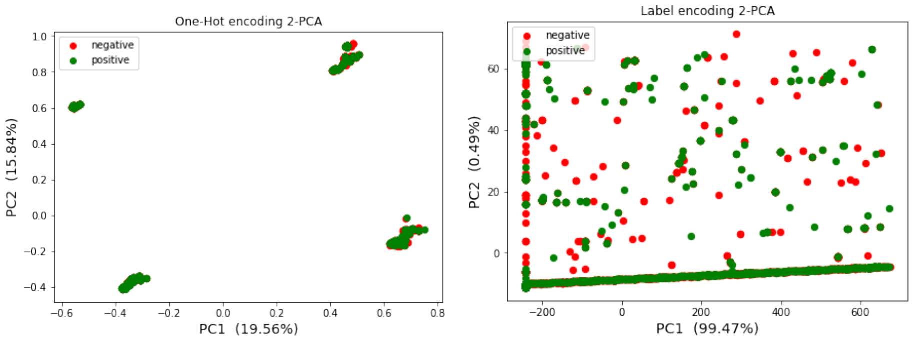


Figure 2: One-hot and labeling encoding using 2-PC

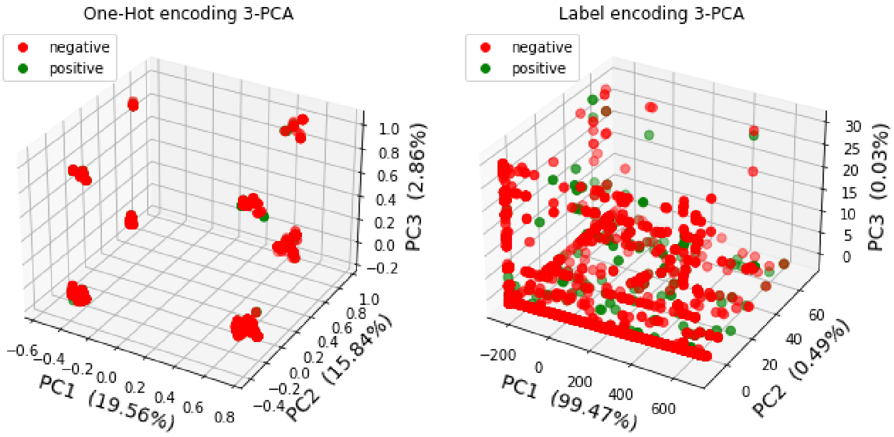
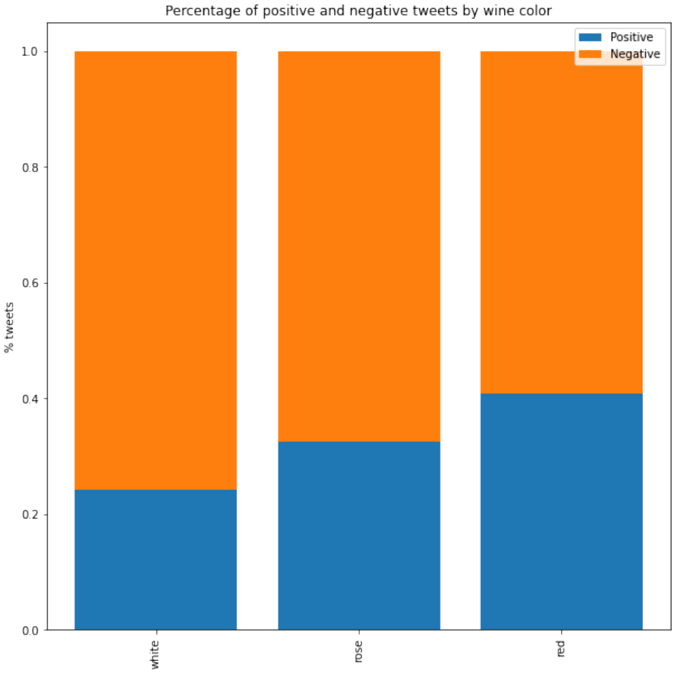


Figure 3: One-hot and labeling encoding using 3-PC

After that, the correlation between each feature and the sentiment classification was examined. For this purpose, the number of positive and negative reviews for each category of each feature was plotted in a stacked bar chart. For the *country* and *wine type* features, only those categories with more than 16 observations were plotted.



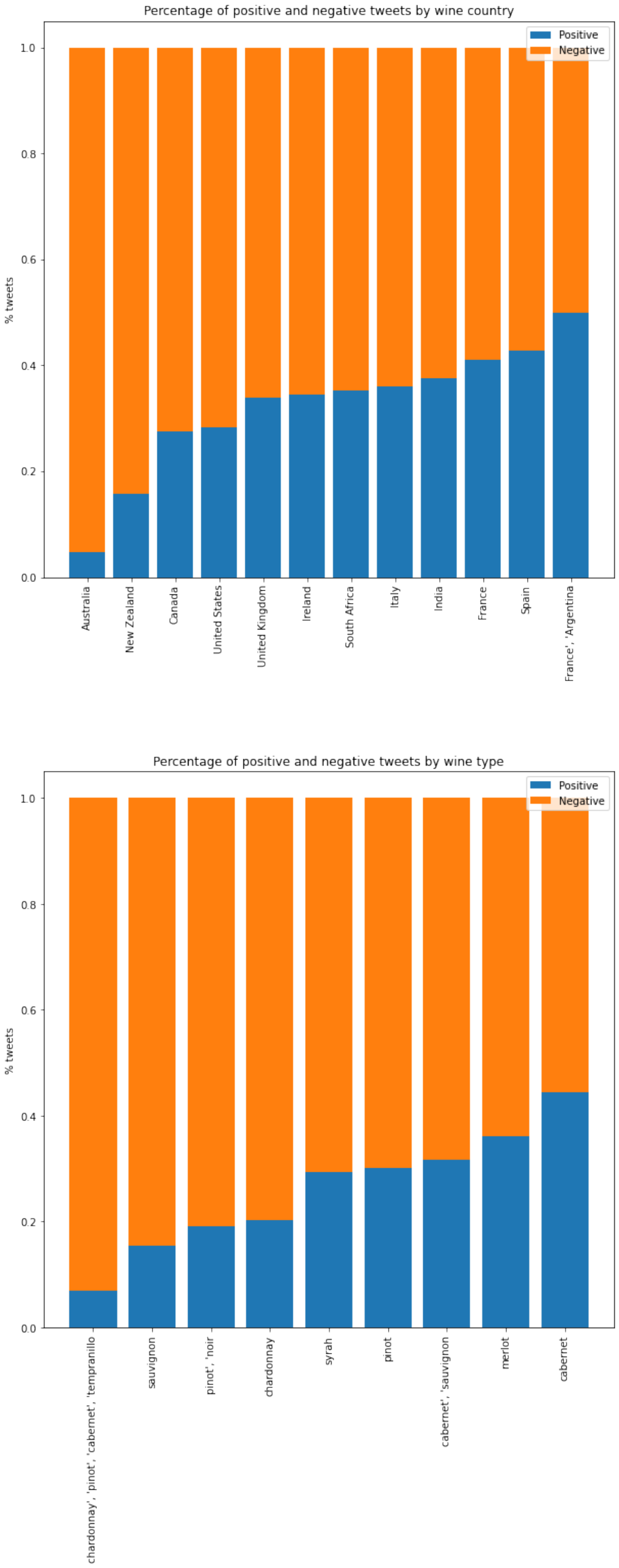


Figure 4: One-hot and labeling encoding using 3-PC

It can be observed that all of the features obtain predominantly negative results. This is due to the fact that there is more negative tweets. From this figure, is possible to conclude that people prefer red wine, *cavernet sauvignon* and *merlot* grape types and wine that originates from France and Spain.

6.2 Vivino dataset analysis

The Vivino dataset offers information about several wine properties, of which this project examines wine structure, wine origin and style.

6.2.1 Exploring wine structure

Wine structure is comprised of three components:

1. acidity
2. intensity
3. sweetness

We examined the difference between structure components for positively and negatively classified data, the correlation between components and the ability to predict sentiment based on different combinations of said components.

Sentiment towards acidity, intensity and sweetness of wine

In this section we explored whether acidity, intensity and sweetness of wine are discriminative for the assigned sentiment and can they serve as predictors of sentiment.

First, we provide statistical motivation for using said structure components to predict sentiment.

In Figure 5 we can observe the boxplot of acidity component for positively and negatively classified subsets. The mean of the positive subset is 3.41019 and the mean of the negative subset is 3.43386. Despite means being relatively close and difference is not visible in the box plot, running a mean equality test was informative.

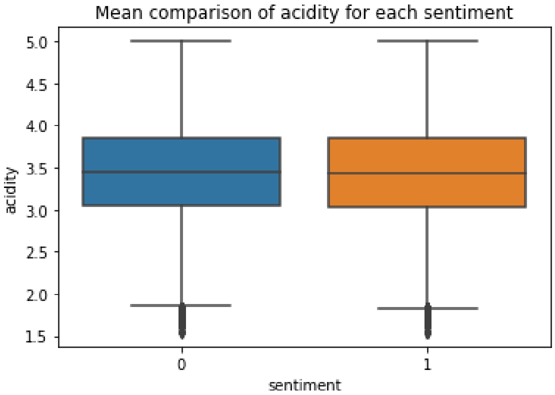


Figure 5: Boxplot of acidity for positive and negative subsets

Source: Insert image source here

Nor positively nor negatively classified subsets of acidity measure passed the KS and Lilliefors test for normality, so we resorted to using the nonparametric Mann-Whitney U test (one-sided alternative) for equality of measures of the central tendencies. The statistic value of the U test is $2.452e^{9}$ with p-value of $8.64e^{-10}$. We can reject the zero-hypothesis about equality of means

$H_0 : \mu_{pos} = \mu_{neg}$in favour of $H_1 : \mu_{pos} < \mu_{neg}$ and conclude there is a statistically significant difference in means of acidity measure for positively and negatively classified examples.

The same procedure was performed for sweetness and intensity and the same conclusions were drawn for all three wine structure components.

Predicting sentiment from structure components

This allowed us to proceed with seeing how well the three components act as predictors of sentiment label. Several models have been trained and tested on the Vivino dataset and the best results are reported in Table 3.

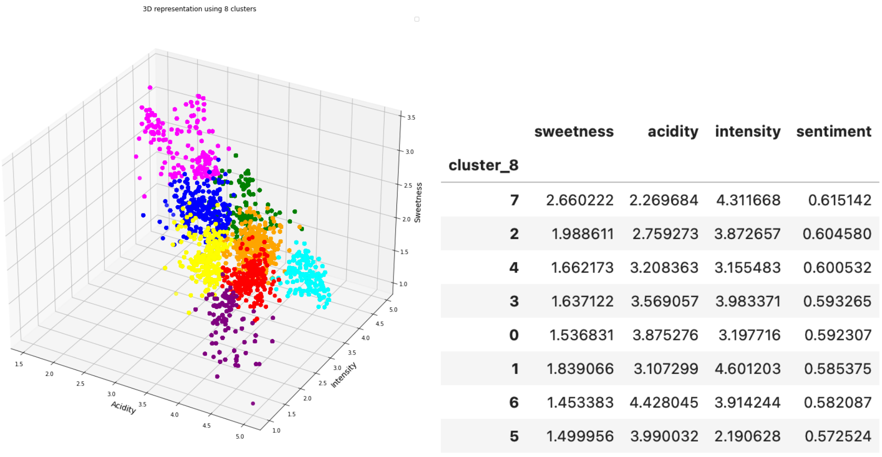
Model	**Hyperparameters**	**Test accuracy**	**Test F1**
Naive Bayes	-	0.59253	0.743294
SVM (linear kernel)	-	0.59533	0.74634
SVM (RBF kernel)	$C = 10 \gamma = 1$	0.60133	0.74818
SVM (RBF kernel)	$C = 0.1 \gamma = 100$	0.60089	0.74769
SVM (RBF kernel)	$C = 100 \gamma = 1$	0.60071	0.74763
XGBoost	$\eta = 0.75$	0.50310	0.69497

Table 3: Test metrics for sentiment classifiers trained and tested on Amazon reviews dataset

Considering that no model accomplished accuracy significantly higher than a random guess, we conclude that these three structural components of wine are not sufficient predictors of sentiment.

Clustering of wine data

To get a more aggregate insight into how different wine structure components are perceived by customers, we performed clustering using k-means algorithm. 8 clusters can be observed in left of Figure 6. The optimal number of clusters was determined using the elbow method.



	sweetness	acidity	intensity	sentiment
cluster_8				
7	2.660222	2.269684	4.311668	0.615142
2	1.988611	2.759273	3.872657	0.604580
4	1.662173	3.208363	3.155483	0.600532
3	1.637122	3.569057	3.983371	0.593265
0	1.536831	3.875276	3.197716	0.592307
1	1.839066	3.107299	4.601203	0.585375
6	1.453383	4.428045	3.914244	0.582087
5	1.499956	3.990032	2.190628	0.572524

Figure 6: Average sweetness, intensity, acidity and sentiment for wine clusters

In Figure 6 we can observe how average cluster sweetness, acidity and intensity change with decreasing average sentiment. We can notice that clusters of wine with high average sentiment also have high average sweetness and intensity. For acidity, we can observe the opposite. Thus, we conclude that more sweet and less acid wines are preferred by consumers.

Correlation between wine components

Positive correlation of sweetness and sentiment and the negative correlation of acidity and sentiment observed in clusters incited us to see if there is any correlation between the three structure components.

In Figure 11 we can see that acidity and sweetness truly are inversely correlated, while intensity is slightly positively correlated with sweetness. This tells us that it is common for sweet wine to be less acidic and vice versa.

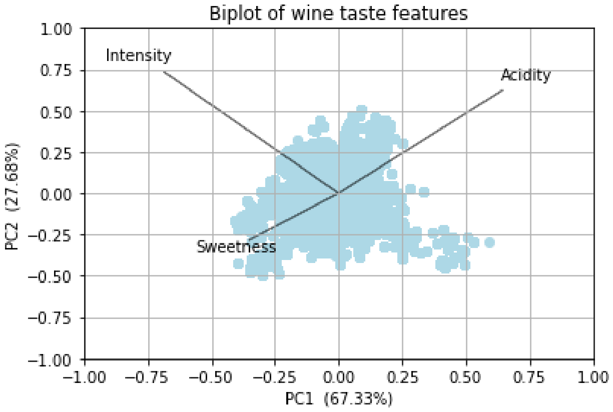


Figure 7: Correlation of sweetness, intensity and acidity

6.2.2 Wine preference by wine origin

The Vivino dataset provides the country of origin of each bottle of wine. The scraped dataset contains wines from 17 different countries. The amount of wines and reviews, however, varies a lot depending on the country.

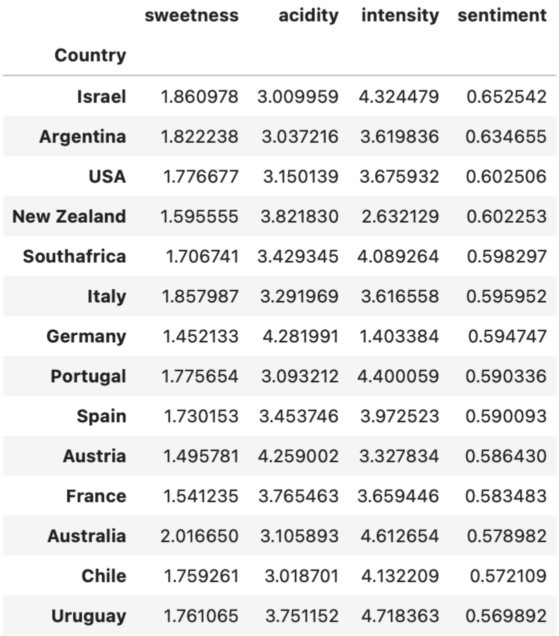
Country	**Num. of wines**	**Num. of reviews**
Italy	797	64975
French	314	35043
Spain	141	12032
Argentina	63	6758
USA	50	6109
Australia	40	5753
South Africa	28	5268
Chile	25	4083
New Zealand	14	1470
Portugal	13	952
Germany	10	619
Austria	6	533
Israel	2	236
Uruguay	2	93

Table 4: Amount of wines of each country in the dataset

With these data, it was analyzed whether the origin of the wine determined the taste features as well as the sentiment. Figure 8 shows the mean of the features for each country.

6.2.3 Wine components with regards to origin

This section studies difference between structural components of wines of different origins, whilst taking the average sentiment in account.



Country	sweetness	acidity	intensity	sentiment
Israel	1.860978	3.009959	4.324479	0.652542
Argentina	1.822238	3.037216	3.619836	0.634655
USA	1.776677	3.150139	3.675932	0.602506
New Zealand	1.595555	3.821830	2.632129	0.602253
Southafrica	1.706741	3.429345	4.089264	0.598297
Italy	1.857987	3.291969	3.616558	0.595952
Germany	1.452133	4.281991	1.403384	0.594747
Portugal	1.775654	3.093212	4.400059	0.590336
Spain	1.730153	3.453746	3.972523	0.590093
Austria	1.495781	4.259002	3.327834	0.586430
France	1.541235	3.765463	3.659446	0.583483
Australia	2.016650	3.105893	4.612654	0.578982
Chile	1.759261	3.018701	4.132209	0.572109
Uruguay	1.761065	3.751152	4.718363	0.569892

Figure 8: Mean of taste features and sentiment per country

We can observe how Israeli wine is the one that has the best reviews while Uruguay's is the one that was liked the least. This is interesting considering the fact that they are both similar in terms of intensity and sweetness. The main difference in terms of taste is its acidity since Uruguay wines scored 3.75 in acidity while Israel scored 3.00.

Besides this, wine intensity might also have a say in Uruguayan wine ratings since they scored the highest intensity among all the countries (4.71 out of 5). Israeli wines are also intense (4.32), but maybe the excessive intensity in the Uruguayan wine made the customers write a bad review.

6.2.4 Wine preference by wine style

The differences between the types of wine have also been extracted and analyzed. The Vivino dataset contained 105 different types of wines but many of these types only contained a few bottles. For this reason, visualizations only show the most common types of wine.

Once again, the ratings of each of these wines were especially interesting for the project. The plot below shows the average sentiment of the 21 best-ranked wine styles.

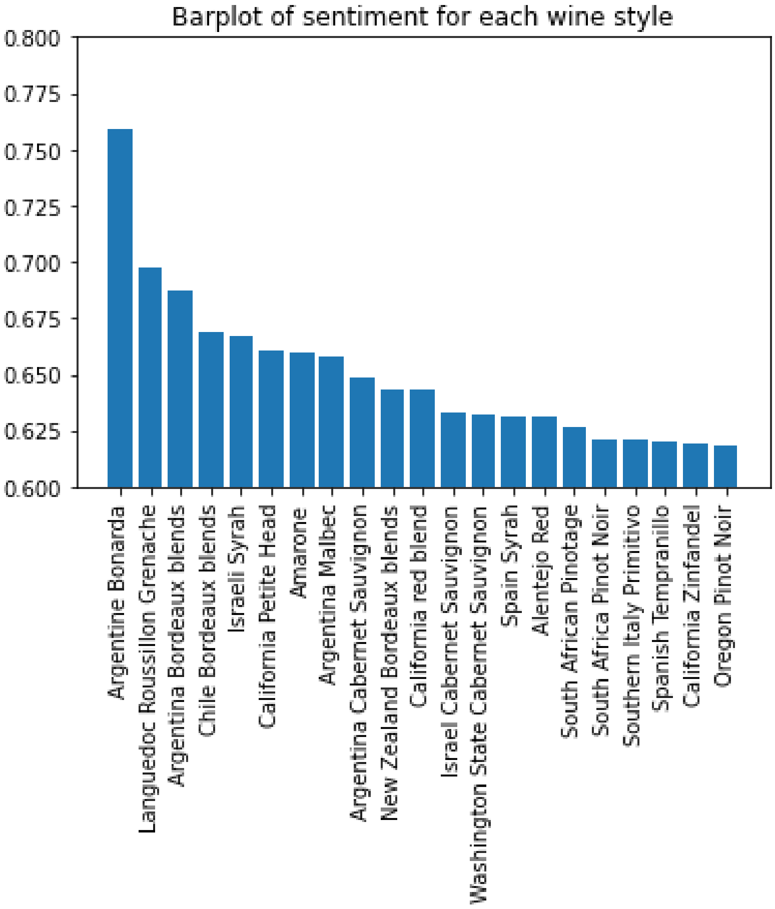


Figure 9: Wine styles that scored the best ratings

The results showed that Argentinian Bonarda and Languedoc-Roussillon Grenache scored the highest percentages of good reviews 75.9% and 69.7%. On the other side, Northern Portuguese Red and Spanish Cabernet Sauvignon received the worst scores with only 35.7% and 42.1% of positive opinions. The difference in taste of these wines was analyzed using a biplot.

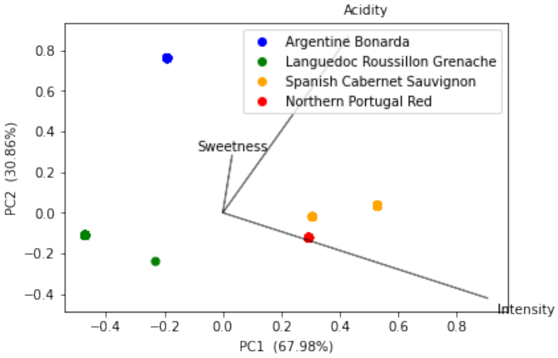


Figure 10: Projection of the two best and two worst types of wines

The yellow and red dots represent the wines that had the worst review while blue and green indicate the opposite. It can be seen how intense wines got bad reviews. Languedoc-Roussillon is below average in terms of acidity and Argentinian Bonarda is very sweet but also acid and not very intense.

Moreover, the 3 most common types of wines were also analyzed by using a biplot.

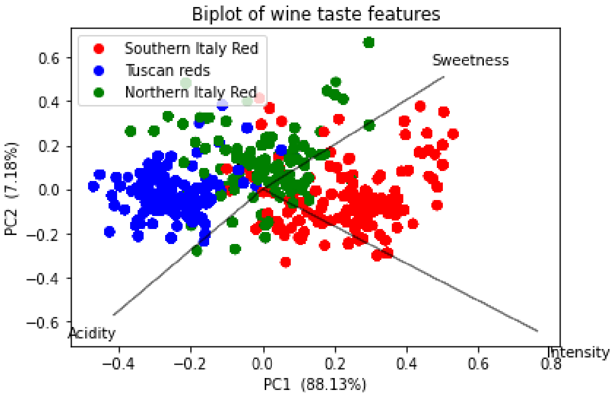


Figure 11: Projection of the 3 most common types of wines

It can be seen how in general the wines of type Southern Italian Red are more intense than the average and some of them are very sweet too. Tuscan Red is an acid type and not very intense.

7 Discussion and reflection on the project

The purpose of this project was to provide wine producers with information about wine customers' wine preferences. This comprises preferences regarding wine taste, country of origin and style. By analyzing sentiment of wine reviews and performing exploratory data analysis the research has concluded that customers generally have preferences towards more sweet wines, that taste components of wines are correlated and that country of origin plays a role in sentiment towards it.

The central problem of data collection was finding a dataset with enough relevant review-like texts. Even though Twitter seemed like a suitable datasource for such dataset, the Twitter API scraping is limiting and the obtained reviews were very sparse in terms of topics, sometimes not at all wine related. Generally, data from Twitter could be described as poor quality. On the other hand, the Vivino dataset was easy to scrape, offered a plethora of information to choose from and getting a decent-sized dataset was manageable.

Even though the sentiment analysis results were satisfying, we have not reached state-of-the-art sentiment classification. In order to improve this, resorting to more resource consuming methods, such as training distributed vector representations and deep classifiers, would be required. Numerical user ratings could also be utilized to improve the sentiment classifier. Also, incorporating them in the analysis of positive and negative reviews has potential to yield meaningful realisations.

We believe our main contribution to wine producers in terms of preference insights lies with the preferred styles of wine and structure. This information can be used to switch production to more sweet wines of specific styles, of course given the availability of suitable grape sorts.

7.1 Future work

For future work, we believe exploring the sentiment towards wine types might bring beneficial insights. Furthermore, this research has only inspected Vivino data irrespective of any user information, resorting to aggregation methods over the whole dataset. We believe performing analysis of users has the potential to build relevant user profiles and serve as rudimentary recommendation systems to producers on geographic or individual preference level.

8 Conclusions

The main findings of our project indicate that the taste features, the origin of the wine and the style are very useful when predicting whether a bottle of wine will be liked. In our dataset, sweeter wines are well received as opposed to very intense wines, which tend to get bad ratings. The origin of the wine also seems to determine its quality. We have found out from the scraped data that Israel wines tend to obtain better reviews. Instead, data implies that Uruguay has the worst quality wine. This result, however, can not be considered as very representative since we do not have a big amount of wines from neither of this two countries.
Finally, the style of the wine has proven to play a big role and the features of the preferred styles have been analyzed concluding that styles that are below average in terms of acidity and above in terms of sweetness are preferred.

We hope Ciù Ciù finds our insight useful and that it aids them in their future business decisions.

References

Chen, Tianqi and Carlos Guestrin (2016). 'XGBoost: A Scalable Tree Boosting System'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

Al-Daoud, MB, NB Venkateswarlu and SA Roberts (1995). *Fast K-means clustering algorithms.* University of Leeds, School of Computer Studies.

Gauthier, Thomas D. (2001). 'Detecting Trends Using Spearman's Rank Correlation Coefficient'. In: *Environmental Forensics* 2.4, pp. 359–362. ISSN: 1527-5922. DOI: https://doi.org/10.1006/enfo.2001.0061. URL: https://www.sciencedirect.com/science/article/pii/S1527592201900618.

Ghosal, Attri et al. (2020). 'A Short Review on Different Clustering Techniques and Their Applications'. In: *Emerging Technology in Modelling and Graphics.* Ed. by Jyotsna Kumar Mandal and Debika Bhattacharya. Singapore: Springer Singapore, pp. 69–83. ISBN: 978-981-13-7403-6.

Kemp, Sarah E et al. (2018). 'Introduction to descriptive analysis'. In: *Descriptive analysis in sensory evaluation* 1.

Lawless, Harry T. and Hildegarde Heymann (2010). 'Descriptive Analysis'. In: *Sensory Evaluation of Food: Principles and Practices.* New York, NY: Springer New York, pp. 227–257. ISBN: 978-1-4419-6488-5. DOI: 10.1007/978-1-4419-6488-5_10. URL: https://doi.org/10.1007/978-1-4419-6488-5_10.

MacFarland, Thomas W. and Jan M. Yates (2016). 'Mann–Whitney U Test'. In: *Introduction to Nonparametric Statistics for the Biological Sciences Using R.* Cham: Springer International Publishing, pp. 103–132. ISBN: 978-3-319-30634-6. DOI: 10.1007/978-3-319-30634-6_4. URL: https://doi.org/10.1007/978-3-319-30634-6_4.

Massey, Frank J. (1951). 'The Kolmogorov-Smirnov Test for Goodness of Fit'. In: *Journal of the American Statistical Association* 46.253, pp. 68–78. ISSN: 01621459. URL: http://www.jstor.org/stable/2280095 (visited on 24th Oct. 2022).

Mishra, Himanshu, Shuchi and Shashi Tripathi (May 2017). 'A Comparative Study of Data Clustering Techniques'. In: DOI: 10.13140/RG.2.2.18076.54401.

Mishra, Sidharth et al. (Jan. 2017). 'Principal Component Analysis'. In: *International Journal of Livestock Research*, p. 1. DOI: 10.5455/ijlr.20170415115235.

Natekin, Alexey and Alois Knoll (2013). 'Gradient boosting machines, a tutorial'. In: *Frontiers in neurorobotics* 7, p. 21.

Srinivasarao, Ulligaddala and Aakanksha Sharaff (2021). 'Email Sentiment Classification Using Lexicon-Based Opinion Labeling'. In: *Intelligent Computing and Communication Systems.* Ed. by Brahmjit Singh et al. Singapore: Springer Singapore, pp. 211–218. ISBN: 978-981-16-1295-4. DOI: 10.1007/978-981-16-1295-4_22. URL: https://doi.org/10.1007/978-981-16-1295-4_22.

Suthaharan, Shan (2016). 'Support Vector Machine'. In: *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning.* Boston, MA: Springer US, pp. 207–235. ISBN: 978-1-4899-7641-3. DOI: 10.1007/978-1-4899-7641-3_9. URL: https://doi.org/10.1007/978-1-4899-7641-3_9.

Syakur, M A et al. (Apr. 2018). 'Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster'. In: *IOP Conference Series: Materials Science and Engineering* 336, p. 012017. DOI: 10.1088/1757-899x/336/1/012017. URL: https://doi.org/10.1088/1757-899x/336/1/012017.

Webb, Geoffrey I, Eamonn Keogh and Risto Miikkulainen (2010). 'Naıve Bayes.' In: *Encyclopedia of machine learning* 15, pp. 713–714.

Williams, R. B. G. (1984). 'The Binomial Test and Lilliefors' Test'. In: *Introduction to Statistics for Geographers and Earth Scientists.* London: Macmillan Education UK, pp. 218–225. ISBN: 978-1-349-06815-9. DOI: 10.1007/978-1-349-06815-9_16. URL: https://doi.org/10.1007/978-1-349-06815-9_16.

Zhao, Bo (May 2017). 'Web Scraping'. In: pp. 1–3. ISBN: 978-3-319-32001-4. DOI: 10.1007/978-3-319-32001-4_483-1.

A Appendix

The main issue we have been faced with is the low quality of tweets scraped and limitations posed by the Twitter API policy regarding the amount of tweets we can scrape. We have remedied that by changing the scope of our project and scraping Vivino wine website for more convenient data.

Considering the team already had experience with machine learning and data science techniques, the main learning takeout for us was Twitter and webpage scraping, filtering of data and assessing quality of the collected data.

Team dynamics was up to the mark, with all teammates contributing to the project, which made this working on the project a beneficial and pleasant learning experience.

All of the Python code, Twitter and Vivino datasets used in this study can be found here.

The Amazon dataset can be found here.