DM2583 Big Data in Media Technology Fall 2022

Using Big Data Analytics to conduct market analysis for Ciù Ciù Tenimenti Bartolomei

October 12, 2022

Group 10: Senane Zineb, Oumhamed Younes, Oliver Gui

Emails: {senane, oumhamed, ocwgui}@kth.se

Professor: Haibo Li

1 Abstract

Our team aims to help Ciù Ciù Tenimenti Bartolomei harness the potential of Big Data to disrupt the centuries-old wine industry. We would like to provide a market analysis such that Ciù Ciù will be able to provide a better wine consumption experience for their customers, increase their customer pool, and generate increased revenue. Our team's objectives were to analyse the research keys that guide consumers when looking for wine online which offers a greater insight into the behaviour and mindset of consumers, information that is invaluable to Ciù Ciù in the future development of their product lines. We planned to use this information would be used in tandem with data on countries in Europe where there is a rising popularity or consumption in wine, and the features of a wine around the price of 25 euros that consumers are inclined towards. These insights would be interesting as we process text data in the form of search queries and wine reviews to derive insights for consumption patterns, which is data that is rarely utilized in this field. We have collected the wine consumption data from the World Health Organization - European Region. For regions with rising popularity in wine, we have collected the data from the Google Trends API. Additionally, we created data visualisations to illustrate trends.To investigate features of a wine that consumers prefer, we have scraped data from the websites Vivino and WineEnthusiast. We solved the data analysis problem by first cleaning the data and conducting exploratory data analysis. Next, we vectorized the text data with several different techniques such as bag-of-words (BoW), TF-IDF (term frequency-inverse document frequency), bag-of-n-grams. Our team also used this data as input to our classification models which were the Naive Bayes Classifier, Long Short Term Memory (LSTM) Recurrent Neural Networks, and Vader (Valence Aware Dictionary for Sentiment Reasoning) to predict which wines consumers are more likely to prefer.

The models implemented and trained on WineEnthusiast reviews gave a bad performance. The runtime for some deep learning architectures such as SVM and LSTM was huge and hence we were unable to test them on test and evaluation dataset. Overall, LSTM give the best results on the training set, and Naive Bayes performed better than VADER. Additionally, the model created from the Vivino dataset had poor performance as well. This paper shows that in order to benefit from using NLP in wine industry field, researchers are required first to construct a good dataset feasible for sentiment analysis tasks and collect useful and significant reviews that have a clear sentiment and not just a neutral description.

2 Introduction

The field of data analytics has gained and sparked a growing interest in various industries since the increasing complexity of life's challenges. Even Though Big Data Analytics (BDA) has been used already in multiple industries like health and finance, BDA is attracting nowadays concerns in different fields such as agriculture and viticulture. Recently, analytics has been used in sales and marketing departments to help these teams to define metrics for their sales process at the beginning of a marketing campaign and a new label sale in the market to best identify customer interaction with the business and track their performance and progress. Consequently, wineries realized the benefits of utilizing BDA and AI in the making-wine process, they will address and guide their business decisions in their growing operations and yield a growing benefit [1].

Contemporary, many companies use data analytics or AI to create and capture value by creating new products or services or to integrate with the process of some part of their current business. For instance, Palmaz Vineyards is pursuing a modern digital strategy in an ancient analog industry, the wine industry [2].

Big data (BD) is data that is machine-readable as opposed to human-readable. There is no official size that makes data "big". It consists of massive amounts of digital information, collected from all sorts of sources that are too large, raw, or unstructured for analysis using conventional relational databases and techniques. On the other hand, Big Data above all is a multidisciplinary and evolutionary fusion of new technologies in combination with new dimensions in data storage and processing (volume and velocity), a new era of data source variety (variety), and the challenge of managing data quality adequately (veracity) [3].

BDA can be identified as a critical technology to support data acquisition, storage, and analytics in data management systems in modern manufacturing. In other terms, it is the use of advanced analytic techniques against very large, diverse datasets that include structured, semi-structured, and unstructured data from different sources and in different sizes.

In the Wine Industry Data Summit, Paul Mabray states: "Data doesn't always tell the truth". As a result, researchers sow the seeds of challenges and opportunities the wine industry faces for using data to better understand and engage with customers effectively and purposefully [4]. The first challenge is that the data is hard to come by. Antonnio Galloni, CEO of Delectable, created a database for relevant, valuable, and meaningful data for wineries provided by wine consumers that are more engaged and enthusiastic about wines. The data can be accessed through Emetry platform, and it will help wineries to target, reach and influence their customers [4].

BigDataGrapes is another project involving big data and artificial intelligence in the vine and wine sector [5]. It aims to help wine companies to become more competitive in the international market by taking advantage of big data tools and providing decision support resulting from the real-time and transversal analysis of diverse data sources. Moreover, they use deep learning algorithms to develop a new counter that takes as input an image and outputs the number of leaves per vine stock. In the "Innovations in Agricultural and Wine Production Sector" review the authors also point out the importance of innovation in the wine industry including the data analysis process to help companies to adapt to market conditions and requirements and remain competitive [6]. Another specific application of big data is in the viticulture domain. The proposed system detects leaf disease as part of plant pathology [7]. It uses a grape leaf dataset that consists of healthy images combined with abnormal ones, and a feature extractor Gray Level Co-occurrence Matrix (GLCM) is applied to track the most important features of these images. Then, K-Nearest Neighborhood (KNN), a classification model, was trained on this dataset with reduced computational complexity. The system accuracy is 82% and it works quite well for vineyards since it also gives recommendations on how to save grapes at the right time given the predicted label by the model.

Another research conducted in the wine industry focused on wine quality [8] [9] [10]. The first paper discusses the data mining and classification algorithms used for wine quality recognition and identification. The comparison they made between SVM, Logistic Regression, and BP neural networks yields that the most feasible and effective one for that purpose is SVM [8].

The second research work analyzes 12 factors influencing red wine quality through data mining algorithms along with the use of data visualization tools available in Python. The concussion was that alcohol, sulfate, citric acid, and volatile acidity are the most decisive factors that impact red wine quality [9]. The third work aims to evaluate both white and red wine quality using Gradient Boosting applied to a dataset containing chemical and physical features. Moreover, the model used was optimized using hyperparameter tuning. The model achieved a relatively high accuracy at the end of predicting wine quality [10].

Another trend in BDA for the wine industry is to analyze user search queries related to wine to get some insights and conduct a study in the wine market. An example of such work is the study made

by three researchers in the US market. They analyzed online searches and examined user-generated data in the largest wine market to help wineries over the world to support their content marketing and digital strategy decisions [11].

Our project aims to generate a market analysis report for Ciu Ciu Tentimenti Bartolomei based on sentiment analysis applied to web-scraped data. Through this work, we want to tell them how data-driven approaches bear fruit to winemaking, by identifying wine products more appealing to their target audience.

3 Research question

Our team's objective was to conduct research in an innovative way that fully harnesses the potential of Big Data Analytics. This research would give Ciù Ciù a deeper understanding of the market that they are targeting and help Ciù Ciù increase their revenue through more effective marketing and product development strategies. We formulated some questions that we thought Ciù Ciù would like to know the answers to, which are as follows:

- When people are looking for wine online, what are the research keys that guide consumers? What are the words/subjects/topics that are associated with wine research?
- As Ciù Ciù aims to launch a new wine on the market which will be around 25 euros per bottle, what are the features of such a wine that more consumers would be more inclined towards?
- What is the potential market of Ciù Ciù? Which countries in Europe have the highest consumption per capita of wine and which countries have an increasing/decreasing consumption per capita of wine?

Our team hypothesised that web scraping and mining text data would provide us with data with which we could conduct insightful descriptive analysis from a unique angle. This project is important as it aims to utilize text data which has not been used to a great extent in the wine industry. Furthermore, our team hypothesized that building models using text data would be effective in regression and classification tasks, with our independent variables being the text review and the dependent variable being the user's rating of the wine.

4 Method Description

4.1 Programming Framework

Our team did all of our coding on Google Colaboratory in the Python language. Using Google Colaboratory allowed us to work with notebooks which is ideal for data analysis and also share the code and results we used with our teammates easily. A variety of useful libraries are also available in Python and this greatly helped us in our data analysis.

4.2 Data Collection

The data our team has collected are from various sources. Firstly, we downloaded data from the World Health Organization (European Region) on the Wine consumed in pure alcohol in litres per capita for individuals above 15 years of age. This data was collected for years up to 2019 as the data for years 2020 and beyond have not yet been published. To fill in this gap, our team also collected data for wine consumption in 2020 from a Forbes report, which is the most recent data available.

Our team also tried to scrap data from a website called Wine enthusiast. This website contains reviews for different types of wines. However, we were not able to successfully scrape the data due to the disproportionately long time it took to scrape a limited amount of data. Therefore, we used an available dataset on Kaggle, which was named 'Wine Reviews' for classification tasks. We were given two datasets: a training dataset with 130K samples and a test dataset with 150K samples. Each sample has 14 features including: 'country', 'description', 'designation' (the vineyard within the winery where the grapes that made the wine are from), 'points', 'price', 'province', 'region 1', 'region 2', 'taster name', 'taster twitter handle', 'title', 'variety', and 'winery'. Our team has also scraped data from an online wine marketplace Vivino. Vivino is the world's largest online wine marketplace and most downloaded wine app. The Vivino community is made up of millions of wine drinkers from around the world, coming together to make buying the right wine simple, straightforward, and fun. Vivino was created for everyone who likes wine — from the wine curious to the wine enthusiast. Due to the diversity of users of the Vivino application, there are millions of ratings and text reviews on the different types of wine. Using the Python Requests module, our team has scraped approximately 200,000 records for wines around 25 euros in price with the features 'note' (which is the text review) and 'user rating'.

4.3 Data Exploration

WineEnthusiast Dataset doesn't have any missing description and the target column correspond to a score between 80 and 100. After analyzing the data, we found some decsriptions containing hashtags, and URLs. Moreover, after removing stopwords and wine-related words such as "Drink", "black" and "wine", and when displaying the word clouds (visual representation of popular words in texts based on frequency and relevance), intuitively, we saw that the obtained words are quite similar (cf Fig 1. left). In order to turn the project into classification task, we mapped the 'points' column to the 'class' columns created by setting a range of points to -1, 0, 1 for Naive Bayes and to 0, 1, 2 for LSTM, SVM and VADER because these models cannot have negative values as labels. The thresholds were set in order to have balanced classes in the training dataset, and the distribution obtained is slightly skewed between different labels (cf Fig 1. right).

Similar to the WineEnthusiast Dataset, the **Vivino Dataset** was cleaned so that any records with missing text reviews was removed. Furthermore, the dataset was split into 3 classes (cf Fig. 2), those with a user rating of less than or equal to 3.0 were given a score of -1, those with a user rating of less than or equal to 4 but greater than 3 were given a rating of 0 and those with a user rating greater than 4 were given a rating of 1.

4.4 Data Pre-processing

First, we removed non-ASCII words and replaced '&' characters with 'and'. Then, as there are still some hashtags and URLs, we suppressed them as they are not relevant to the given task. We also removed punctuations only for Naive Bayes, SVM and LSTM because they work at the word level. Then, we converted emojis into texts as models can only interpret texts. Furthermore, we removed stopwords, which are commonly occurring words in the English language such as "the", "a", "an", and "in", because they do not add much meaning to sentences. However, as it was better to keep words like "not" or "no" to detect more negative reviews, we removed them from the stopwords list of the NLTK.corpus library.

Finally, an important step is to reduce the inflectional forms of each word into a common base or root in order to be able to analyze the meaning behind a word. To do that, we used lemmatization and stemming methods. Stemming uses the stem of the word, i.e. the part of a word responsible for

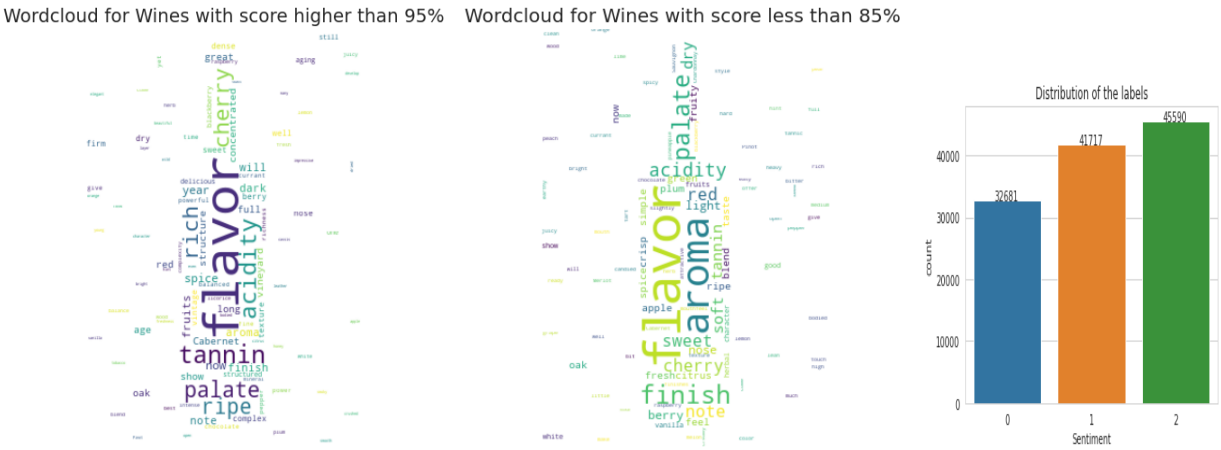

Figure 1: WordClouds for high and least rated wines on WineEnthusiast (Left) and labels distribution (Right)

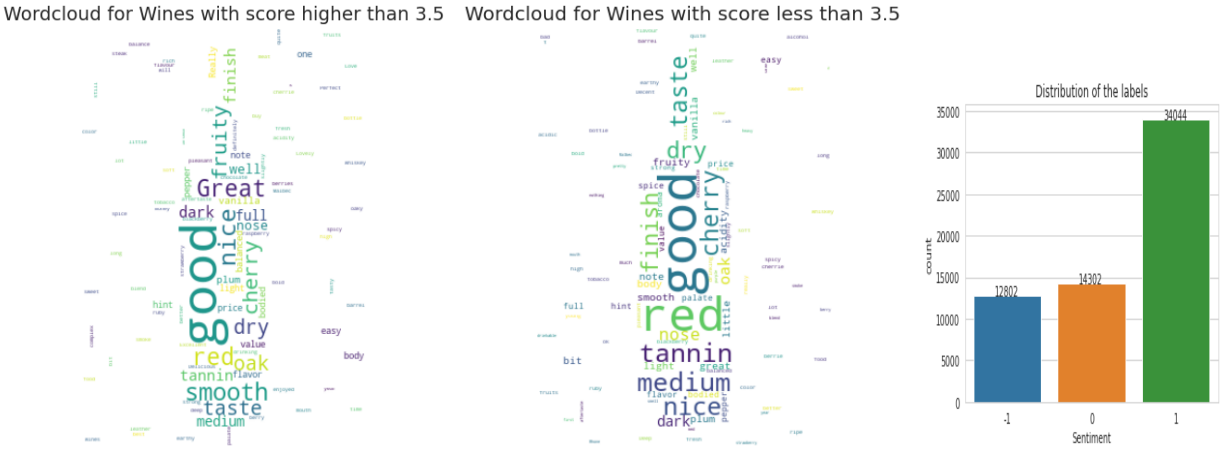

Figure 2: WordClouds for high and least rated wines on Vivino (Left) and labels distribution (Right)

its lexical meaning, and removes the suffixes (e.g. the stemmed version of "wineries" is "winery"). Instead, lemmatization uses the context in which the word is being used and returns its base form (e.g. lemmatized version: "winery"). Stemming is simpler to implement and faster than lemmatization because it gives shorter vocabulary, but it only takes a word and produces a normalized form, without considering the context. Whereas lemmatization, although taking more time to run, can return either a noun, a verb, or an adjective depending on the context of the text. Thus, we applied lemmatization first as it gives more qualitative performance and then we used stemming to have the stem, except for LSTM because we will apply Word2Vec which does not need this improvement as we will see later.

4.5 Features Extraction

To analyze preprocessed data, a major step in NLP is to find a numerical representation for the input data by extracting features. Our team tried different techniques to extract the desirable features: the Bag of Words representation for Naive Bayes, TFIDF vectorizer for SVM, and the Word2Vec representation for LSTM.

Bag-Of-Words: Bag-Of-Words gives a **sparse** representation of text that describes the occurrence of words within a text. It involves a vocabulary of known words and a measure of the presence of known words. It is the simplest way to vectorize a record, but its main problem is that all the words

are weighted uniformly which is not true in all scenarios as the importance of the word differs with respect to the context. In our setting we used 2-grams.

TF-IDF Vectorizor: The TF-IDF vectorizor converts text into vectors by measuring the importance of words for statistical analysis. As the occurrences of a word in a document increases, TF increases the statistical importance of this word. However, the IDF decreases the statistical importance of a word if it appears in many documents. Hence, common words that appear in most sentences will be given a lower importance, whereas a word that is present many times in a few documents might be more indicative of the context of the document. In our setting we used unigram.

Word2Vec: The Word2Vec algorithm uses a neural network model to learn word associations from a large corpus of text and generate a **dense** representation of a text by also taking into consideration the similarities. Once trained, it can detect synonyms or suggest additional words for a partial sentence. The Word2Vec model is more advanced than Bag-Of-Words since it uses a combination of two techniques: Continuous Bag Of Words (CBOW), which predicts the current word from a window of surrounding context words (where the order of context words does not influence prediction) and the continuous skip-gram architecture. It uses the current word to predict the surrounding window of context words.

4.6 Model Training

To perform the classification of reviews, we used different models: Naive Bayes classifier, SVM, LSTM, and VADER. To compare them, we used the accuracy and F1-score.

Naive Bayes classifier: A naive Bayes classifier is a probabilistic machine learning model based on the Bayes theorem and used for classification with discrete features (e.g., word counts for text classification). It is among the fastest and simplest models but its biggest disadvantage is that it requires features to be independent which is not guaranteed in most real-life cases. However, we chose to use this model for its simplicity and its efficient time complexity, which is useful to refer to as a baseline. Here, we applied a Bag-Of-Words algorithm to the processed review and passed the result to the multinomial naive Bayes classifier.

LSTM: LSTM stands for Long Short Term Memory. It is a recurrent neural network that overcomes the problem of vanishing gradients. It has a memory cell at the top which helps to carry the information from a particular time instance to the next time instance in an efficient manner as it has feedback connections. We used this model in our case since it will enable us to treat sequences of data such as texts (sequence of words).
The architecture we implemented is composed of: an embedding layer with weights set from the Word2Vec model's learned word embeddings; three consecutive structures of an LSTM layer with 128 nodes and 13158 parameters followed by a dropout layer to avoid overfitting and finally a dense layer with the softmax activation function, 129 parameters, and 3 neurons since we want to predict the sentiment of the text and we have 3 different classes. Before passing the labels in the fitting part, we used one-hot encoding as we have 3 nodes in the output layer, one for each class. We used the categorical cross-entropy as the loss function, and the Adam optimizer to optimize the algorithm. We trained the model with 5 epochs and with batches of size 128 to reduce the running time.

VADER: VADER stands for Valence Aware Dictionary and sEntiment Reasoner. It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.It uses a combination of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative and tells gives a Positivity

and Negativity score but in addition to how positive or negative a sentiment is.The motivation behind using it is the well understanding of sentiment of unprocessed text containing emojis, punctuations and reduced execution time since it doesn't require any training.

5 Research Process

5.1 Planning of Research

Our team wanted to find the answers to a few research questions, which were stated earlier that would allow Ciù Ciù to better understand the market and their customers. To accomplish this, we had to conduct some research on the different types of data we could collect and the different techniques we could use for analysis.

Firstly, we needed to find data on the consumption per capita of wine for European countries. After some research, we found the data on the World Health Organization (WHO) website. However, it appeared that the data was outdated as it only included the consumption data from years 2020 and earlier. Furthermore, our team wanted to know what consumers were searching for online when they searched for wine. We believed that obtaining data on their search queries could give us invaluable insight into the behaviour of consumers. As the most commonly used search engine in the world is Google, we researched for a Google API that could allow us to access data regarding the consumers' search queries. We discoved the pytrends api, which was not an official Google API but took advantage of the Google Trends service to retrieve data regarding users' search queries. Next, we conducted research to find wine websites with text reviews that could be scraped, as we wanted to approach the issue of finding the features of a wine that consumers prefer from a novel approach of Natural Language Processing (NLP). The 2 websites that we considered were the WineEnthusiast and Vivino. These websites differed as the WineEnthusiast is a wine review magazine with curated wine reviews by professional wine tasters. On the other hand, Vivino is an online wine marketplace where consumers could leave casual, uncurated reviews. By using these 2 different datasets, our team wanted to explore the difference between the tastes of more sophisticated wine consumers compared to the average wine consumer.

Our team was confident that our story would be appreciated as it would be valuable for Ciù Ciù to have up-to-date statistics of the trends for wine popularity in each country to formulate their marketing and production strategies. Furthermore, a deeper understanding of their consumers' tastes and preferences would allow Ciù Ciù to fine-tune their wines. We aim to accomplish this through the use of classification and regression models to explore which wines consumers would higher predicted rate of enjoying, with their level of enjoyment represented in the rating they give on the wine websites. This would lead to Ciù Ciù expanding their market share and hence increasing their revenue. We connected our story to our research questions by asking research questions regarding the consumer demographics and behaviour, and displaying these results in a story such that Ciù Ciù would be able to use this information directly in their business strategies.

Some open ended sub-questions we asked in our research were:

- When people are looking for wine online, which wine (between red wine, white wine, sparkling wine, and rosé wine) was the most commonly used search query?
- Between different countries, is there any difference in the search patterns for the different types of wine (red wine, white wine, sparkling wine, and rosé wine)?
- Are there any terms in the words/subjects/topics associated with wine research that correspond to the e.g. the brand, the geographical area of wine production, the grape variety, etc.?

- As wine consumption data takes a substantial amount of time to collect, could we explore potential methods (e.g. by using the volume of search queries) for Ciù Ciù to have a more up-to-date view on the changes in consumption of wine per capita in the different countries?

5.2 Challenges faced

The first challenge we faced was in the collection of data for countries in Europe with the highest consumption per capita of wine. Such data is not readily available on the internet and our team only managed to find data on wine consumption in 2020. Hence, the data on wine consumption was likely to be stale and not indicative of the current consumption figures, which would be problematic for our analyses. To fix this, we used the volume of search queries in each country as a proxy for the popularity of wine and hence the consumption of wine. However, there are some issues with this approach as it is is not clear if the correlation proposed above is valid. Furthermore, the value of the search queries is the total amount and not per individual as is in the official data. Our team believes that this was the best approach given our circumstances and the only other approach that would be more superior to this would be a partnership with the World Health Organization or governments to obtain up-to-date data.

Furthermore, our team used the pytrends API to retrieve data from Google Trends on the volume and content of users' search queries. However, the pytrends API was open-sourced and free to use but was not an official Google API. Hence after a short period of time using the API, Google would return with a response code 429 which meant that we made too many requests in a short period of time and Google would stop serving our requests. This prevented us from using the pytrends API for a period of time following the timeout. The only ways we managed to tackle this issue were to either wait for the timeout to end so that Google would serve requests from our IP address again or simply go to the Google Trends web application and manually key in the parameters for our queries.

Additionally, the scraper that we implemented for the WineEnthusiast website did not perform very well. Hence, we decided to go with data that was scraped by another user on Kaggle.

Lastly, we encountered challenges after the scraping of data from the Vivino website. As the text reviews scraped from the Vivino website are of different languages, our team was presented with 4 options. What we did was filter out the non-english reviews using the 'langdetect' module in Python. After this process, we applied our pre-processing steps such as removing punctuation, removing emojis, removing stopwords, stemming, and lemmatization. Alternatively, as some of the libraries we used were specific to the English language, we could have split the data into datasets containing text reviews of different languages. Following which, we could have used different libraries built for processing the specific language and conducted sentiment analysis for these reviews by language. The third option was to use the Google Translate API to translate non-English reviews to the English language. However, we encountered difficulties with the use of this API as there were bugs in it that had been pointed out in community forums over a year ago and have not been resolved. Lastly, an alternative would have been to use multilingual models. Such models include but are not limited to the XLM (Cross-lingual Language Model) and BERT (Bidirectional Encoder Representations from Transformers). However, our team decided not to proceed with these models as they are black box models and we would not be able to interpret the results easily and present them as business strategies to Ciù Ciù.

6 Results

6.1 Consumption Indicators

rank	Country	consumption, million hl	change on 2019
1	France	24.7	0%
2	Italy	24.5	7.50%
3	Germany	19.8	0.20%
4	Spain	9.6	-6.80%
5	Portugal	4.6	-0.60%
6	Romania	3.8	-1.90%
7	Netherlands	3.5	-0.3
8	Belgium	2.6	-3.10%
9	Austria	2.3	2.20%
10	Sweden	2.2	-2.30%
11	Czech Republic	2.1	2%
12	Hungary	1.9	-10.20%

Figure 3: Top wine-consuming countries in the EU

We can see from Figure 3 that the top 3 wine consuming countries are France, Italy, and Germany. Therefore our team has decided to focus on the consumption habits of these 3 countries in the further analyses.

6.2 Google Searches in Top 3 Countries

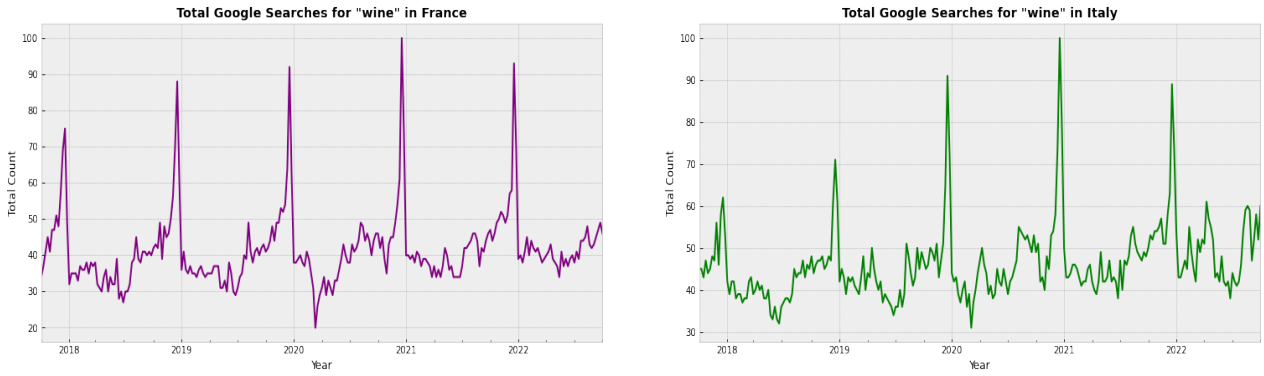

Figure 4: Trend of google searches for wine in France (Left) and Italy (Right)

As wine consumption data from each country takes a significant amount of time to be collected and published (as can be seen from our latest data regarding wine consumption being that of the year 2020), companies are not able to adjust their production or marketing strategies to deal with increases or decreases in demand. Our team aimed to use the number of google searches as a proxy for the "live" popularity of wine in the various countries which would also be influence the amount of consumption of wine. As Google searches tend to be in the country's native language, we translated the word "wine" to that of the native language (e.g. using vin which is French for wine). As can be seen from Figure 4 and 5, We have displayed such results for the top 3 countries in consumption of wine: France, Italy and Germany. The figures show that the search of the word "wine" tends to increase

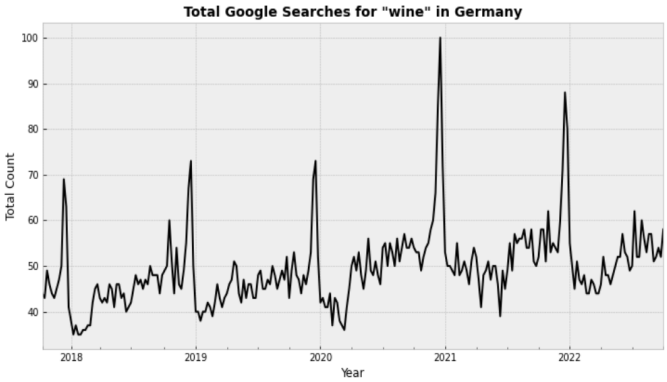

Figure 5: Trend of google searches for wine in Germany

during the beginning and end of the year, which corresponds to the festive season in these countries. While the amount of search queries for wine tends to drop after these festive periods, we can see that the decrease in search queries for wine during the beginning of 2020 was particularly substantial. This could likely be due to the Covid-19 pandemic which prevented consumers from gathering and drinking wine together.

6.3 Queries For Different Wines In Top 3 Countries

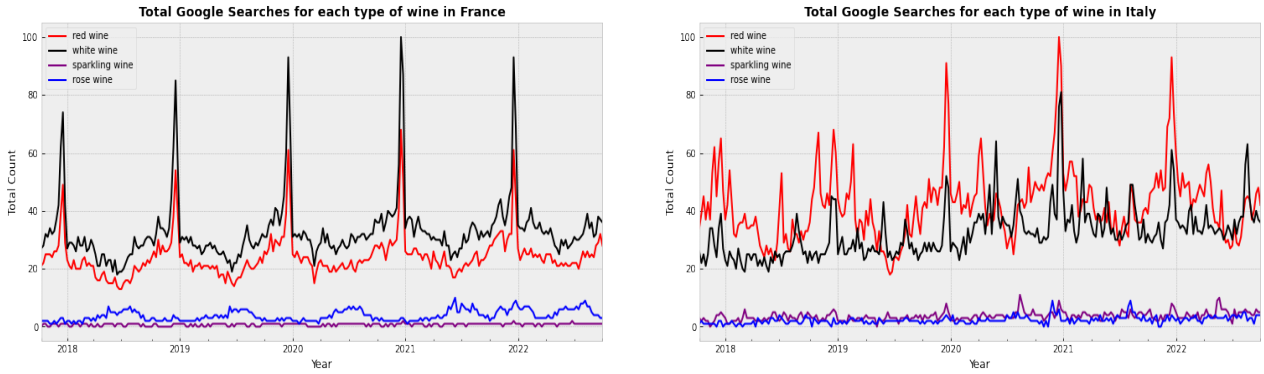

Figure 6: Popularities of wine in France (Left) and Italy (Right)

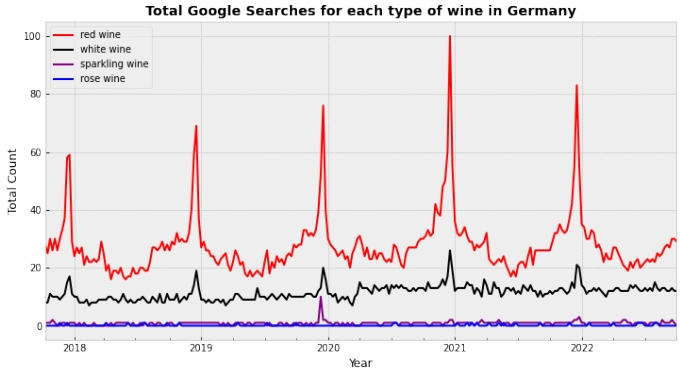

Figure 7: Popularities of wine in Germany

Having identified these 3 countries of interest for Ciù Ciù, we have also used the popularity of search queries for Ciù Ciù's products, red wine, white wine, sparkling wine, and rosé wine to determine which type of wine is the most popular for a particular country. This might help Ciù Ciù in deciding

which wine to promote to consumers from a certain country, which could increase sales. For example, Figure 6 shows that red and white wine are approximately equally popular in France, with white wine generally slightly more popular than red wine. Figure 6 shows that red and white wine are approximately equally popular in Italy as well, with red wine generally more popular than white wine. Figure 7 shows that red wine is significantly more popular than the other types of wine in Germany.

6.4 Research Keys

	top query	top query value	related query	related query value
0	recette vin rouge	100	enlever tache vin rouge incrustée	4700
1	sauce vin rouge	96	sauce bourguignonne au vin rouge avec des lardons	4550
2	vin blanc	95	vin rouge de toscane	450
3	tache vin rouge	50	vivino	400
4	recette au vin rouge	49	boeuf bourguignon cookeo	250
5	vin rouge bordeaux	41	vinatis	140
6	sauce au vin	40	vin rouge sans sulfite	90
7	sauce au vin rouge	39	quel vin rouge avec raclette	80
8	bouteille vin rouge	37	saint joseph vin rouge	80
9	verre vin rouge	35	saint joseph	80
10	tache de vin	35	saint joseph vin	70
11	bourgogne vin	34	vin rouge sans alcool	70
12	tache de vin rouge	34	quel vin rouge avec une raclette	60
13	vin bourgogne rouge	34	vin rouge brouilly	60
14	bourgogne	30	brouilly	60
15	bon vin rouge	29	sauce échalotes vin rouge	50
16	bouteille de vin rouge	26	vacqueyras vin rouge	50
17	bouteille de vin	26	cubi vin rouge	50
18	bourguignon vin rouge	24	vin rouge cote de boeuf	50
19	bourguignon	24	vin rouge avec raclette	40
20	lapin vin rouge	21	morgon vin rouge	40
21	verre de vin rouge	20	température vin rouge	40
22	meilleur vin rouge	20	cubi de vin rouge	40
23	fondue vin rouge	20	crozes hermitage rouge	40
24	boeuf au vin rouge	18	sauce au vin rouge rapide	40

	top query	top query value	related query	related query value
0	vino bianco	100	vino rosso coronavirus	29250
1	brasato vino rosso	64	vino rosso siciliano docg cruciverba	24700
2	brasato	61	ottimo vino rosso del veneto	20450
3	brasato al vino rosso	55	figlio di nettuno ucciso da ercole	18100
4	risotto	50	svetta in gran bretagna	14350
5	risotto vino rosso	50	ottimo vino rosso veneto della zona del garda	12500
6	macchie vino rosso	40	sorta di vino rosso frizzante	12250
7	calorie vino rosso	37	un vino rosso del nord italia	11850
8	vino siciliano rosso	35	un negozio	11300
9	macchie di vino rosso	31	vino rosso del monferrato cruciverba	11200
10	bicchiere vino rosso	31	spezzatino toscano con pepe e vino rosso	10050
11	vino veneto rosso	27	uno scienziato come rubbia	9900
12	vino frizzante rosso	25	requisito indispensabile perche si verifichi q...	9850
13	ciambelline al vino	24	roccia simile al granito	9800
14	ciambelline al vino rosso	24	un tipico vino rosso sardo	9650
15	riduzione vino rosso	24	padre di achille	9400
16	vino dolce rosso	24	dava i responsi nel santuario di apollo	9300
17	risotto al vino rosso	22	il ragu di selvaggina a base di vino rosso e c...	9100
18	bicchieri vino rosso	22	il mitico figlio di nettuno ucciso da ercole	8650
19	torta al vino rosso	22	set di spareggio nel volley	8500
20	bicchiere di vino rosso	21	ciambelline al vino rosso fatto in casa da ben...	8100
21	salsa vino rosso	19	scienza delle cose divine	7900
22	vino rosso veronese	18	affollato viale di cannes	7850
23	calice vino rosso	17	antica arte di predire il futuro con i dadi	7700
24	vino rosso piemontese	17	nero di pregiato vino rosso di puglia	7700

Figure 8: Top wine related queries in France (Left) and in Italy (Right)

	top query	top query value	related query	related query value
0	wein	100	trockener rotwein aus piemont	48500
1	kalorien rotwein	40	rotwein aus piemont 6 buchstaben	30300
2	rotwein trocken	31	trockener rotwein aus piemont 6 buchstaben	21150
3	weißwein	25	trockener rotwein aus dem piemont	14100
4	glas rotwein	20	ital. rotwein 7 buchstaben	13600
5	rotwein kaufen	20	heller bordeaux rotwein mit 7 buchstaben	7850
6	primitivo rotwein	19	tessiner rotwein	7650
7	primitivo	19	zweitgrößte stadt polens	4350
8	aldi rotwein	16	kanadischer tennisspieler daniel	3350
9	rotwein rewe	16	heller bordeaux rotwein 7 buchstaben	2450
10	gulasch rotwein	16	norditalienischer rotwein	2150
11	spätburgunder	15	italienischer rotwein 7 buchstaben	1550
12	rotwein cola	15	gewürzter rotwein 7 buchstaben	1150
13	gulasch	15	franz. rotwein 5 buchstaben	1100
14	spätburgunder rotwein	15	schlaftabletten rotwein 5	950
15	rotwein edeka	15	span. rotwein 5 buchstaben	950
16	rotwein lieblich	15	rotwein aus piemont	750
17	trockener rotwein	15	wie viel kalorien hat ein glas rotwein	750
18	dornfelder rotwein	13	heller bordeaux rotwein 7 buchstaben kreuzwort...	700
19	dornfelder	13	jacques wein depot	650
20	guter rotwein	13	alligatoah schlaftabletten rotwein v	650
21	rotwein sauce	12	span. rotwein kurzwort 5 buchstaben	600
22	rotwein soße	12	span rotwein kurzw	450
23	italienischer rotwein	12	gewürzter rotwein rätsel	400
24	spanischer rotwein	12	rotwein kalorien 100 ml	400

Figure 9: Top wine related queries in Germany

Lastly, for each country we also explored the common queries that users searched for along with key words such as 'wine', 'red wine', 'white wine', and 'sparkling wine' to understand what users who searched for these terms also searched for. This could help Ciù Ciù understand what kind of research

keys (words/subjects/topics) are associated with online wine research. For example, as can be seen from Figure 8, bordeaux red wine is a popular search query in France which might prompt Ciù Ciù to explore selling more wines from this region. From Figure 8, we can also see that Sicilian red wine is a popular search query in Italy. Lastly, from Figure 9 we can see that people in Germany have searched for red wine in supermarket chains such as Aldi, Edeka, and Rewe. This might mean that most Germans get their red wine from supermarkets and Ciù Ciù should aim to get their wines in these supermarkets rather than targeting the Germans through online sales.

However, some limitations of these results are that there are many queries that are irrelevant to a wine company such as Ciù Ciù. These queries include the amount of calories in red wine, how to remove red wine stains, red wine in cooking, etc.

6.5 Classifiers performance on WineEnthusiast Data

For Naive Bayes classifier, we implemented two models based on the simple representation given by the Bag-Of-Words and TFIDF using 2 grams. As expected, the model with Bag of Words worked better on both train and test sets. For testing, the accuracy and F1-score were both around 76%. Looking at the confusion matrix in Fig. 11, the classifier trained on processed texts worked moderately as the true positive sentiments predicted are more numerous but it struggles to differentiate between neutralor negative/positive (quite important numbers for this case).

For LSTM, we fit the model by the processed text of the training set directly as it has its own embedding layer that uses the Word2Vec weights. The accuracy and F1-score for the training set were around 91%. Given these results, we noticed that LSTM performs quite well on the training set for this sentiment analysis problem as expected. This is due to the nature of the architecture used for our LSTM model, as it involves deep learning neural networks that work well for this kind of NLP tasks, and also because it includes dropout layers to prevent any overfitting scenario. Even though we could not assume if it will perform better on unseen data, because of the huge execution time we were unable to test the model on the test and evaluation datasets.

From our experiments, we can see that VADER provided significantly bad results. The accuracy for testing was around 37% and the test F1-score was 30.04%. The results confirmed that Naive Bayes gave a better result on this task. However, due to its simplicity and our relatively big dataset, we find during testing and evaluation that Naive Bayes classifier overfit easily. Overall, we can clearly see that all the models have weak performance. Moreover, looking at the confusion matrix in Fig. 11, we can see that the labels are predominantly on the diagonal meaning that the classifiers worked well on classifying almost all positive sentiments however they predict other review to be neutral or positive as well. The results for SVM classifier are not included in the table because of time constraints, since the model training time exceeded 4 hours.

	Models	**Accuracy**			**F1-Score**		
		Training	Testing	Evaluation	Training	Testing	Evaluation
Accuracy	Naive Bayes	86.88%	76.15%	63.63%	86.74%	76.25%	62.82%
	LSTM	91.31%	%		93.75%	%	
	VADER	42.47%	37.30%	37.30%	30.04%	30.04%	30.04%

Figure 10: Comparison between different models' performance

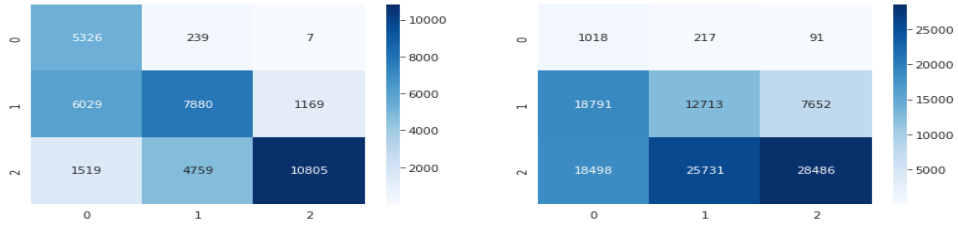

Figure 11: Evaluation Confusion Matrix for Naive Bayes (Left) and VADER (Right)

7 Discussion

In this project, we implemented basic NLP techniques starting with tweet preprocessing. We compared lemmatization and stemming techniques. Then we did three methods of feature extraction from processed reviews that gave a static word representation: a sparse distribution using Bag-Of-Words for Naive Bayes classifier and a dense distribution using Word2Vec for LSTM, and TFiDF for SVM. We also implemented VADER which doesn't need any processed text as they learn from the lexical level and deal automatically with punctuations and non textual characters. Because the moderate quality of the data, the models didn't performed quite well and hence as a future work, a good dataset for collected wine reviews is needed.

7.1 Contributions

Senane Zineb : Data analysis, processing, Feature extraction. Implementing Naive Bayes Classifier, SVM Classifier, LSTM and VADER classifiers, classification using WineEnthusiast Dataset.

Gui Oliver: Google Search Keys, Web-scraping Vivino, classification using Vivino Dataset and all the afforementioned models.

Oumhamed Younes: Web-scraping WineEnthusiast. Implementing LSTM classifier and using it to perform a classification on WineEnthusiast Reviews.

All the members contributed to the report.

7.2 Limitations of Results

1. The text data from reviews on Vivino do not have sufficient information for machine learning models, e.g. those text reviews that are only 1 word long. This leads to our models having poor performance.
2. Scores seem to be arbitrarily decided (e.g. scores that have the exact same word i.e. 'good' can have different scores), especially for the Vivino website, which leads to poor results, hence given a text review it is difficult for us to predict whether a wine is good or bad.
3. As we mentioned previously, our team did not manage to scrap the data off wine enthusiast, hence the data might be stale.
4. Our team could only build weak models for Wine Enthusiast reviews since the descriptions doesn't have a sentiment and they are interpreted as neutral in general.

5. The model created using the Vivino data had extremely poor performance, and as such was not included in the report.

7.3 Future Research

As a future, we aim to tune the hyperparameters to ameliorate the models performance and regularize the algorithms to prevent the overfitting problems. Furthermore,we can try to determine which words are the reason for each classifier to determine the classification (negative, neutral, or positive) by using the Jaccard coefficient to evaluate the overlap between the selected words and the ground truth.

Additionally, we can build different models for different languages, which would allow us to preserve more information instead of simply filtering the data that is not in English. This might increase the performance of the models and provide more insight to Ciù Ciù as their market is mainly countries in the EU, most of which do not speak English as a first language.

8 Conclusion

In conclusion, the findings of our project adequately answered the research questions on the potential market of Ciù Ciù and the trend of wine popularity in the countries that are the top consumers of wine. Our project also found out the research keys that guide consumers when looking for wine online. However, we did not manage to build any models that performed sufficiently well to classify whether a wine will be treated as high quality by a consumer or not. Despite not building a strong model, we believe that our findings on the research keys will still prove useful to Ciù Ciù as they look to expand their market share.

9 Appendix

9.1 Project Assessment

Zineb: From the beginning, we assigned to each team member a set of tasks to do in this learning experience and we fixed our own deadlines for different parts such as data collection, data processing , etc. The team functioned well according to the predefined plan. Nevertheless, the training of different models took more time than what we expected. As I worked with WineEnthisiast Dataset, I conducted the whole data analysis, processing and feature extraction parts with addition to models training and Testing. Since this dataset is too big, the training phase took a huge time especially for SVM because of the big number of training samples. Furthermore, the performance results we obtained as the end didn't reached our expectations as well and the models performed moderately.An explanation to this gap could be the nature of the given reviews. In fact, the description column just describe the wine features and attributes and hence it doesn't reflect explicitly a sentiment. Lastly, this project helped me to deal with a real industry problem where no good and dedicated dataset is available. We learned how to think about and implement an end-to-end solution and handle unexpected problems along with time constraints.

Younes: Before starting the work on the project, each one of the team was aware of the task asked to do. We agreed that we have to focus on the data collection because it's the part that takes time the most. Then, I started to code a scrapping program for collecting data from Wine Enthusiast. But, it was too slow. I tried to run it within Github Actions but the time limit was exceeded and it failed to scrap the data. Furthermore, I also worked on the LSTM model. For this purpose, I added the W2V Vectorizer in the features extraction part. But, I realized that the training of this model takes too much time, then I stopped the training in the second Epoch. Finally, this project is my first real project in the Machine Learning world. And it was a good experience, I learned a lot from it starting with how to collect data and clean it and how to choose the models. Without forgetting, how to handle unexpected problems in such projects.

Oliver: What went well in the project was that we split the tasks up at the start so the whole team was clear on our different roles and responsibilities. During the planning phase, we also had a clear idea of the solution we wanted to propose and how we wanted to go about implementing it. However, during the actual implementation of the project there were some hiccups. As I was tasked with using the pytrends API to work with Google search queries, I only realised that the API had issues taking in many requests as the service would time out. Furthermore, I conducted webscraping on the Vivino website. As Vivino text reviews came in different languages, I wanted to translate them to English but the Google Translate API was not working as intended. Lastly, the models we used were not able to provide satisfactory results possibly due to the quality of the text reviews being subpar. However, this project was still a good learning experience as we had to handle every aspect from data mining to model building and taught me what problems to expect from such projects in the future and how to deal with these problems.

9.2 Datasets used

WHO data on wine consumption per capital in Europe
Web-scraped Vivino data
Web-scraped WineEnthusiast data

9.3 Link to code

All code related to the Google Trends API

All code related to WineEnthusiast Reviews classification
All code related to Vivino Reviews classification

References

[1] Nathaniel K. Newlands. "Artificial Intelligence and Big Data Analytics in Vineyards: A Review". In: (2021).

[2] Matt B. *Palmaz Vineyards: Can big data analytics disrupt the centuries-old wine industry?* 2020. URL: https://d3.harvard.edu/platform-digit/submission/palmaz-vineyards-can-big-data-analytics-disrupt-the-centuries-old-wine-industry/.

[3] Hans Ulrich Buhl et al. "Big Data: A Fashionable Topic with(out) Sustainable Relevance for Research and Practice?" In: (2013).

[4] The Wine Industry Advisor. *Making Big Data Work for Wine*. 2018. URL: https://wineindustryadv
com/2018/10/30/making-big-data-work-for-wine.

[5] Pascal Neveu and Coraline Damasio. "BIG DATA AND ARTIFICIAL INTELLIGENCE: EXAMPLES OF APPLICATIONS FOR THE VINE AND WINE SECTOR". In: (2020).

[6] Jovana Gardašević, Ivana Brkić, and Tamara Krstić. "Innovations in Agricultural and Wine Production Sector". In: (2020).

[7] Jitali Patel et al. "Big Data analytics for Advanced Viticulture". In: (2021).

[8] Zhang Lingfeng, Feng Feng, and Huang Heng. "Wine quality identification based on data mining research". In: (2017).

[9] Shuhao Zhang, Caixing Shao, and Wei Xiao. "Research on Red Wine Quality Based on Data Visualization". In: (2020).

[10] Yizi Liu. "Optimization of Gradient Boosting Model for Wine Quality Evaluation". In: (2021).

[11] Carlos Gonzalo Penela, Patrizia de Luca, and Giovanna Pegan. "Insights from Google search user-generated data: a study on European Wine in the US Market". In: (2017).