Palate Perception:
Unraveling Wine Sentiments through Deep Learning

Dylan Jorling

7/6/2023

Introduction

2023 revenues in the United States wine industry are estimated to approach $57 billion. Despite an average American spending $160 annually on wine, competition within the industry is fierce and capturing market share can be an arduous task[1]. In addition to roughly 14,000 domestic wine producers, approximately 30 percent of U.S. wine sales are imported [2, 3]. With so many consumer options, thoroughly understanding the customer is critical to success in the wine industry.

Although online wine sales currently account for only 2% of total sales, experts predict this figure will rise to 3% by 2028, reflecting the ongoing shift towards digital consumption [1]. One of the most prominent players in this space is Vivino, proudly boasting itself as the "world's largest online wine marketplace."[4] With a vast database encompassing over 15 million wines, Vivino not only offers buying options but also provides invaluable information about the wine's producer, region, and vintage. Furthermore, its user base of approximately 60 million individuals contributes a wealth of reviews that shape the decision-making process for wine buyers worldwide.

Informed purchasing decisions rely heavily on reviews. According to a 2021 study, a significant 84% of consumers consider reviews to be either "important" or "very important" when evaluating a product [5]. In an industry teeming with choices–Vivino claims to have over 200,000 producers total–reviews become a critical factor in differentiating wines from one another. Thus, understanding the impact of reviews on consumer preferences becomes paramount when it comes to selecting wines for purchase.

The primary objective of this paper is to leverage advanced deep learning natural language processing (NLP) techniques to uncover hidden qualities within distinct wine subsets that strongly correlate with positive and negative reviews. These insights can empower wine producers to align their offerings more effectively with consumer desires, enhancing their market positioning.

Additionally, this paper explores the development of a review generator capable of producing diverse types of reviews in the distinctive style of Vivino users. While a review generator may appear as a narrow pursuit, it serves as a foundation for various recommendation systems and offers producers invaluable insights into the preferences of their consumer base.

Through the combination of sophisticated NLP techniques and the creation of a powerful review generator, this study aims to unlock valuable knowledge within the wine industry, enabling producers to enhance their offerings and better cater to the evolving needs of wine enthusiasts.

Data Collection and Cleaning

The wine reviews used in this study were obtained through the official Vivino API, which provided access to a vast collection of user-generated reviews. Initially, metadata for 75,000 unique wines was scraped, encompassing a wide price range from as low as $1 up to a cap of $80. The $80 price cap was chosen to ensure the relevance of the study to the vast majority of consumers. As a point of reference, the average unit price for a bottle of wine sold in the US during 2022 was reported as $11.88 [1]. The collected metadata included essential information such as the average rating, region, winery, varietal, and price of each wine.

Subsequently, the reviews for each specific wine were gathered, with a self-imposed limit of 500 reviews per wine to allow for a diverse sample. As a result, the final dataset comprised a comprehensive collection of 9.5 million reviews, providing a robust foundation to analyze consumer preferences.

It is worth noting that the dataset initially contained reviews in various languages. To maintain the focus on English-language reviews, the Python "langdetect" module was employed to detect and filter out non-English reviews.

Overall, the dataset provides a rich source of information for exploring the factors influencing wine reviews and sentiments among consumers. The inclusion of extensive metadata and a substantial number of reviews enables a comprehensive examination of the relationship between wine characteristics and consumer wants.

Methodology

A. Classification Task and Word Embeddings

The wine reviews were split into 6 different categories, with each data segment set to be individually trained to yield 6 unique word embeddings. Initially, data was split between red and white wines which not only differ in color, but also in aroma and taste profiles. Second, to cater to various types of wine producers, price bands were defined as 1) Under $15, 2) $15-$35 and 3) $35-$80, resulting in a total of 6 categories.

To determine polarity scores for descriptive words, a classification task was designed to classify reviews as "positive" or "negative" based solely on the text content. Given the positive skewness of ratings in the dataset, a cut-off point of 4.0 was used, considering ratings of 4.0 or above as "positive" and ratings below 4.0 as "negative".

Prior to training, several text preprocessing steps were performed, including lowercasing and removal of special characters, punctuation, hyperlinks, emojis, and numbers. Common "stopwords" that provided little sentiment information were also removed. The text was then tokenized and padded to a fixed sequence length to ensure consistent input for the model.

Despite the cut-off point of 4.0, "positive" labels outnumbered "negative" labels by a ratio of 2 to 1. To address this class imbalance, stratified sampling was employed, ensuring that each training and testing set for every data segment contained an equal number of positive and negative labels.

Using the TensorFlow framework, a basic transformer decoder model was designed and trained to predict the review labels. Each data segment underwent 4-fold cross-validation, with 5 training epochs for each fold. Mean accuracy across all models ranged between 68.9% and 71.3%, with an overall mean of 70.6%.

The weights from the embedding layer of each trained model were extracted, yielding 256-dimensional vectors for each word in the model vocabulary. Identifying the descriptive words to analyze–referred to henceforth as the descriptors–involved creating word dictionaries sorted by total frequency for each of the 6 data segments and manually selecting. It became readily apparent that the most commonly used descriptors were essentially identical across price-bands and that only two sets of descriptors needed to be analyzed: red wine descriptors and white wine descriptors.

Descriptive words were manually selected using domain knowledge, with the goal of choosing words clearly describing the wine and that possessed no obvious sentiment. The computation of polarity scores for each descriptor first required identifying two buckets of high-frequency terms with either clear positive or negative sentiment. Next, cosine similarity scores were calculated for each descriptor with both the positive and negative word buckets, with an average score taken across each positive or negative term. Finally, the average of cosine similarities within both the positive and negative word buckets were calculated, giving a total of 2 sentiment scores for each descriptor.

One caveat that should be disclosed is that while finding high-frequency positive words was straightforward, the low prevalence of clearly negative words forced some words with only somewhat negative connotations such as “ok” and “fair” to be included in the negative word bucket. Therefore, the positive polarity scores are of considerably higher confidence than the negative polarity scores.

B. Review Generator

After attempting to design and train a review generator from scratch, which yielded subpar results, the decision was made to employ transfer learning using a "small" version of Open AI's GPT2 model, containing over 127 million parameters. Transfer learning involves training a pre-trained model further on new data with different characteristics than the original dataset, a method often referred to as domain adaptation.

Four main types of reviews were identified as the target for the model to generate: positive and negative reviews for both red and white wines. For this task, negative reviews were defined as any review below a 3.0 rating, while positive reviews were defined as any review above a 4.0 rating. Additionally, only reviews containing 10 or more words were included in the training dataset. Due to the low percentage of reviews meeting the criteria for negative reviews, along with the large amount of computing power required to train such a network, the final dataset consisted of just 280,000 reviews, equally distributed among the four review types.

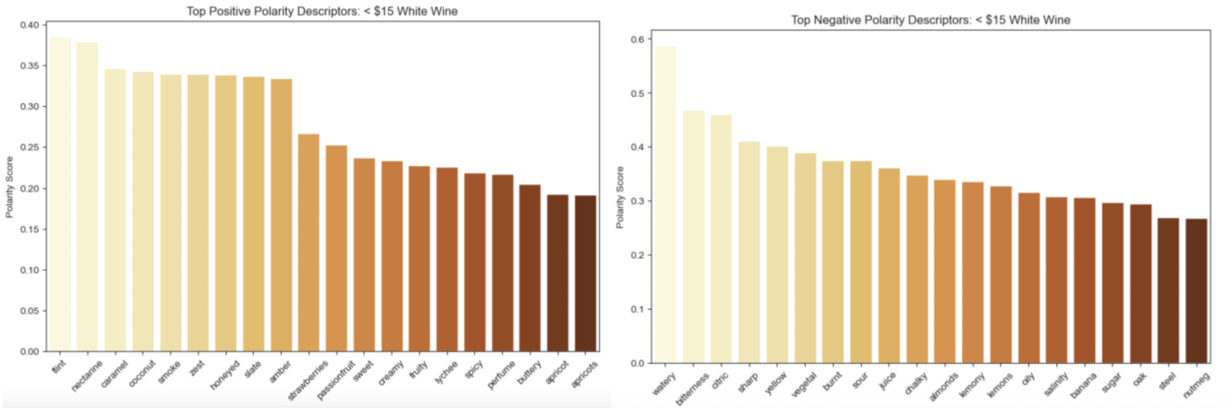
The text underwent preprocessing to remove special characters, emojis, and hyperlinks, while capitalization, stopwords, and punctuation were retained to ensure the model could generate coherent and grammatically correct reviews. Additionally, four specific beginning-of-sentence tokens were added to each review based on their review type. The texts were then tokenized using the GPT2Tokenizer, a sub-word tokenizer originally employed during the training of GPT2.

The training process began by initializing the pre-trained GPT2 weights, and the model was trained on the training set for three epochs using a linearly decaying learning rate. The training loss reached a plateau towards the end of the second epoch and remained relatively stable throughout the entirety of the third epoch, indicating that the model had converged.

Analysis

Descriptor Polarity Analysis

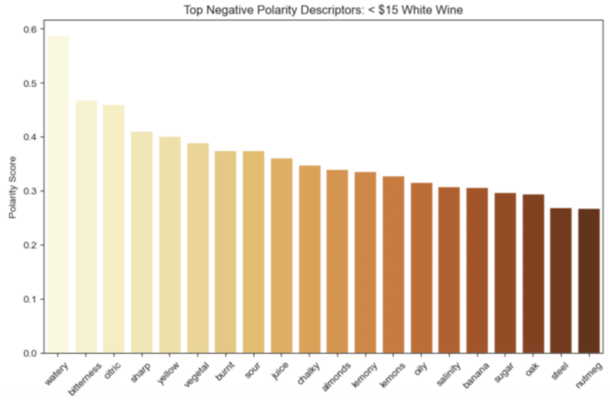
Positive and negative polarity scores were calculated for each of the 186 red, and 136 white wine descriptors. The top-20 most polarizing words were plotted for each data segment and are discussed further below.

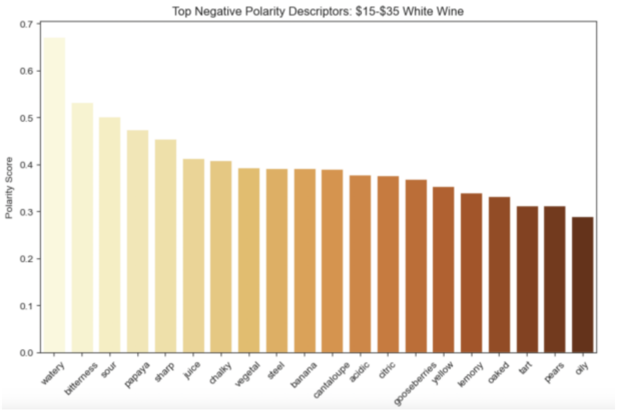


White Wine Below $15 Descriptor Polarities

For white wines priced below $15, the positive descriptors encompass a wide range of characteristics. Prominent positive words include earthy tones such as flint and slate, sweet flavors like nectarine and honey, and terms such as smoke and zest, indicating an intensity of flavor. Tropical fruits flavors, including coconut, passionfruit, and lychee, also exhibit high positive scores. On the negative side, words like watery suggest a wine that lacks body, while bitterness, sourness, citric, and lemon denote an aversion to sour or acidic taste profiles.

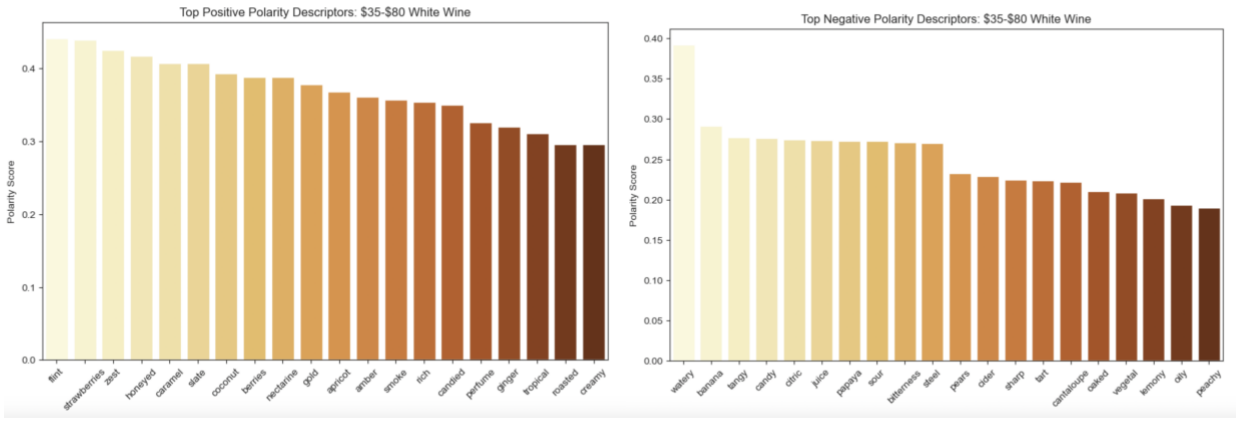
Additionally, terms like oak and steel indicate a dislike for flavors associated with these wine storage methods.





White Wine $15-$35 Descriptor Polarities

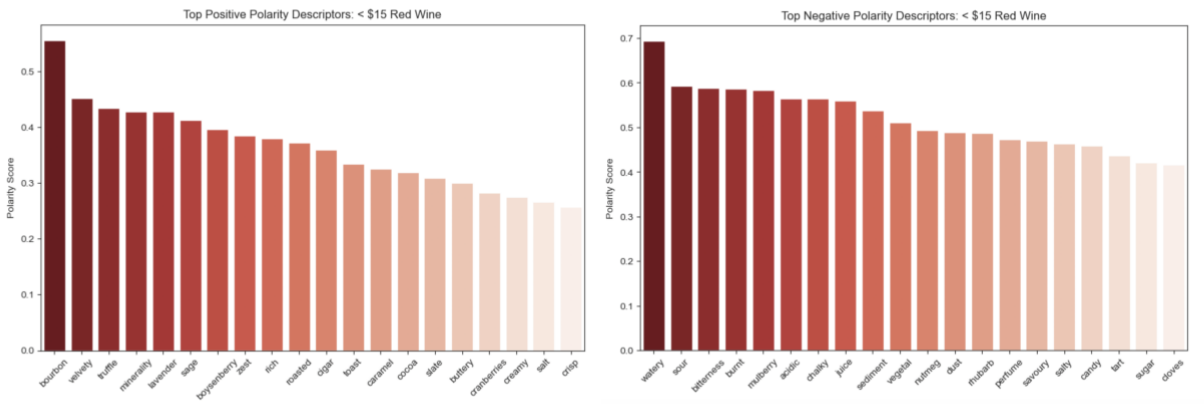
Medium-priced white wines show many similarities in positive and negative polarity scores compared to the cheapest level of white wines. However, there is a modest dip in the preference for sweet flavors, with the emphasis shifting towards tropical fruits like coconut, and descriptors related to the wine's structure, such as rich, dryness, and crisp. Negative polarity descriptors remain relatively consistent across both categories.



White Wine $35-80 Descriptor Polarities

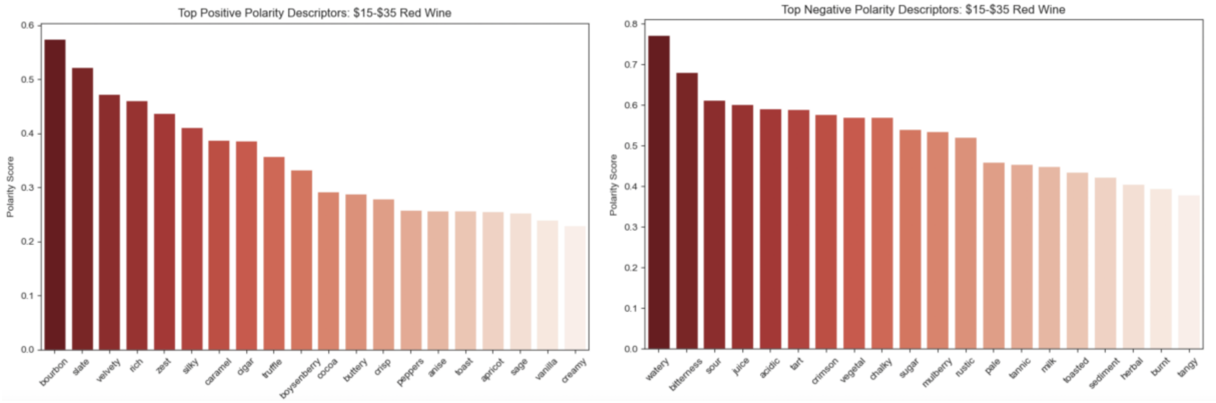
The top-priced white wines exhibit the widest variety of positive descriptors. They feature earth-forward tones like flint and slate, sweet profiles such as strawberries and honey, exotic flavors like coconut, apricot, and tropical fruits, and spice-related terms like zest, smoke, and ginger. Additionally, words related to the color of the wine, such as gold, caramel, and amber, indicate a preference for richer hues compared to various shades of yellow. Furthermore, the top-priced wines, similar to the mid-priced white wines, align with terms like rich and creamy, reflecting consumer preferences for fuller-bodied wines at these price points.

Red Wines



Red Wine Below $15 Descriptor Polarities

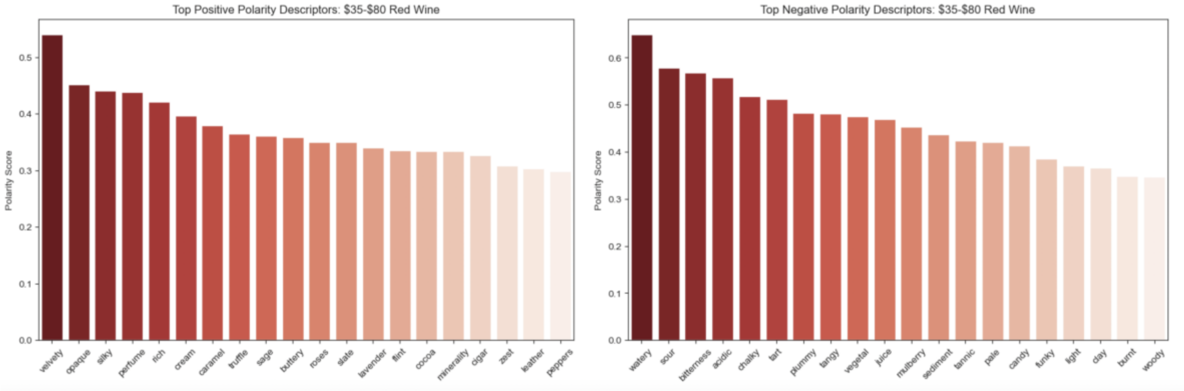
The lowest priced red wines demonstrate a diverse range of positive descriptors, including unique flavors like bourbon, truffle, and cigar, tart berries such as boysenberry and cranberry, earthy tones like minerality and slate, and floral profiles like lavender and sage. Unlike white wines, there are fewer words associated with sweetness that exhibit high positive polarity. Many of the negative terms found in the white wines also appear in the red wines, with “watery” being the most negative term, followed by “bitter”, “sour”, and “chalky”.



Red Wine $15-$35 Descriptor Polarities

While words such as “bourbon” and “velvety " retain their high positive polarity scores from the cheapest red wine tier, terms related to the wine’s structure have a significantly greater presence in the rankings for mid-priced reds. For instance, words like "rich," "buttery," and "crisp" have significantly higher scores for mid-priced wines, and additional structure-related terms like "creamy" also appear. This trend aligns with the preference for richness and smoothness observed in white wines at higher price points. Additionally, floral profiles exhibit less positive orientation in the mid-priced category, with lavender absent from the top 20 and sage scoring lower.

Negative descriptors for mid-priced reds encompass terms that seemingly oppose some of the positive descriptors. For example, while "toast" and "sage" appear in the positive descriptors list, "toasted" and "herbal" appear in the negative descriptors list. This indicates that toasted and floral/herbal profiles elicit polarizing opinions among consumers.



Red Wine $35-80 Descriptor Polarities

In the top-priced red wines, positive descriptors heavily emphasize the structure and aromas of the wine. Smoothness is considered crucial, with an aversion to bitterness and tannins. Other essential factors include a full and rich body, along with floral aromas such as perfume, sage, roses, and lavender. Negative descriptors in this category remain consistent with the other two red wine segments.

Review Generator

Red Wine: Positive Review (Good)	**Red Wine: Negative Review (Good)**
“Light ruby color. Oak, cherry and cranberry on nose. Vanilla, oak and plum on palate. Medium dry with medium acidity. Good balance of flavor and length. A good wine.”	“Tasted like a pinot noir from Argentina. Maybe not what I was looking for but not bad for the price. It was a very light wine, good for those that like a lighter Pinot.”
“Super enjoyable wine. Aromas of cherry and blackberry, notes of cigar box, and dried fruits. Smooth body, medium tannins, and a well balanced acidity. A lovely wine, very easy to drink.”	“Fruity, medium to heavy alcohol body. Smells like blackcurrant. Tastes like it would get better with decanting, but definitely doesn't in my opinion. “
Red Wine: Positive Review (Poor)	**Red Wine: Negative Review (Poor)**
“Dark fruit, leather, dark chocolate and chocolate! Not too dry for me. Nice balance.”	“Light red. Taste of sour cherries, earth and earthy. A hint of tobacco. Nice flavor, though. Good value at 10 a bottle. But for a cheaper, I'm not a fan. “
“It's definitely a light Chianti. Very easy to drink. Not your typical Napa Bordeaux.”	“Quite closed to me when first opened, but mellowed the next day after a day.”

Table 1: Generated Review Examples: Red Wines

Table 1 displays examples of positive and negative wine reviews artificially generated for red wines. Overall, the generated reviews provide detailed descriptions of hypothetical wines, including various flavors, color, acidity, alcohol content, and finish. Many of the reviews clearly indicate that they are referring to a red wine, and it is easy to distinguish between the positive and negative reviews. The writing style of the reviews resembles that of actual wine reviews, with short, often incomplete sentences.

While many of the generated reviews are indistinguishable from actual reviews, some of the reviews show clear indication that they are artificially generated. One recurring pattern is the repetition of words consecutively or in close proximity. For instance, the first “poor” positive and negative examples feature repeated use of the words "chocolate" and "earth"/"earthy," respectively. The next second “poor” positive review is well-written but it becomes evident that the review contains errors when it mentions Chianti (an Italian wine), Napa (in California), and Bordeaux (in France) together. The model clearly lacks sufficient knowledge about each wine region resulting in such mistakes. Furthermore, the second poor-performing negative example exhibits gibberish-like language and incoherently repeats the word "day."

White Wine: Positive Review (Good)	**White Wine: Negative Review (Good)**
“One of my favourite whites, great balance between buttery acidity and ripe fruit flavours and a long finish.”	“Bland for a Sauvignon Blanc, nothing spectacular, but a nice citrus forward wine”
“Fresh crisp pear, lime, melon, green apple, a bit of honey and apricot on the nose and palate. Light medium bodied, high acidity, medium alcohol 13.1 ABV. Paired really well with fresh fish and salad. Very refreshing and easy drinking.”	“Very disappointing. Overripe pears, unripe peaches and hints of lemon. Off dry and high acidity.”
White Wine: Positive Review (Poor)	**White Wine: Negative Review (Poor)**
“4.25 Nice medium full body, deep red. Nice intensity of red fruit on the nose, with some herbs on the palate. Medium bodied, medium acidity. Great wine with a light finish. This is one to explore! “	“This was an okay wine at best. The initial flavor is a bit overwhelming. I don't remember what I would have had but I would recommend drinking the rest of the bottle.”
“I'm not a huge Riesling fan but this is really good! Lemon zest, grapefruit. I just love it. I'm not a Sauvignon fan.”	“A little tart, some fruitiness, a bit oaky. Very nice. I'll buy it again. I'll finish the bottle. “

Table 2: Generated Review Examples: White Wines

Table 2 presents both positive and negative generated reviews for white wines. Similarly, the first two rows of reviews demonstrate the model's ability to produce in-depth and well-written reviews. The “good” reviews are clear and follow a logical order, covering flavors, acidity, and food pairings. However, the “poor” reviews once again highlight the limitations of the model. The first “poor” positive review describes a red wine instead of a white wine, while the second review initially refers to the wine as a Riesling (a German white wine varietal) but ends up discussing Sauvignon (a different white varietal). On the other hand, the first “poor” negative review appears to contradict itself, and the second sounds more like a positive review rather than a negative one.

Despite having an initial dataset of 9.5 million reviews, the scarcity of reviews with ratings below 3.0–especially in white wine–significantly limited the size of the training dataset for the review generator. This limitation likely contributed to some of the poor-performing instances observed.

Conclusion

This study explored two key aspects related to wine reviews and consumer preferences: identifying positive and negative attributes within different wine subsets and developing a review generator based on user style. By analyzing a comprehensive dataset of wine reviews from Vivino, non-obvious qualities associated with positive and negative sentiment for different wine segments were uncovered. These findings provide valuable insights for wine producers, enabling them to better understand consumer desires and tailor their products accordingly.

Furthermore, a review generator was successfully trained using transfer learning techniques, leveraging OpenAI's GPT2 model. While the generated reviews demonstrated an ability to mimic the style of Vivino users, there were limitations and patterns that indicated the artificial nature of the generated content, in some cases. Despite these limitations, the review generator serves as a foundation for further advancements and recommendation systems, such as a wine pairing recommender, which can help producers gain deeper insights into consumer preferences and enhance user experience. Overall, this study contributes to the field of wine analysis by combining deep learning techniques, sentiment analysis, and generative models to provide actionable insights and open possibilities for future research and industry applications.

Sources

[1] Statista. https://www.statista.com/outlook/cmo/alcoholic-drinks/wine/united-states
[2] Stacker. https://stacker.com/food-drink/spectacular-wineries-and-vineyards-around-world#:~:text=An%20overnight%20stay%20at%20a,of%20those%20are%20in%20California).
[3] The Wine Economist.
https://wineeconomist.com/2019/04/09/imports/#:~:text=Imports%20account%20for%20about%20a,it%20was%2025%20years%20ago.
[4] vivino.com
[5] Wine Direct. https://www.winedirect.com/learn/blog/winery-care-about-product-reviews-2/
[6] Stack Exchange
https://stats.stackexchange.com/questions/343763/fine-tuning-vs-transferlearning-vs-learning-from-scratch