University of South Wales
Prifysgol De Cymru

Faculty of Computing, Engineering and Science

MSC in data science

Data Mining and Statistical Modelling - MS4S08

Practical coursework -1

Assessment No: **1**

Wine Review Analysis Report

Student Name:

Thayanantham Nitharshana

Student Number:

30119539

Table of Figures

1. Task A - Text Mining

Data Preprocessing and Analysis of Wine Reviews Dataset

The objective of this task is to clean, preprocess, and analyze wine reviews to extract meaningful insights and visualize key trends. The dataset consists of wine reviews written by tasters, primarily focusing on the description column.

1.1 Handling Missing Values

Missing values in numerical column price and points were imputed using median values. Missing textual descriptions rows were removed.

...1	country	description	designation
0	4396	4336	44608
points	price	province	region_1
8672	17289	8732	29021
region_2	taster_name	taster_twitter_handle	title
84824	33646	38399	8672
variety	winery		
8674	8673		

Figure 1 Missing values in our dataset

...1	country	description	designation
0	60	0	40272
points	price	province	region_1
0	0	4396	24685
region_2	taster_name	taster_twitter_handle	title
80488	29310	34063	4336
variety	winery		
4338	4337		

Figure 2 Missing values after handling

The points and price columns initially had 4,336 and 12,953 missing values, respectively, but imputation reduced them to zero. The description column had 4,336 missing values, which were removed. Other columns with significant missing values included designation (40,272), region_2 (80,488), and taster_twitter_handle (34,063), region_1, region_2, and taster_name were retained as they were not used in the analysis.

1.2 Handling Outliers

There are only two numeric columns in the dataset such as price and points. In this data set price column contains outlier values, whereas the points column does not have any outliers. Outliers in price were identified using the Interquartile Range (IQR) method and handled through capping techniques.

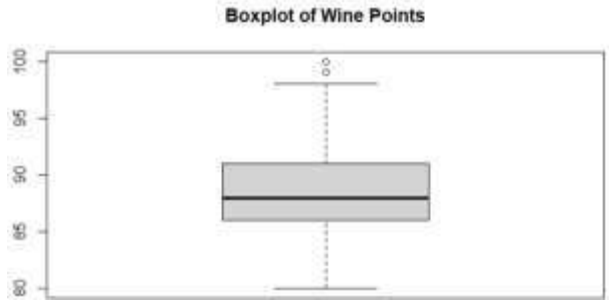

Figure 3 Boxplot of point before handling outliers

Figure 4 Boxplot of price before handle outliers

Replacing the outlier values with their respective boundary values (lower or upper bounds) to ensure the data remains consistent and usable for analysis.

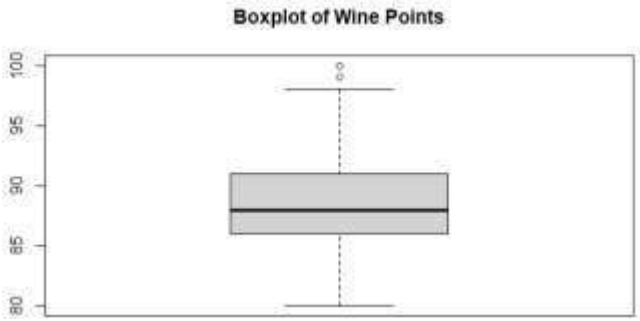

Figure 5 Boxplot of points after handling outliers

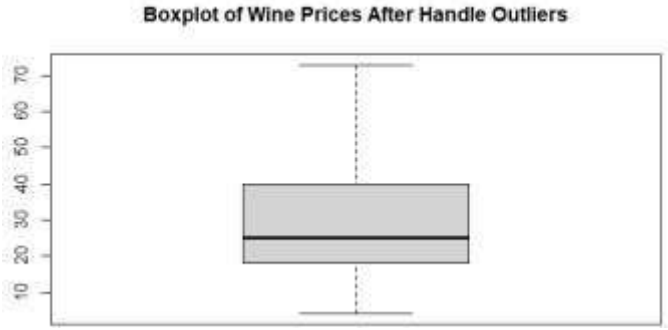

Figure 6 Boxplot of prices after handling outliers

1.3 Selecting and Sampling the Data

The dataset, containing 133,189 rows, was considered quite large. For computational power concerns, a subset of 20,000 reviews was sampled for quicker analysis. The description column, containing wine reviews, is the main analysis focus. Title may have useful keywords but is less relevant. Variety is used for further topic modeling. Description is correlated with points for sentiment classification.

1.4 Data Preprocessing

To prepare the text data for analysis, the following preprocessing techniques were applied:

- Lowercasing: Convert all text to lowercase for uniformity and maintain consistency.
- Removing Punctuation: Remove all punctuation marks to focus only on meaningful words.
- Removing Numbers and Special Characters: Remove numeric values and non-alphabetic characters unless they are relevant.
- Removing Stop Words: Eliminate common words like "the," "and," "is," and some custom words like "wine", "flavors", "taste", "bottle", "glass", and "drink" that do not add meaning.
- Tokenization: Split text into individual words or tokens.
- Stemming: Reduce words to their root forms.
- Lemmatization: Convert words to their base forms based on context (e.g., "running" to "run").
- N-grams: Generate sequences of n words to identify common phrases.

```
[1] "Happily, each new vintage of BÃ©soleil is a bit larger than the previous year, so more people
can taste what a joy this wine is. In 2008 it's 82% Grenache, 9% Syrah and 9% MourvÃ¨dre. As with
all the Betz wines, it is intensely aromatic. The wine glass fairly explodes with scents of grape,
plum, raspberry and cherry. In the mouth grace notes offer hints of leaf and fresh herb. Plush and
irresistible."
[2] "Tropical aromas of lychee and banana come with a varietally familiar note of bath soap. This
is oily feeling, but with integrated acidity it isn't ponderous. Pithy citrus flavors are backed by
a hint of bitterness, while the finish is juicy, a touch burning and tastes of nectarine and
mango."
[3] "This big, dramatic wine has a dark color and bold aromas of wood smoke and black pepper.
Flavors are concentrated but also dry and appetizing. Firm tannins add grip and seem to help the
peppery, smoky accents linger on the finish. Best after 2020."
[4] "This is appealing from the beginning, with fresh raspberry, strawberry and mineral aromas. It
feels nice and bright, with flavors of citrus, pink grapefruit and green herb. Finishes with
forward acidity and mild flavors of sage and tarragon."
[5] "Central Otago Pinot Noir doesn't come along very often at this price. This medium-weight,
peppery wine doesn't show a ton of fruit, but does deliver some pleasantly spicy complexity and a
mouthwatering chocolaty finish. Drink now."
[6] "Reduced on the nose, this massively intense and tannic wine is big, dense and juicy, broadly
etched in leather, blackberry, currant and black licorice. Full-bodied and oaky, it shows the
concentration of the vintage and the ripeness of the appellation and style."
```

Figure 7 Sample of Description Before data preprocessing

word <chr>	stemmed_word <chr>	lemmatized_word <chr>
happily	happili	happili
vintage	vintag	vintag
bsoleil	bsoleil	bsoleil
bit	bit	bite
larger	larger	large
previous	previou	previou

Figure 8 Sample of Description after data preprocessing

1.5 Exploratory Data Analysis (EDA) and Visualization

The goal is to understand **frequent words** in wine reviews before moving to deeper analysis. The following visualizations will be generated.

1.5.1 Word Frequency Analysis

Identified common words such as "fruit," "aroma," "finish," and "acid." In this result, N/A values, stop words such as "is," "are," "the," and meaningless custom words like "wine," "bottle," and "glass" have been removed, while words like "fruit" and "fruits" have been treated as the same word during lemmatization to improve accuracy.

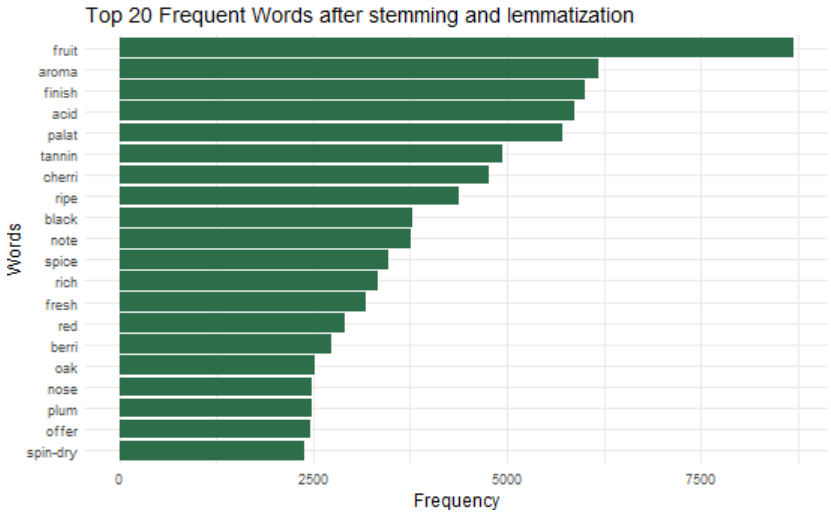

Figure 9 Bar Chart of Top 20 Word Frequencies

1.5.2 Word Cloud Visualization

Displays words in varying sizes based on their frequency.

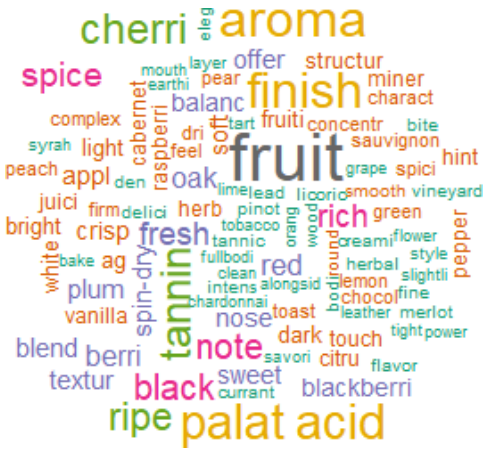

Figure 10 Word Cloud after lemmatization.

1.5.3 Bigram Analysis

Revealed common word pairs like "black cherry" and "fruit flavors."

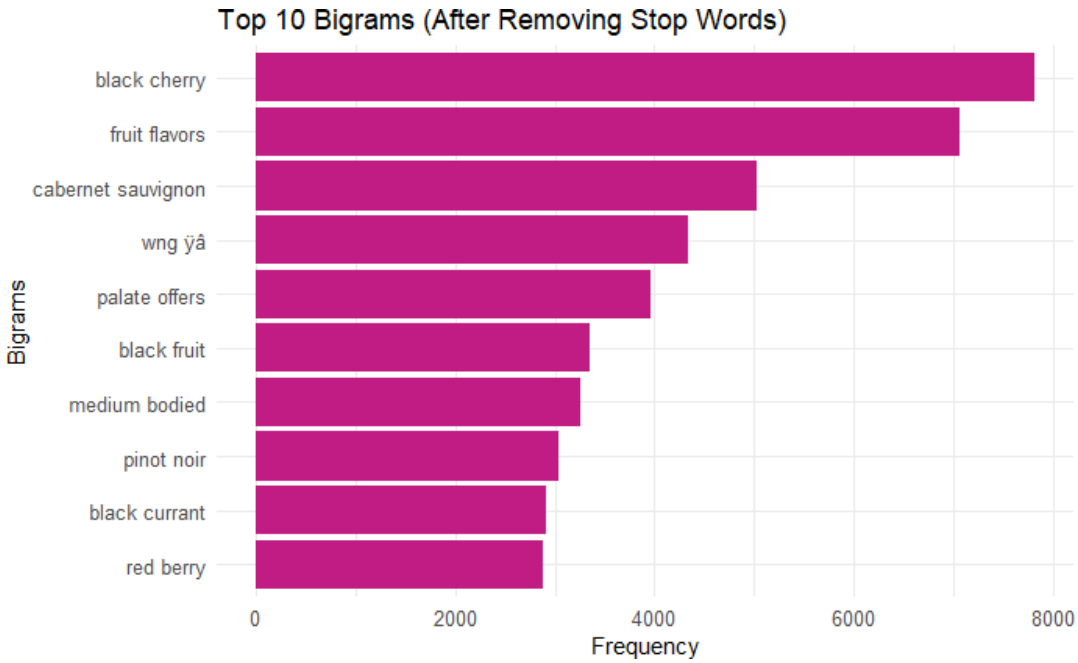

Figure 11 Bar Chart of Top 10 Bigrams

1.5.4 Network Graph

Showed relationships between frequently co-occurring words.

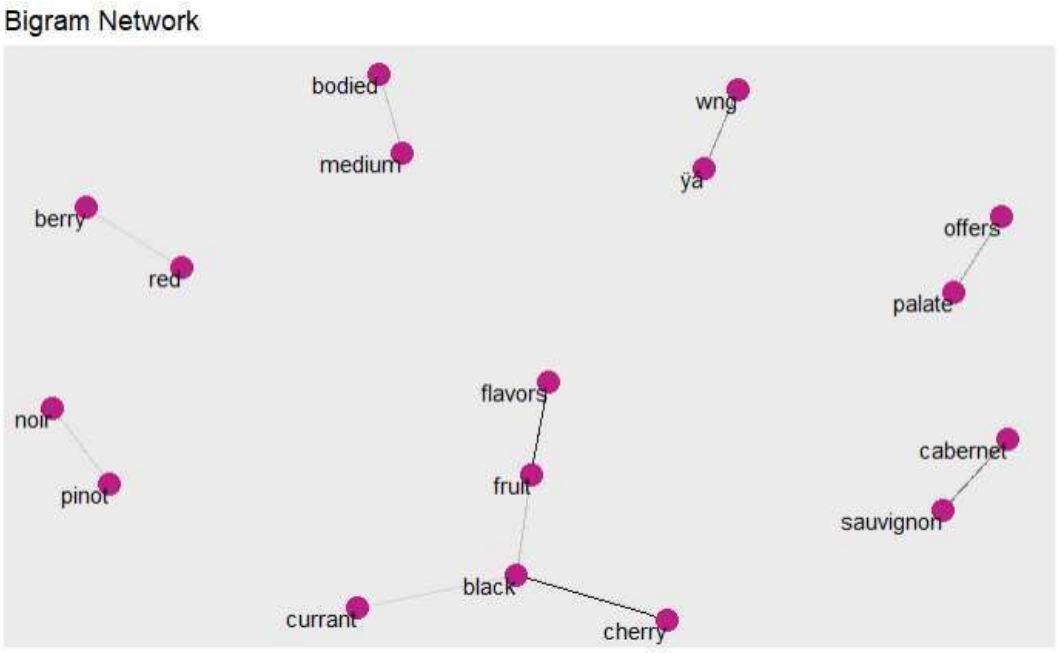

Figure 12 Network Graph for co-occurring words

1.6 Insights from Text Mining and Conclusion

The word cloud highlights key wine-related terms like "fruit," "aroma," and "palate." After removing common stopwords, domain-specific words were preserved, providing insights into consumer preferences and expert evaluations. This targeted approach ensures a focus on relevant vocabulary for a more effective analysis of the wine industry. Common bigrams emphasized key wine descriptors, improving interpretability. Preprocessing enhanced data quality by removing noise. Overall, text mining provided a structured understanding of wine characteristics, aiding in sentiment analysis and topic modeling.

2. Task B - Sentiment Analysis

The goal is to analyze the sentiment of wine reviews and classify them into positive, negative, or neutral categories using sentiment analysis techniques.

2.1 Sentiment Classification Techniques

Sentiment analysis was performed using multiple lexicons.

2.1.1 AFINN Lexicon

Assigned sentiment scores to words.

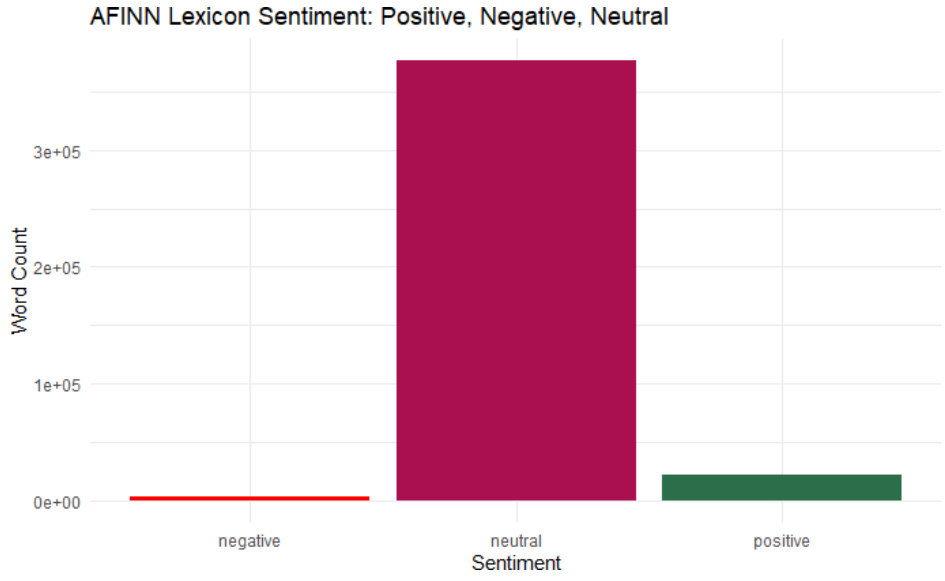

Figure 13 Bar chart of AFINN Lexicon

Neutral Sentiment Dominates: The tallest bar, in dark maroon, represents neutral words, indicating that most of the words analyzed have no strong positive or negative sentiment.

Positive vs. Negative – The green bar represents positive words, while the red bar represents negative words. There are more positive words than negative ones, but both are significantly lower compared to neutral words.

This could suggest that the text analyzed is mostly neutral, with a small proportion of positive and negative sentiments. This is common in general text analysis, where many words do not carry strong emotional weight. **If we omit neutral reviews, positive reviews take the dominant position in this analysis**.

2.1.2 Bing Lexicon

Classified words as positive and negative. This graph shows that **most people write positive reviews** about wines, while fewer reviews express negative sentiments. The **higher green bar** means more customers enjoyed their wine, whereas the **shorter red bar** shows fewer complaints or criticisms.

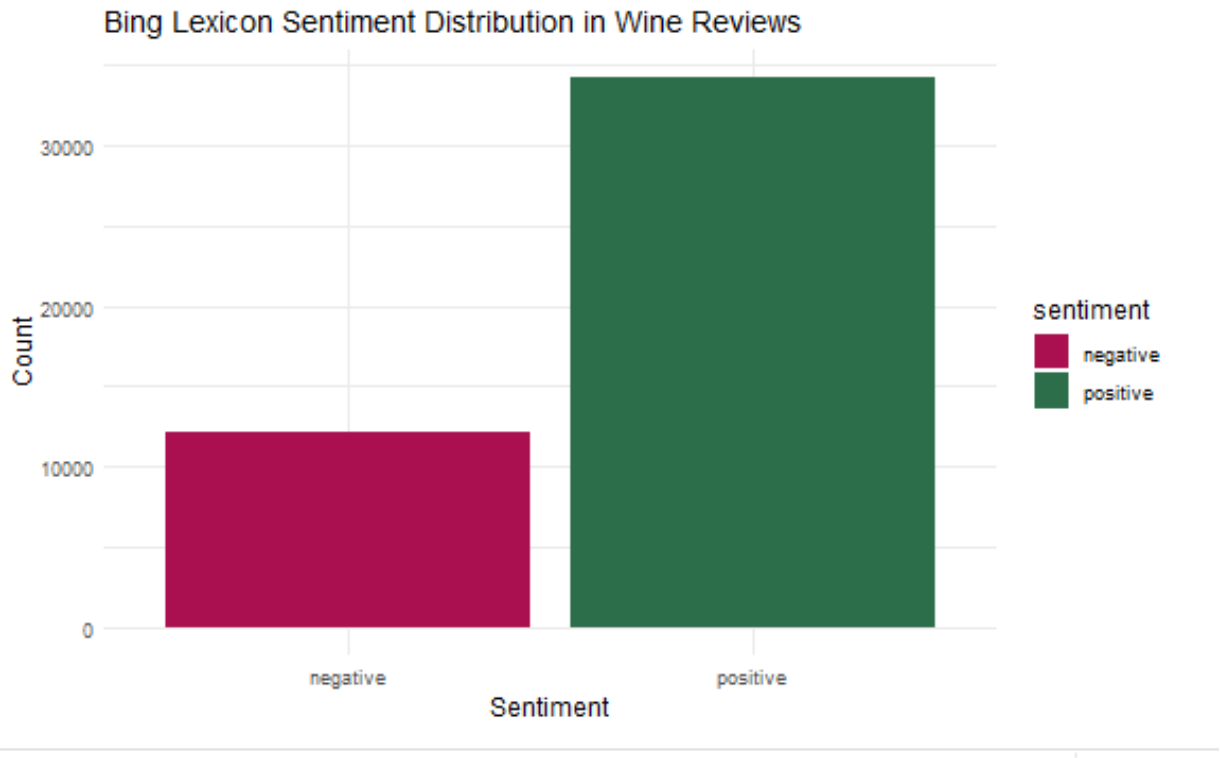

Figure 14 Bar chart for Bing Sentiment analysis

2.1.3 NRC Lexicon:

Classified words into emotions like joy, anger, and fear. The bar chart shows "Positive" as the most frequent sentiment, followed by "Negative." "Joy" and "Anticipation" are also common but less dominant. Other emotions disgust, anger, trust, sadness, fear, and surprise are infrequent. Overall, the text has a strong positive bias with notable anticipation.

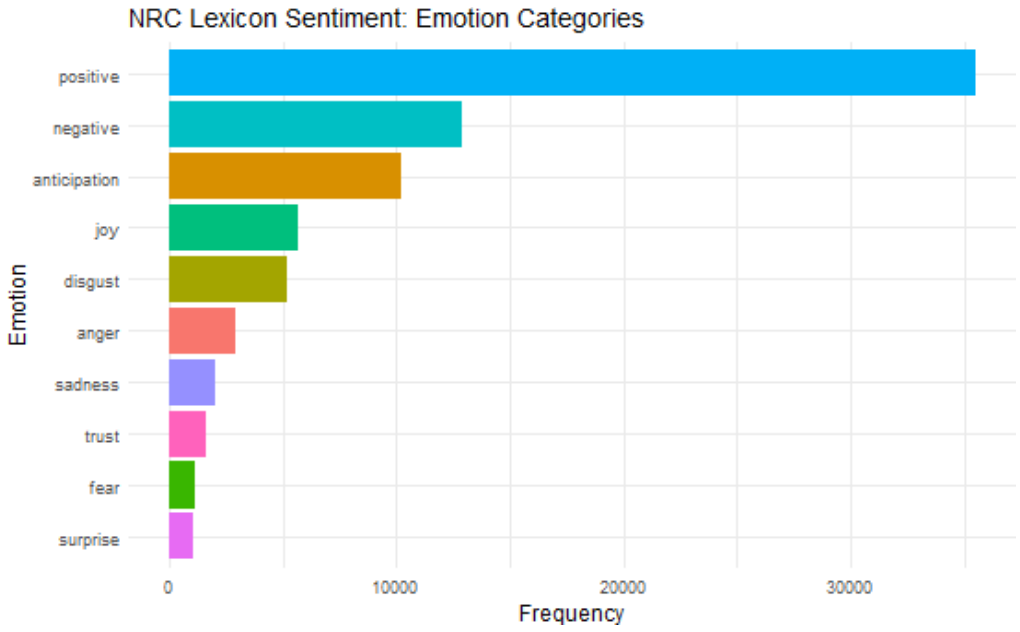

Figure 15 Bar chart of NRC Lexicon Sentiment

2.2 Visualization using Sentiment Scores

2.2.1 Word Clouds

Illustrated frequent positive, negative, and Neutral words.

Figure 16 Positive Word Clouds

Figure 17 Negative Word Clouds

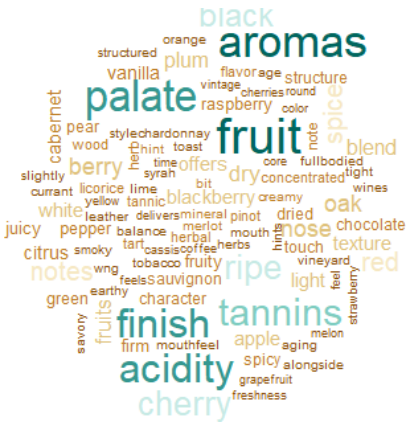

Figure 18 Neutral Word Cloud

The word cloud highlights key themes in wine reviews. Positive words like "fresh," "soft," and "rich" dominate, emphasizing sensory appeal, quality, and emotional warmth. Negative words such as "lemon," "hard," and "bitter" suggest unpleasant experiences and challenges. Neutral terms like "fruit," "aromas," and "tannins" focus on objective wine descriptions, reflecting a balanced vocabulary without strong sentiment bias.

2.2.2 Bar Charts

Displayed the most common sentiment-bearing words.

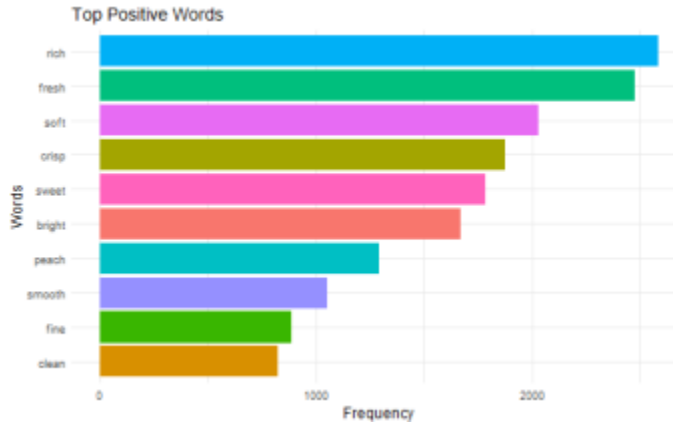

Figure 19 Top Positive words

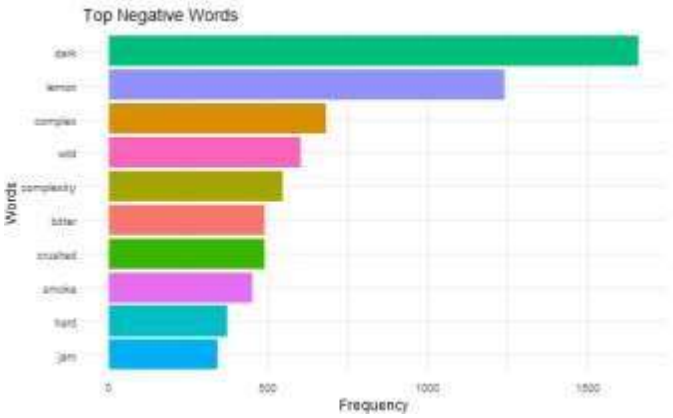

Figure 20 Top Negative words

The bar charts highlight the most common positive and negative words in wine reviews. Positive terms like "fresh," "rich," and "soft" dominate, reflecting customer appreciation for these qualities. Negative words such as "dark," "lemon," and "dense" indicate common complaints. The charts emphasize key descriptors, showcasing preferences for freshness and richness while noting concerns about intensity and density.

2.2.3 Boxplot Analysis

Showed correlation between sentiment scores and wine points. The boxplot shows that higher-rated wines (90+ points) tend to have more positive reviews, while mid-range wines (85-90 points) exhibit greater variability, indicating mixed opinions. Top-rated wines (97-100 points) have positive sentiment outliers, while lower-rated wines (80-84 points) receive mostly negative feedback. This suggests a strong correlation between ratings and sentiment, with higher ratings reflecting better reception, while mid-range ratings reveal subjective variation. These insights help wine producers understand how customer sentiment aligns with expert ratings.

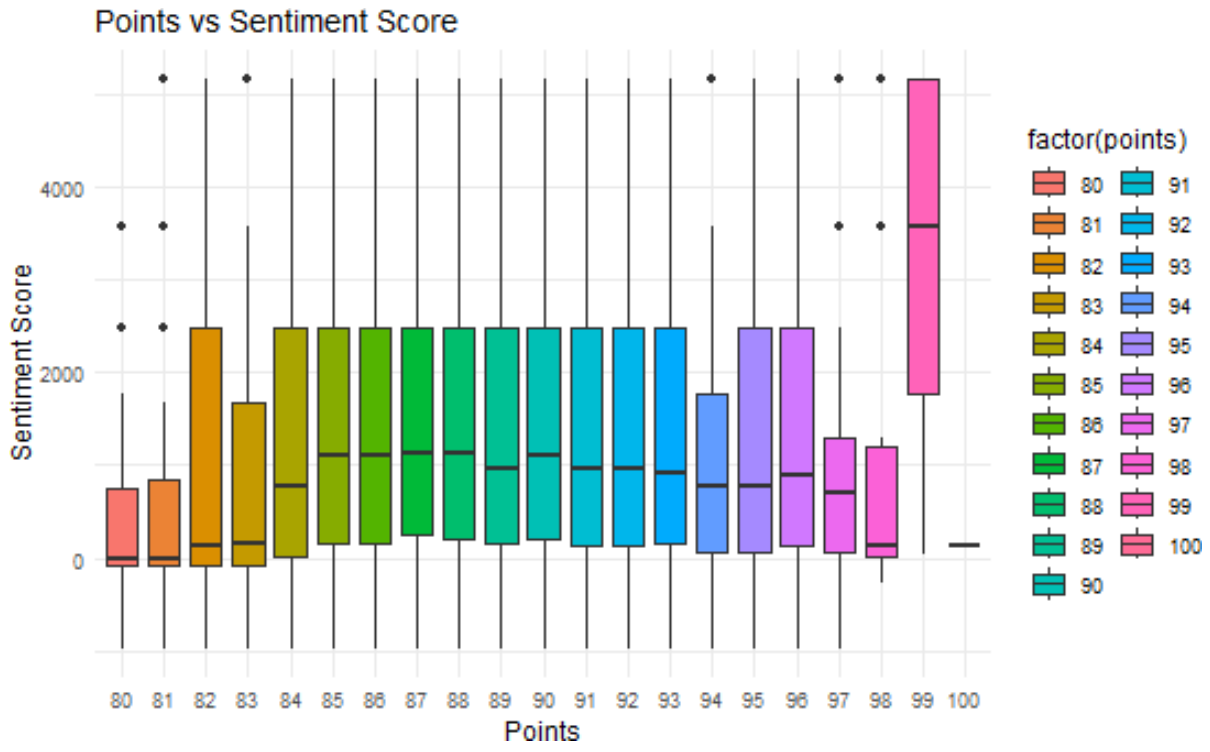

Figure 21 Box plot of Points vs. Sentiment Score

2.3 Insights from Sentiment analysis and Conclusion

The sentiment analysis shows that most words are neutral, but positive sentiment dominates when neutral terms are excluded, indicating a favorable perception. Words like "fresh," "rich," and "soft" emphasize quality, while negative terms like "dark," "lemon," and "dense" reflect common complaints. Emotion analysis highlights a positive bias, with anticipation and joy being prominent. The correlation between ratings and sentiment confirms that higher-rated wines receive more positive feedback, while mid-range wines show mixed opinions. Overall, the sentiment is mostly positive, with some negative aspects influencing perceptions.

3. Task C - Topic Modeling

The goal of topic modeling is to uncover hidden themes in wine reviews, allowing us to understand common topics that wine tasters discuss. This helps identify trends, preferences, and key characteristics that influence customer opinions.

3.1 Topic Identification Using Latent Dirichlet Allocation (LDA)

LDA was applied to cluster reviews into distinct topics. The optimal number of topics was determined based on coherence scores and interpretability.

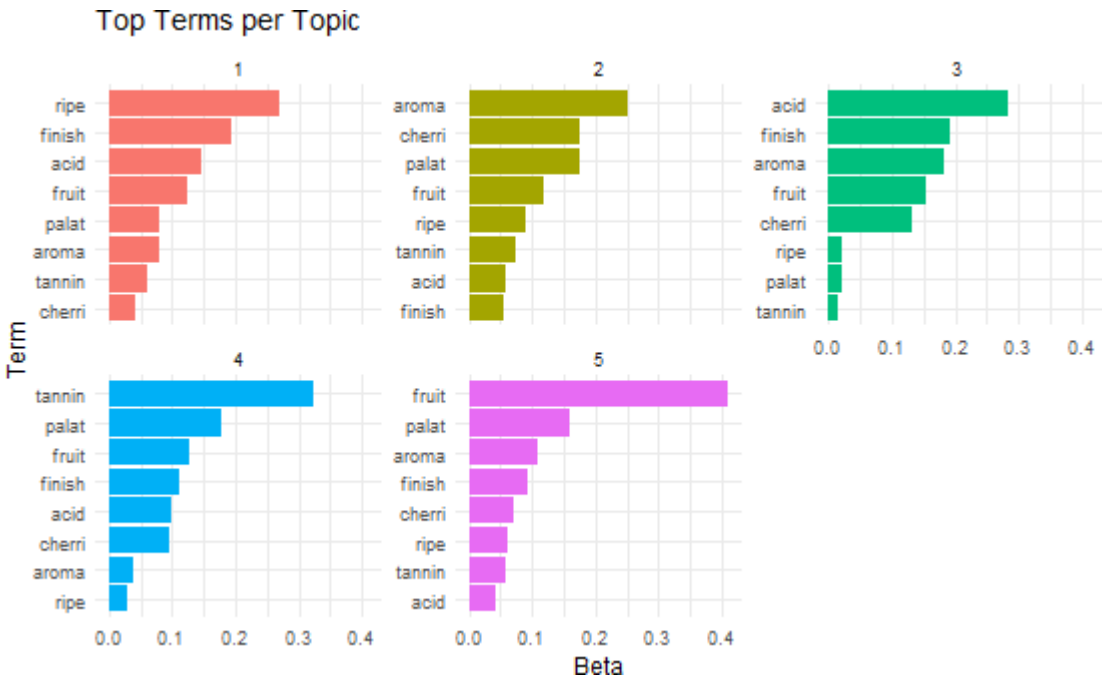

Figure 22 Bar chart of Top terms per Topic

Dominant Topics Identified

- **Topic 1 (Red) – Ripeness & Acidity**
  - Key terms: ripe, finish, acid, fruit, aroma, tannin, cherri.
  - Possible theme: This topic likely discusses the ripeness of fruits and their acidity levels, possibly in the context of wine or fruit tasting.
- **Topic 2 (Olive Green) – Aroma & Flavor Profiles**
  - Key terms: aroma, cherri, palat, fruit, ripe, tannin, finish.
  - Possible theme: Focuses on aromas and tasting notes, possibly describing the sensory aspects of wine, food, or beverages.
- **Topic 3 (Teal Blue) – Acidity & Mouthfeel**
  - Key terms: acid, finish, aroma, fruit, cherri, ripe, palat, tannin.
  - Possible theme: Discusses acidity and finish, which are important factors in wine or beverage quality. The mention of "palat" suggests it relates to mouthfeel and taste balance.

- **Topic 4 (Blue) – Tannin & Structure**
  - Key terms: tannin, palat, fruit, finish, acid, cherri, aroma, ripe.
  - Possible theme: This topic likely focuses on tannins and their effect on texture, particularly in wine, as tannins influence astringency and aging potential.
- **Topic 5 (Pink/Purple) – Fruit-forward Characteristics**
  - Key terms: fruit, palat, aroma, cherri, ripe, tannin, acid.
  - Possible theme: Likely describes fruit-forward wines or beverages, emphasizing flavors of cherries, ripeness, and overall fruit intensity.

3.2 Insights from Topic modeling and Conclusion

The topics are strongly linked to wine tasting and flavor analysis, possibly from a dataset of wine reviews or descriptions. LDA has successfully grouped related tasting attributes (acidity, tannins, aroma, and fruitiness) into distinct clusters. These insights could be valuable for wine critics, sommeliers, or beverage analysts looking to categorize products based on tasting notes.

4. Task D: Further Exploration

This section will extend our analysis by applying advanced techniques, integrating all dataset variables, and discussing potential future work.

4.1 Additional Techniques Explored

4.1.1 Topic Coherence Evaluation

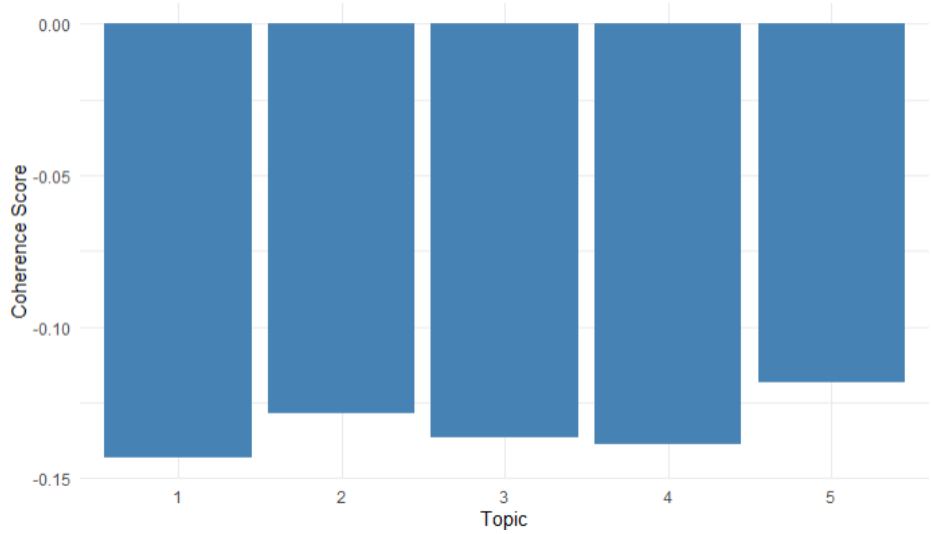

Figure 23 Bar Chart of topic coherence

Measured topic interpretability. The bar chart shows topic coherence scores for five topics, with the x-axis representing the topics and the y-axis displaying their coherence scores. The negative scores suggest that the topics are poorly defined, with weak word associations, making interpretation challenging. This analysis is used in topic modeling (e.g., LDA) to assess the meaningfulness of topics.

4.1.2 Interactive Visualization (LDAvis)

Enabled dynamic exploration of topic distributions. The interactive visualizations (Figures 24-28) dynamically explore topic distributions in wine reviews using LDAvis. Each figure represents a distinct topic, illustrating key terms and their relationships. This enables deeper insights into how wine tasters describe different aspects, such as fruitiness, acidity, tannins, and oak influences.

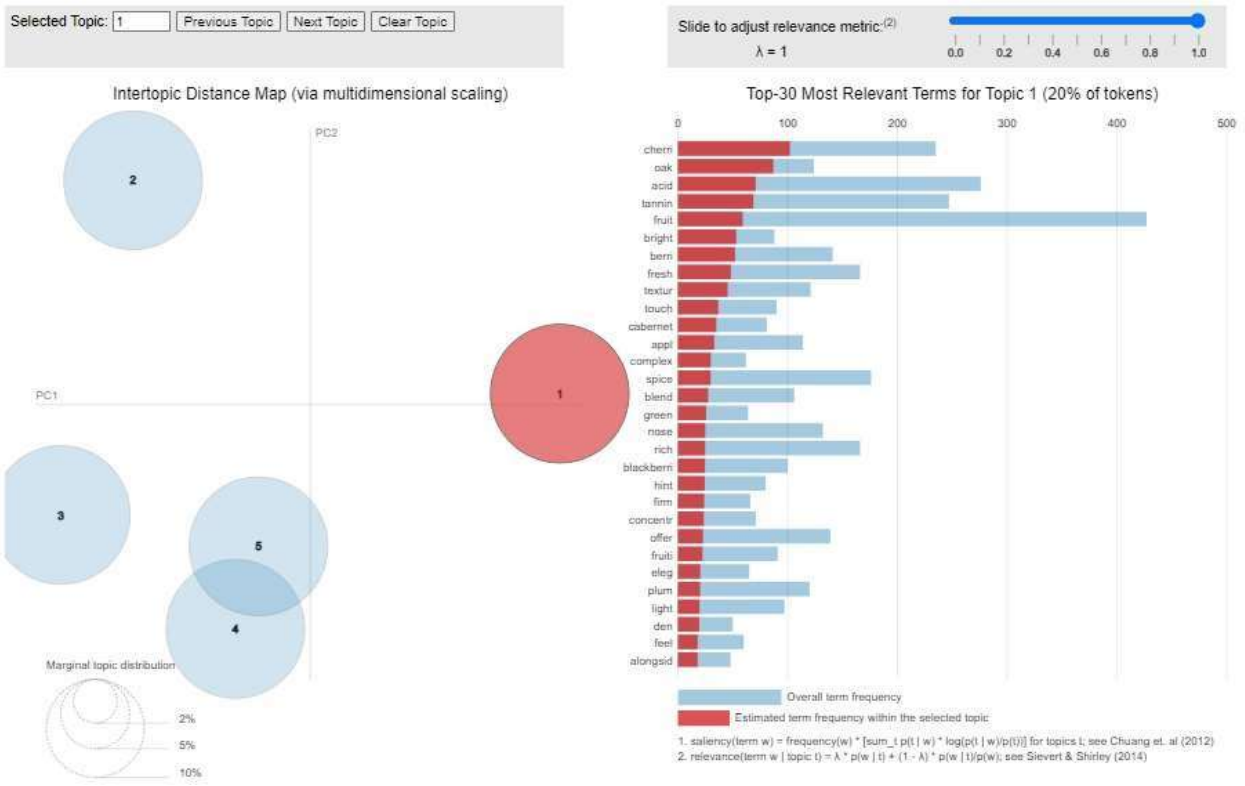

Figure 24 Dynamic exploration of topic 1

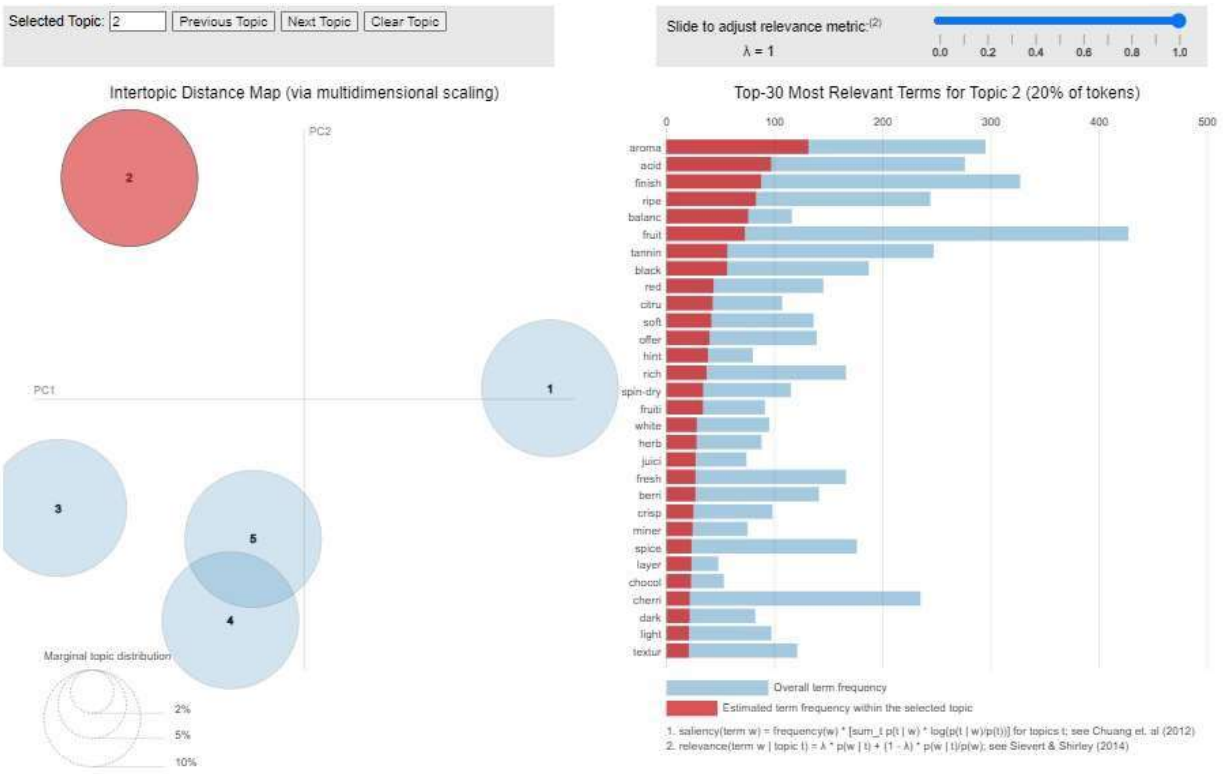

Figure 25 Dynamic exploration of topic 2

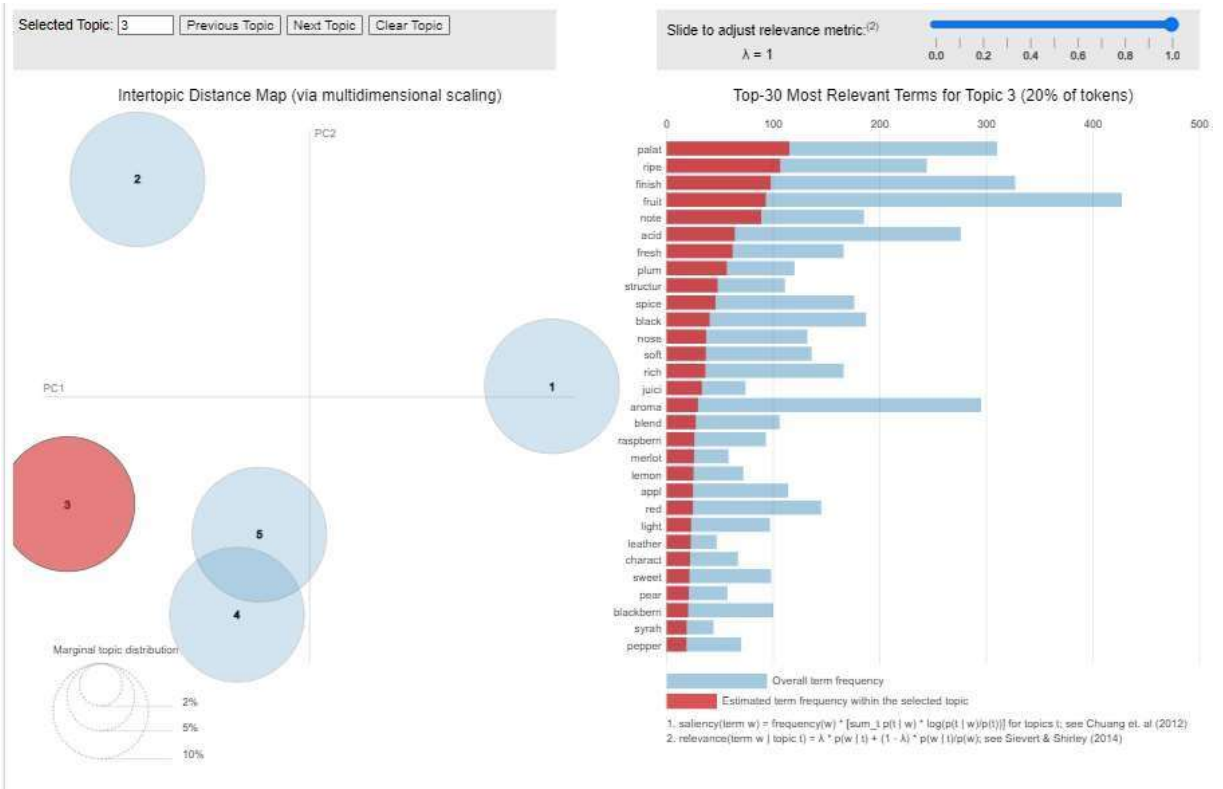

Figure 26 Dynamic exploration of topic 3

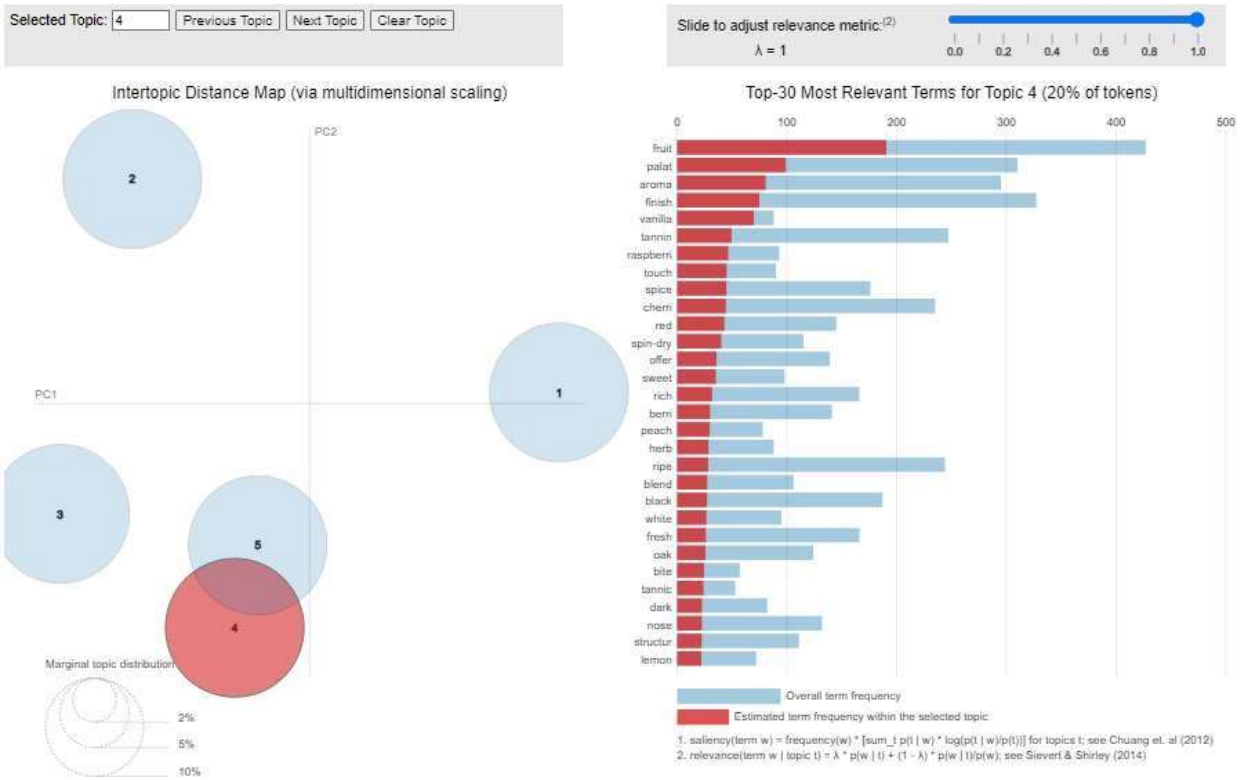

Figure 27 Dynamic exploration of topic 4

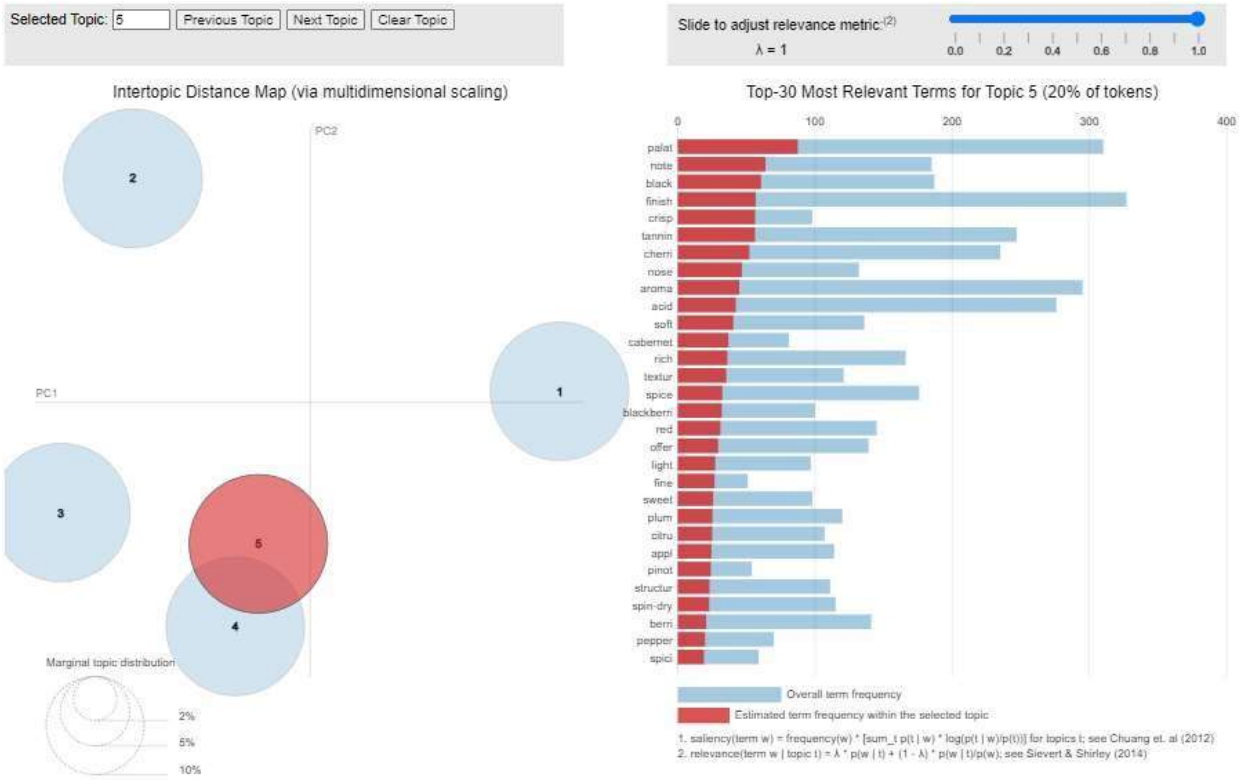

Figure 28 Dynamic exploration of topic 5

4.1.3 Topic Distribution across Wine Types

Analyzed correlations between wine varieties and topics.

Topic 1: Fruity & Floral – Wines like Pinot Noir and Riesling with cherry, peach, and citrus notes. *Example:* "Fresh cherry and raspberry flavors with a floral aroma."

Topic 2: Oak & Vanilla – Aged wines like Chardonnay and Cabernet Sauvignon with vanilla, butter, and spice. *Example:* "Buttery texture with vanilla and toasted oak notes."

Topic 3: Earthy & Tannic – Old World wines like Bordeaux blends with firm tannins and earthy notes. *Example:* "Tobacco, mineral finish, and structured tannins."

Topic 4: Bright & Crisp – High-acid whites like Sauvignon Blanc with zesty, refreshing qualities. *Example:* "Lively citrus notes with crisp acidity."

Topic 5: Bold & Spicy – Reds like Syrah and Zinfandel with deep fruit and spice. *Example:* "Blackberry, peppery spice, and a smoky finish."

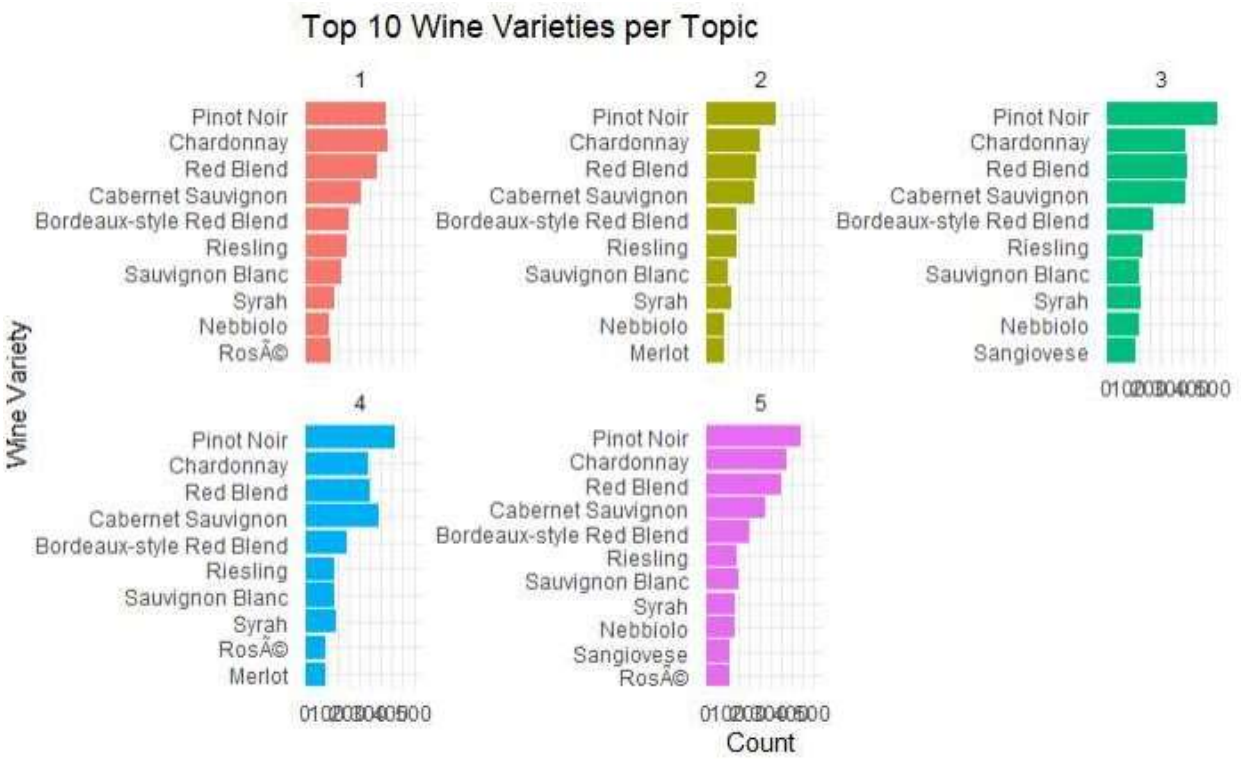

Figure 29 Bar Charts of wine varieties per topics

4.2 Future Work and Recommendations

Sentiment Analysis Refinement - Future improvements could involve fine-tuning the sentiment analysis model with domain-specific wine-related lexicons or custom-trained classifiers to improve accuracy in sentiment classification.

Advanced Topic Modeling - Using more sophisticated models like BER Topic or hierarchical clustering could provide better topic differentiation and reveal deeper semantic relationships.

Wine Feature Analysis - Investigate how specific wine features (e.g., acidity, tannins, flavor notes) influence overall sentiment, potentially providing wine producers with insights to improve their products.

User Feedback Incorporation - Future work could involve incorporating user demographic data (age, location, etc.) to understand how different types of tasters perceive wines differently, adding a layer of segmentation to the analysis.

Cross-Model Comparison - Compare the results from sentiment analysis and topic modeling to see how well the sentiment aligns with the topics identified, providing a clearer understanding of the relationship between wine descriptions and tasters' emotions.

4.3 Conclusion

This study applied text mining, sentiment analysis, and topic modeling to wine reviews, revealing key insights into wine characteristics and consumer sentiment. While the results were meaningful, refining topic coherence, expanding dataset coverage, and integrating additional features could further enhance the analysis. Future research should focus on leveraging advanced models and exploring new segmentation techniques for a deeper understanding of wine consumer preferences.

References

Lecture Notes.

Wickham, H. (2017). *1 Introduction | Advanced R*. [online] Hadley.nz. Available at: https://adv-r.hadley.nz/introduction.html [Accessed 3 Feb. 2025].

Grolemund, G. and Wickham, H. (n.d.). *2 Introduction | R for Data Science*. [online] *r4ds.had.co.nz*. Available at: https://r4ds.had.co.nz/explore-intro.html.

Manning, C.D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Rpubs.com. (2022). *RPubs - Wine Analysis Project*. [online] Available at: https://rpubs.com/kamriefoster/984584 [Accessed 3 Feb. 2025].

Video File Path

https://www.youtube.com/watch?v=HnXtRwa2Tg4