Sales Forecasting – Basis Consumer behaviour using Sentimental Analysis

Amarnath Venkataramanan
MSc FinTech
National College of Ireland
Dublin, Ireland

***Abstract*— This article describes how sentiment analysis places a vital role in the sales of a product. Depicting Consumer behavior and opinion aids firms to improve their customer satisfaction by using these machine learning techniques firms to tend to follow the strategies and opportunities in their products and services. Recent growth in retail industry attempts to improve the long established SFT techniques which has a gap when analyzing large data of consumer reviews efficiently, to determine the effectiveness of this approach models such as SVM, Linear Regression and KNN. Although there has been a procedure of using sentiment analysis for tracking the sales forecast by considering the customer's tastes, preferences and perception towards the product by using algorithms and AI methodologies, it is still in a very nascent stage as complete exploration have still not been adequately done. Also, the senior management does not get a complete insight of how these ratings derived by perceptions and expressions affect the sales forecast, these are still used to predict to a large extent, and this definitely helps in product improvements. Although several methods have been proposed to use product reviews for business intelligence like the KNN, SVM and Linear Regression, still great deal of work are kept in anvil. The aim of this study is to extract and derive patterns and develop them into business insights based on aspect-based sentiment analysis which can be utilized for product improvement recommendations for higher market penetration and effective pricing mechanism.**

Keywords— Supply chain management, Blockchain, GDPR, RegTech.

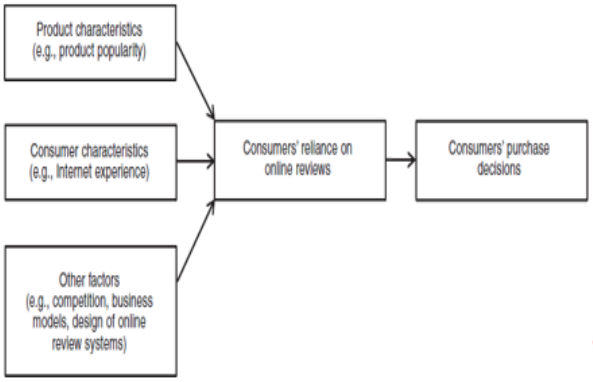


Figure1 -

I. INTRODUCTION

Background: Traditionally, before the penetration of internet, the modus operandi of opinion and sentiment was based on the review from family, friends and relatives and placed critical factors for individual decision-making process. For examples, asked friends to recommend an auto mechanic or who they are planning to vote in elections, reference letter for jobs applications or to buy consumer durables. However, the trend has changed since the internet penetration in later part of 90's and made possible to get reviews and opinions from people that we have never heard. The reviews are based on people neither personal connections nor well known to individual. Besides through social media, more and more people are making views and reviews to strangers and help them to take informed decision.

Currently, most of the companies sell the product online to consumer by B2C and creates the demand for the information on opinion and sentiment and from CEO to marketing people are curiosity to know on the product "what consumer think" and has been play critical for the product acceptances and sales forecasting

Business Problems: The long-established sales forecasting techniques (SFT) used by the company poses challenge considering the big data. Additionally, the current SFT is not serving company to predict the accurate forecasting due to fact that the reviews and comments are structured or unstructured through online. Moreover, these techniques are difficult to distinguish the timely isolated problem and failure to select right activities to deepen the customer relationship in turn to increase sales.

The solution: The sales strategy has been changing in a paradigm shift from reactive to proactive and from intuition feelings to data driven approach. Rather than people to put their effort to source and analyze the data to understand the prediction, the systems itself be automated in such a way to analyze the data in real time and guide the user on the way forward. Additionally, the systems should capable of distinguish from activities that does not work based on the learning and self-optimize to guide the user to improve on continuous basis. Therefore, considering the challenge faced by organization on sales forecasting with the big data and unstructured form, it is pertinent to convert the challenges into opportunities; the automation of the machine language using statistical algorithm enables organization to increase the top line.

Based on the above, the project aims to focus the answer on below question
"How can firm forecast the sales based on sentiment analysis?"

The sentimental analysis is the process of analyzing the text data of the set of people to extract the emotion of the people. The sentiment analysis plays a major role to predict the sales. The key driver is due the fact that people all over the world shares all their thoughts and interests via social networking like Face book, twitter, etc. and even in mobile applications like What's App and Instagram. For Example- In e-commerce, based on the reviews and ratings the buyer will predict the quality of the product and decides whether to buy the product or not and here fake reviews could misguide the

buyers either way like make them to buy the fake products or not to buy the good products

This paper includes addressing the following Specific Items
Firstly, we are going to look the data sets to support the questions with the help of web scraping and extract and perform those data under CRISP –DM approach (data cleaning and data exploration) to prepare data sets for the sentimental analysis.

Secondly, we are going to select the models to be applied for forecast analysis. The approach is based on Customer's opinion on the product/goods/services which can be mined via sentiment analysis using Machine Language. Based on earlier literature review that models like KNN, SVM and other ML are suitable for applications and data sets for separated into training and testing to standardize the model to forecast the sales based on sentiment analysis.

Finally, other aspects to looking into the project is regarding the impact of fake reviews cause the customers to buy the fake products which in turn ends up in the loss of trust of customers on the ecommerce.

To have an empirical study on above situation, this report chosen wine data sets from Kaggle's and attempt to determine the connection and implication of variables between Sales and price, demographic as per information available in data sets. The motivation of this report is to predict the sales forecasting using three machine languages
to include the limitation & Outcome of the project

This paper is organized as follows. Section II discusses the Literature review based on Sentimental analysis, data sets, 3 Machine Languages. Section III describes data and variables using summary statistics and Data Mining Methodology applied under CRISP-DM model. Section IV deals with the evaluation of all 3 methods applied and presents the empirical results and Section V concludes the paper

II. LITERATURE REVIEW

Sentimental analysis:

Previously customers choose or buy products either by word of mouth or by advertisements, Due Recent popularity of Web 2.0 area which has led to various applications platforms for reviewing or commenting on products such as blogs, social forums and social media such as Reddit, medium and Facebook, Twitter etc. Customers now play an important role in developing the brand value of the company, their review and opinions act as a vital part of buying or selling a particular product. From this reviews and ratings (positive or negative) generated through social media companies are implementing various ways to analyze those reviews based on various scales such as (e.g., extremely positive, neutral, or somewhat negative), The evaluation of sentiment is filled with difficulty

For example, to decipher the exact meaning of the review for instance, this analyzed by using models such as mean absolute percentage error (MAPE), SVM, NB and KNN [1].

A comment like "Great stuff [Laptop Brand Name] – having the battery go flat so quickly is helpful!" an irony review like this will be understandable for humans, similarly simple words such as 'great' and 'helpful' in the comment may indicate positivity but the word 'go flat' may not be selected as a negative term. Some Customers comprehensible enough to get the irony review and know this is a "Negative review" but sentiment analysis reads that it is a "Positive review". These sentiment analyses will help the company to monitor the product and consecutively take the decision on supply and demand of the product.

If numerous comments where bad or negative on social media for a product it will eventually drop the sales, so as consumers examining a product will omit away from viewing the product even though the company popular brand in that area. However, where a product receives a positive comment among the community, which will create a higher demand for products and equally increase the sales in the marketplace [2].

Online reviews create an impact based on factors like Trending products will acquire high reviews which proportionally make the consumer think that it's trustworthy and safe. Hence the reviews of trending products accurately reflect the products quality and its sales [3]. Differentiating significant reviews from the normal reviews aided us to predict the online reviews and sales [4].

To analyze these reviews, we implement NB with Sentiment Analysis (Opinion Mining) lexicons dictionary and we use Bass and Norton model to combine the reviews and measure of customer satisfaction method concludes individual terms that are not the same terms for all other models and NB seems to be less accurate than more complex models namely SVM and KNN. But the NB method helps to define sentiment polarity compared to other models.

SVM models are implemented for both the new and existing products which depict that textual information added and will aid in forming random projections [5].

Outcomes which acquire from evaluating the models are defined in three properties [6],
(i) Accuracy: Overall ratio.
(ii) Sensitivity: Positive Sentiment ratio
(iii) Specificity: Negative Sentiment ratio

III. RESEARCH METHODOLOGIES

A. *Methods used*

K Nearest Neighbor (KNN): -

KNN is an instance-based learning machine-learning algorithm, which chooses the K-nearest neighbors from the training set and proposes to track the volatility of movements of two attributes namely the sentiments and the sales in this instance. It is basically used to test the similarity between 2 attributes which can cause changes. It is a non-parametric method used in classification and regression cases wherein the input consists of K closest training datasets. It is a instance based learning. It is a supervised learning algorithm

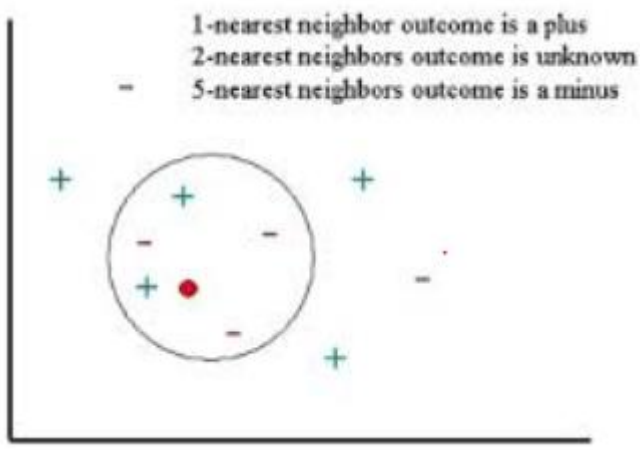


Figure 2 - KNN

Support Vector Machine (SVM): -

It is also a supervised learning algorithm which analyze the data for Classification and Regression analysis. They have the capability to perform linear classifications and non-linear classifications.

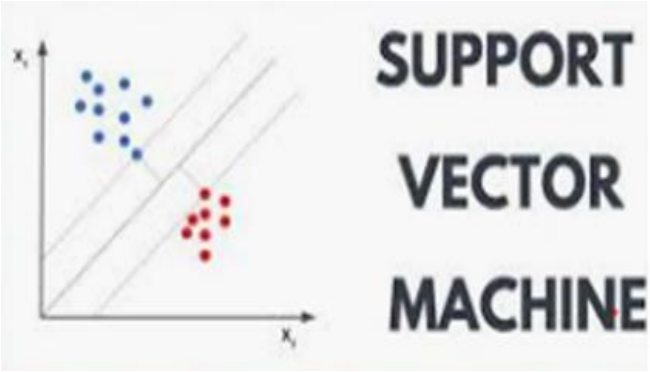


Figure 3- SVM

Linear Regression (LR): -

It is used to explain the relationship between dependent variable and independent variable the case of one explatory variable is defines as Simple linear regression and multiple variables is called as Multiple Linear regression.

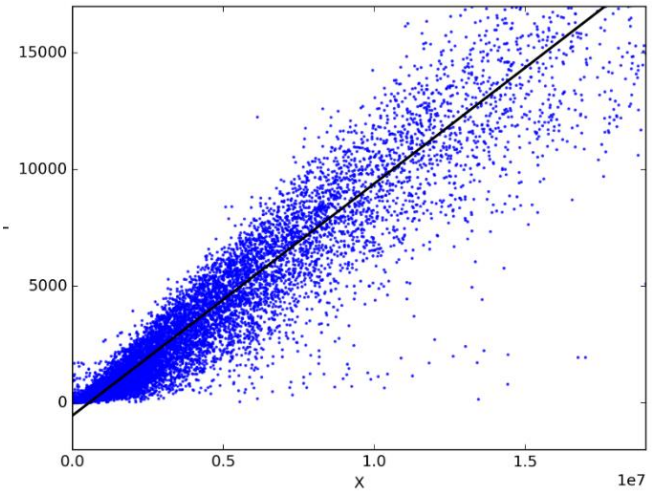


Figure 4 – Linear Regression

.

B. *Datasets*

Motivation: - By adopting this dataset we will be performing sentiment analysis on the customers sentiments on the taste and quality of wines with different characteristics and how are these sentiments affecting the sales of wine.

Wine Dataset: - A predictive model to identify wines through blind wine tasting like a master sommelier would help to give it a review in a scale of 1-100. The initial step in this process is accumulating and assimilating the dataset for training purposes. Post this, a deep learning in wine taste prediction will be derived in order to understand the sentiments of customers and how this sentiment is affecting the sales process.

Variables of dataset:

Country: - This column explain from the wines are derived from which country.

Description: - This column describes the wine's taste texture, smell, look and feel.

Designation: - This column explains the vineyard within the winery from the where the wines are from.

Points: - This column explains the ranking of the wine in a scale of 1-100 by wine tasting enthusiast.

Price: - Cost of the wine bottle post production.

Province: - It describes the province or state from where the wine is from

a) **Region 1: -** This is one of the vineyards from a region numbered as 1 which can primarily be only 1 region.

b) **Region 2: -** In many instances the wine production happens from different regions and more than 1, this column will explain the other region details.

Variety: - Grades and types of grapes that are used for manufacturing wines.

Winery: - The winery from where the wine was extracted.

Data Preparation:

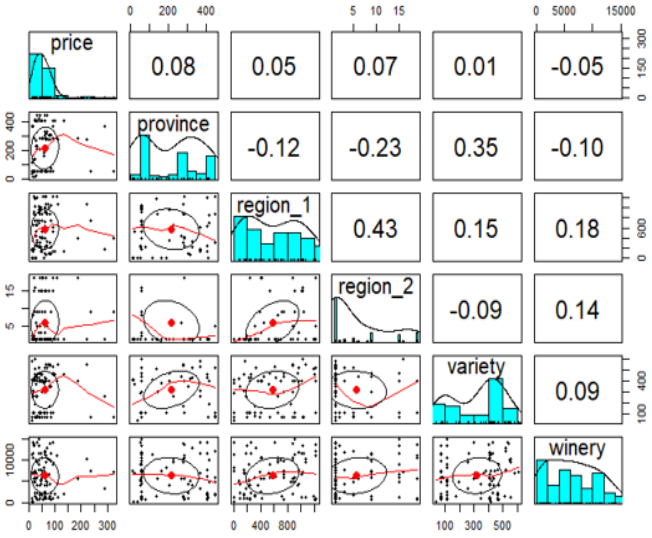
In this dataset we have 129969 rows and 14 columns.

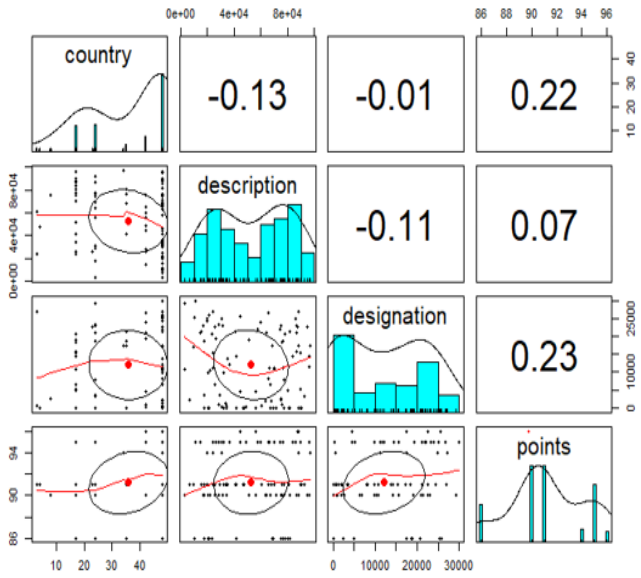
Missing values: - Dataset consists of mere number of missing values so we are omitting missing values from the dataset using the function 'na.omit()'.

Outliers: - We have observed that greater the points, higher are the prices of wine bottles. But we feel that purchasing wine bottles at such high prices (>$40 – greater than average price) is a poor decision (outlier), so we have levelled off those price range of USD 4000 to USD 40 for easy reference.

IV. Results and Evalualtion

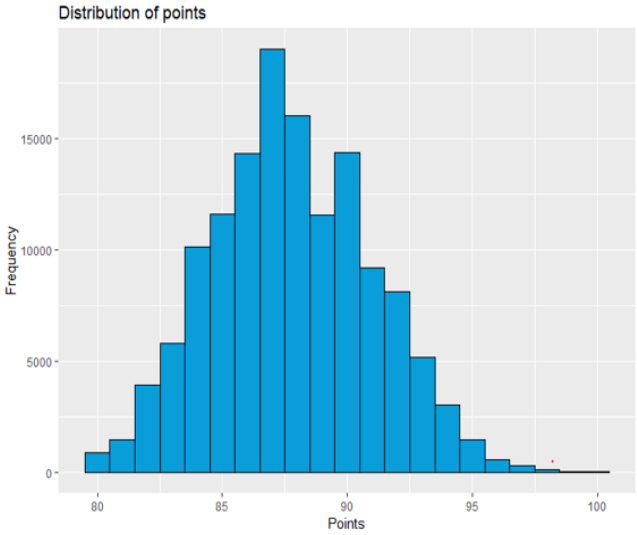
Correlation: -



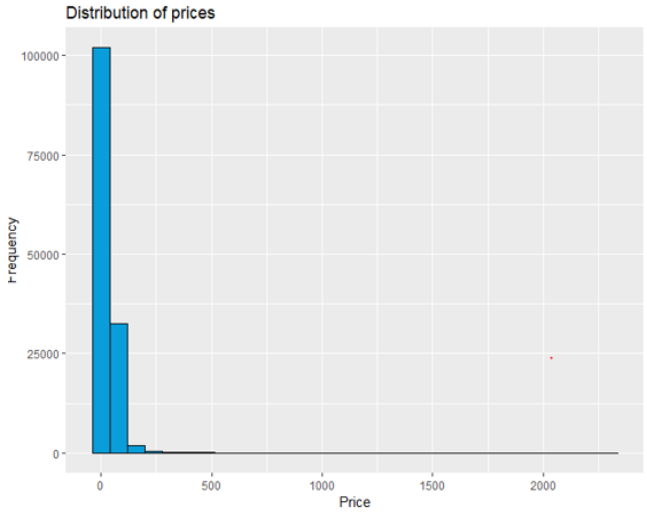


The above pair diagrams seek to explain the co-relation among various variables of the wine dataset. We have identified "points" as to be response variable and rest of the variables are predictive variables. The above chart depicts that the response variable "points" is more correlated to "designation" and "country" predictive variables.

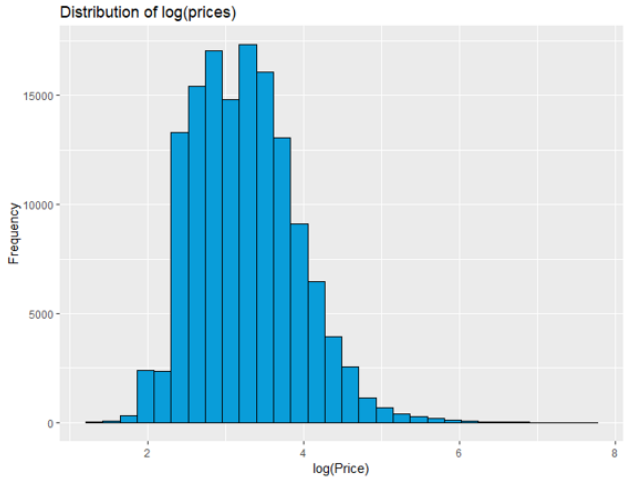
Distribution of Key variables:



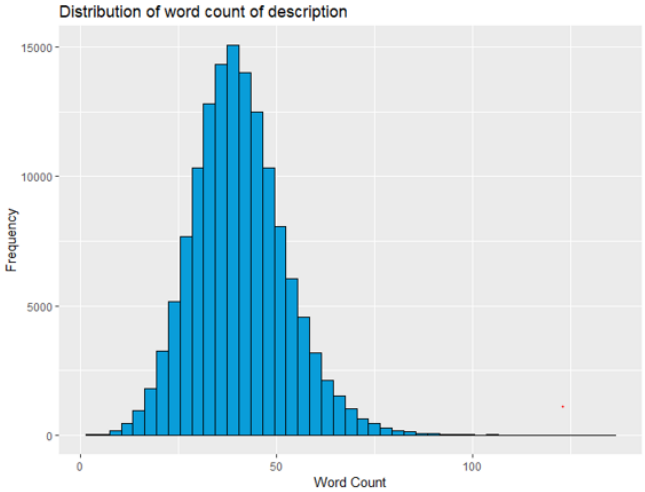
The above bar chart depicts that the response variable "points" is normally distributed. The Y axis denotes frequency of distribution and X axis denotes the number of points with what wine lovers rated on a scale of 1-100. Even though the above chart is found to be normally distributed, the plot looks slightly right skewed as well.



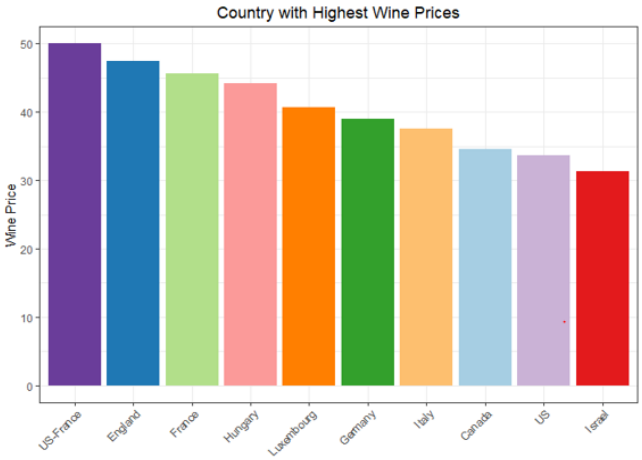
The above bar chart is found to be severely right skewed in which Y axis represents the Frequency and X axis represents the Prices of the wine. Simply, the chart is the distribution of the cost for a bottle of the wine as per the intensity.



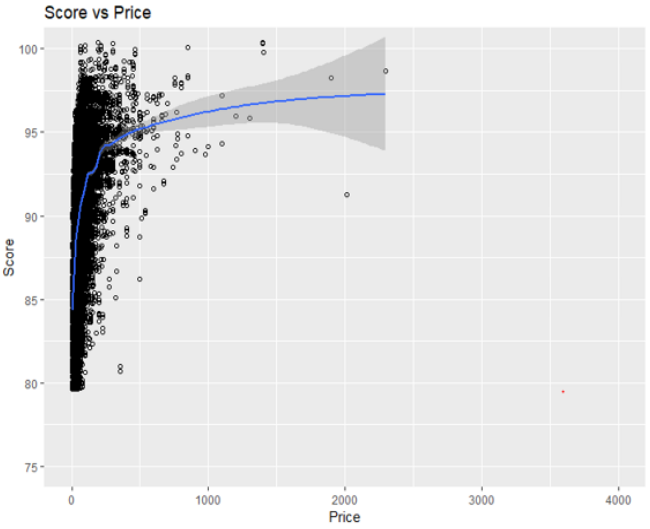
The above bar chart consists of the log(price) which has represented in the X axis and frequency in the Y axis. It is also found to be skewed towards the right slightly.



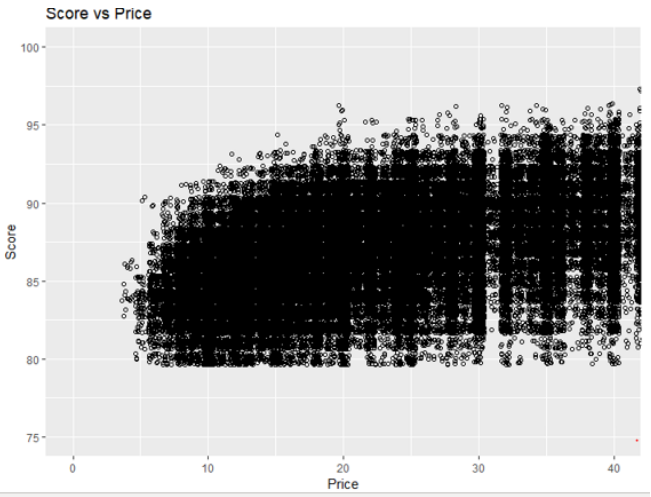
The above bar chart found to be normally distributed with slight skewness towards the right. The Y axis represents the frequency of word count and X axis represents the word count. The above plot represents the word count of each description (tweet) which is nothing, but the review of the wine tasted by customers.



The above graph represents the average price range of wine in countries with most costly wines. It is observed that average prices of wine in jointly produced US-France are the highest when compared to other individual 9 countries. England found to be an individual country with the highest average prices of wines and the average prices of Israel's wine are found to be the lowest.

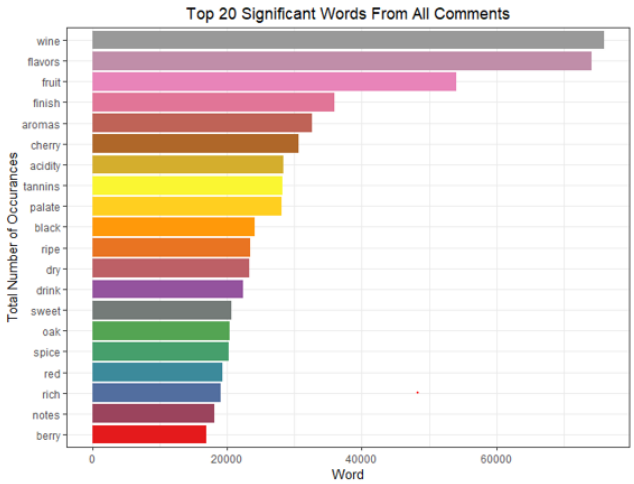


We have observed that greater the points, higher are the prices of wine bottles. But we feel that purchasing wine bottles at such high prices is a poor decision (outlier), so we have levelled off those price range of USD 4000 to USD 40 for easy reference.



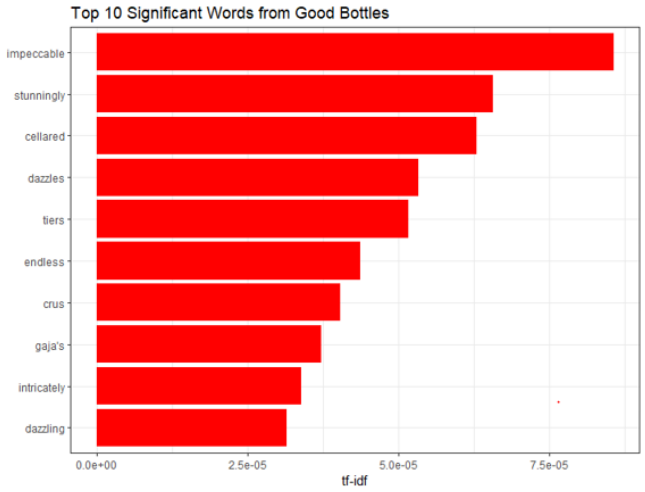
From the previous analysis we have removed the outliers and only 40 observations have been taken for consideration for the analysis. There is a clear indication that there is direct correlation between prices of wines and points. This means that if the points are high, mostly wine prices will be high and vice-versa. Additionally, spending so much of thousands of monies on expensive wine bottles are a bad choice and

buying at second comparable cost of wines with slightly lower points are the best options.

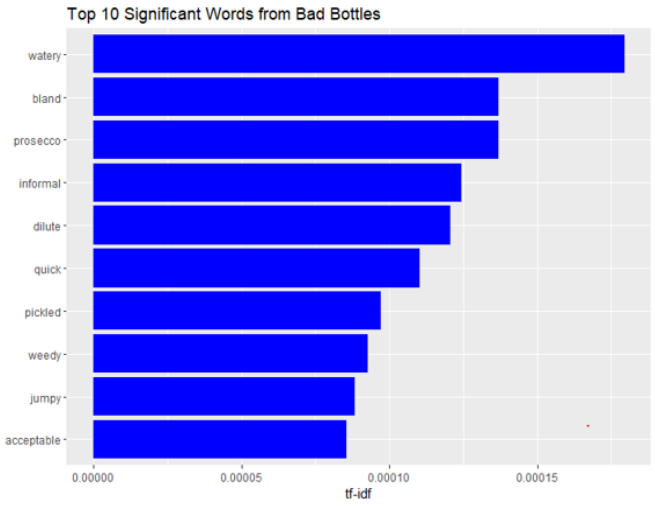


(Code has to be changed to swap the labels of x and y axis).

The above graph represents the significant common words that occur maximum number of times with up to 80,000 being the highest. It is clearly observed that people have given more preferences to types, flavors and presence of fruits followed by presence of berry fruit in wine as the last option. This data is taken from comments by customer for scoring the preferences. This is also called as exploitation of most common words. The most commonly used 20 words to make the selection of tweets more comprehensive in order to do better sentiment analysis.



The above Chart represents the choice of words by people to compliment the quality of good wine. The word ‘Impeccable’ has been used more than 7500 times followed by complimenting words like ‘Stunning’, ‘Cellared’, ‘Dazzle’ etc followed by minimally used word as ‘Dazzling’. The above word has been inferred from the tweet remarks. We have narrowed 10 random words showing how customers are satisfied with the quality of wine on the basis of their tweets.



The above chart represents the most commonly used words for expressing their disappointment regarding the quality of wine. These words are obtained from tweets that are regarded as bad. The word ‘Watery’ has been used the most followed by ‘Bland’ and the least use word was ‘Acceptable’. This is also inferred from the series of tweets that has been expressed to show the unacceptability about the quality of wines. We have chosen 10 most commonly used words which expresses dissatisfaction about the wine quality for doing sentiment analysis.

Results:

MODEL	ERROR VALUE
KNN	2.826455
SVM	2.946409
LINEAR REGRESSION	2.942353

[1] "RMSE FOR KNN model is "

[1] 2.826455

> print("RMSE FOR SVM model is ");RMSE(test$pred_svm,test$points)

[1] "RMSE FOR SVM model is "

[1] 2.946409

> print("RMSE FOR LM model is ");RMSE(test$pred_lm,test$points)

[1] "RMSE FOR LM model is "

[1] 2.942353

RMSE used to measure the prediction by each model we have use and comparatively KNN model performed better.

With the obtained knowledge from this data mining, the preference of the customers can be identified so that we could manufacture different types of wines based on the preference of the people of the country so that the maximum revenue is possible for any manufacturing industries which in turn strengthens the economy.

How data is reproducible?

This paper proposes to record that the datasets that has been applied for this study has been obtained from Kaggle (https://www.kaggle.com/zynicide/wine-reviews/activity).

Though this particular study has been targeted for wine industry for predicting the sales and subsequent wine reproduction using 'Sensitivity Analysis' by applying Root Mean Square Error (RMSE) method, the same kind of analysis with the help of social media can also be extended to Hospitality, Tourism and other leisure industries. With the help of customer feedbacks and tweets, it can act as direct publicity for increasing sales and gain market share.

Advantages of Sensitivity Analysis: -

Increases sales as customers will be assured of quality.

It will act as direct publicity for companies offering services and it can be a form of economical advertisement medium.

Since the feedbacks are real, it can be considered as real time by prospective customers.

Disadvantages of Sensitivity Analysis: -

Fake reviews cannot be restricted, and this may create biases onto reviews.

All customers cannot be expected to give reviews which may impact the output of sensitivity analysis.

There can be instances that where the firms in hospitality and service industry may try corrupt practices in order to get better ratings and reviews ultimately creating biases.

Implications:

This project showcases how sales of the product related to the public opinion. There are various variants of same product can be manufactured by the same manufacturer, (for instance, snacks and beverages with various ingredients). Not all variants are being liked by consumers equally. Different variants can be liked by different set of consumers in different set of regions. So, this study states that manufacturing all the variants in the same quantity is not a wise option. This study helps manufacturer to forecast the sales of the product so that it allows them to make decision on which type of products to be manufactured more, which is to be manufactured less, where to be manufactured more and where to manufactured less. With the help of forecast on the sales, the efficient manufacturing is possible which in turn saves the resources, increases the sales and strengthening the economy of the market.

V. REFERNCES

[1] [1] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends R in Information Retrieval, vol 2, nos 1–2, pp 1–135, 2008.

[2] [2] L. C. Wood, T. Reiners and H. S. Srivistava, "Expanding Sales and Operations Planning using Sentiment Analysis: Demand and Sales Clarity from Social Media", Aut.researchgateway.ac.nz, 2019.

[3] [3] F. Zhu and X. Zhang, "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics", Journal of Marketing, vol. 74, no. 2, pp. 133-148, 2010. Available: 10.1509/jm.74.2.133.

[4] [4] C. Chern, C. Wei, F. Shen, and Y. Fan, "A sales forecasting model for consumer products based on the influence of online word-of-mouth", Information Systems and e-Business Management, vol. 13, no. 3, pp. 445-473, 2014. Available: 10.1007/s10257-014-0265-0.

[5] [5] M. Schneider and S. Gupta, "Forecasting sales of new and existing products using consumer reviews: A random projections approach", International Journal of Forecasting, vol. 32, no. 2, pp. 243-256, 2016. Available: 10.1016/j.ijforecast.2015.08.005.

[6] Z. Fan, Y. Che and Z. Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis", Journal of Business Research, vol. 74, pp. 90-100, 2017. Available: 10.1016/j.jbusres.2017.01.010