Analyzing Taste from Review Word Choices using Natural Language Processing

Fahad Khan Raj
20101250
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
fahad.khan.raj@g.bracu.ac.bd

Radoan Sharkar
20101263
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
radoan.sharkar@g.bracu.ac.bd

Somaya Al Sadia Rahman
20101560
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
somaya.al.sadia.rahman@g.bracu.ac.bd

***Abstract*—There have been several people relying largely on online review platforms to distribute their opinions on foods. Prior research examining the effects of online reviews has primarily concentrated on a single quantitative factor, such as ratings or reviews about foods given by customers. We were able to determine whether or not the consumers liked the food by using a wide variety of algorithms, such as C-Support Vector Classification and Naive Bayes. Examining client feedback helps to discover how individuals feel about foods and their overall responses. We evaluated specific food items to examine client feedback to discover how individuals feel about foods and their overall responses.**

I. Introduction

Food, like all other things in life, has both positive and negative effects. The way in which we consume food can either promote or detract from our overall health and well-being. However, the area that has not gotten much attention is automatically ranking certain menu items in a restaurant based on internet customer feedback. In this paper, we investigate the methods and techniques used to automatically find whether the customers liked or disliked the food using natural language processing. In addition, we have gone through factors that affect the accuracy of our algorithm for food items of a restaurant based on the reviews provided online by users. Finally, we evaluate the accuracy and reliability of the proposed model by comparing its performance with that of other models that were suitable.

II. Methodology

A. Dataset

The dataset we used is from kaggle. The author of the dataset is Mr. Vicky, a data scientist from Chennai, India. The dataset contains, 1000 individual restaurant reviews with human detected good or bad.

B. Data Preprocessing

First of all, we checked if there are any null values with *isnull()* method. After that, we cleared out all the punctuation and kept only the Latin letters using Regular Expressions.

TABLE I
Dataset preview.

Index	Review	Liked
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1

C. Train-Test Split

To train and test the data, we split the data into two parts. For training the models we are going to reserve 80 percent of the data and to test the models, we need to reserve 20 percent of the data.

D. Training Models

1) C-Support Vector Classification: SVC is a supervised learning classification model. SVM first maps all the training data into a vectorized formation in multidimensional planes and then uses an algorithm to propose a hyperplane that maximizes the aggregate distance between the classified vectors in the space. The real capability comes into play in SVM where it can perform both linear and nonlinear classification using a kernel. The default kernel in sklearn sklearn.svm.SVC is 'RBF' (radial basis function) which we are using to train our SVE model. It uses the following formula:

$$K\left(x^i, x^j\right) = \phi\left(x^i\right)^T \phi\left(x^j\right) = e^{\left(\gamma \left\| x^i - x' \right\|^2\right)}$$

Here, $\gamma > 0$

2) K-Nearest Neighbor: When the KNN technique is applied to a dataset, we are capable of making a prediction about the category that a set of points belongs into. This occurs every time a new point is added to the dataset. The initial step which we are likely to have to do in order that will get the process of forecasting underway is to determine how much the optimal solution will be. This is going to be the most important step. If we extrapolate from the data shown in the Figure, we find that green points are associated with Class X, blue points with Class Y, and yellow points

with Class Z. If K is 8, then we choose the eight points that seem to be the furthest from the triangle's new origin. To get the distance that separates the two spots, we must first accomplish this. When K equals eight, as seen in Figure 1, fresh points are considered a threat to one yellow point, three green points, and four blue points respectively. We are able to draw the conclusion that the new point belongs to class Y when K is equal to 8 as a result of the fact that we have a bigger number of additional points than any other color. If we continue with K set to 16, we will have to look for 16 more points that will be closest to the new points. According to the findings of the calculation of distance, new points are found to be closer to three yellow points, five blue points, and eight green points when K is set to 16. These results are derived from the fact that K is multiplied by 16 in the calculation. As a consequence of this, and taking into account the fact that K = 16, we are in a position to assert that the new point belongs to class X. It's possible that we'll need to run a few different K values forward through a cross-validation testing technique before we can establish which one of those K values produces the best results.

3) Logistic Regression: This study opts to employ the logistic regression model, which is not only popular in the field of machine learning but also widely used in real-world industrial environments. For instance, in this study, the heart disease risk factors were discussed, and along with that, a forecast was made regarding the probability that the condition will manifest itself. Logical regression is extensively used for classification, mainly for situations involving two categories, and it is able to estimate the likelihood that each classification event will take place. This is due to the fact that there are only two forms of output, each of which represents a single category. The following is an example of the logistic regression model:

$$\text{prob}(Y = 1) = \frac{e^2}{1 + e^2}$$

Where Y denotes to the binary dependent variable (Y is equivalent to 1 if an event occurs; else, Y=0), e represents the base of natural logarithms, and Z denotes the following:

$$\mathrm{Z} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots\ldots\ldots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p}$$

with constant β_0, coefficients β_j and predictors X_j, for p predictors $(j = 1, 2, 3, \ldots\ldots, p)$

4) Gaussian Naive Bayes: The term "Normal Distribution" refers to another name for the "Gaussian Distribution." A statistical model that illustrates the statistical distributions of continuous random variables found in nature is called the normal distribution. The bell-shaped curve that characterizes the normal distribution is used to define the distribution. The two characteristics of the normal distribution that are considered to be of the utmost significance are the mean () and the standard deviation (). The average value of a distribution is referred to as the mean, while the standard deviation is indeed the "width" of the distribution around the mean. Both terms refer to the same concept. It is essential to be aware that a normally distributed variable (X) is a continuous variable with a distribution ranging from to +, and that the total area underneath the model curve is equal to 1.

It is important to know that a variable (X) that is normally distributed, is distributed continuously (continuous variable) from $-\infty < \mathrm{X} < +\infty$ and the total area under the model curve is 1.

5) Multinomial Naive Bayes: Within the realm of natural language processing, the Bayesian learning strategy that goes by the name of the Multinomial Naive Bayes algorithm sees widespread application (NLP). The computer is able to create an informed estimate on the classification of a text using the Bayes principle. This guess may be used for a text such as an email or a news item. It then provides the outcome of the computation it performed to estimate the probability of each tag being connected with a specific sample. This calculation determines the likelihood of every tag being linked with a particular sample. The Naive Bayes technique is an efficient method for performing text input analysis and finding solutions to problems involving multiple categories. It is necessary to have a firm grasp of the Bayes theorem before moving on to comprehending the Naive Bayes theorem. Thomas Bayes is credited with the development of the Bayes theorem, which calculates the probability of an event taking place based on the circumstances that surround it. We make a determination regarding the likelihood of class A assigned predictor B. The following equation serves as its foundation. It is based on the following formula:
It's based on the formula below:

$$P(A \mid B) = P(A) * P(B \mid A)/P(B)$$

Calculating probabilities is all that is required to accomplish this objective, thus getting there shouldn't be difficult. This technique may be used to either constant or not constant data with equal effectiveness. This simple method makes it possible to anticipate the use of real-time apps. It is readily scalable and able to cope with extremely large datasets without any problems. When compared to other probability algorithms, this method of creating predictions is not as exact as the others. In this particular instance, regression analysis is not applicable. The Naive Bayes algorithm could also be used to classify the information that is provided in the form of text; it cannot be used to estimate numerical values.

III. Results

A. Overall Accuracy Score

1) C-Support Vector Classification: Overall accuracy for SVC: 73.5 percent

2) K-Nearest Neighbor: Overall accuracy for KNN: 58.5 percent

3) Logistic Regression: Overall accuracy for Logistic Regression: 71 percent

4) Gaussian Naive Bayes: Overall accuracy for GaussianNB: 73 percent

5) Multinomial Naive Bayes: Overall accuracy for MultinomialNB: 76.5 percent

B. Classification Report

Macro Average:

Algorithm	Accuracy	Precision	Recall	F1 Score
SVC	0.735	0.78	0.74	0.73
K-Nearest Neighbor	0.585	0.60	0.59	0.58
Logistic Regression	0.71	0.71	0.71	0.71
GaussianNB	0.73	0.75	0.73	0.72
MultinomialNB	0.765	0.77	0.76	0.76

IV. Conclusion

After all the training and testing we can conclude that Multinominal Naive Bayes rendered the best overall accuracy score of 76.5 percent and on the other hand K-Nearest Neighbor performed an overall accuracy score of 58.5 percent which is the worst in this case.