RESTAURANT REVIEWS: PREDICTING CUTOMER REVIEW SENTIMENT AND ITS AUTHENTICITY.

Tajudeen Abdulazeez
IST736 Text Mining
School of Information, Syracuse University
toabdula@syr.edu

INTRODUCTION

Oh yes, there's lots of great food in America. But the fast food is about as destructive and evil as it gets. It celebrates a mentality of sloth, convenience, and a cheerful embrace of food we know is hurting us. Anthony Bourdain

According to an article published on USTODAY website, more than 1 of 3 American eat outside, Findings also showed fast food consumption decreased with age. Forty-five percent of adults ages 20 to 39 ate fast food, while only 24 percent of adults over 60 had fast food as a meal or snack. Men tend to eat more fast food than women, said the survey, and non-Hispanic black adults consumed more fast food (42 percent) compared to other races.
A person's income also played a key role in how often they ate fast food. People with higher incomes were more likely to consume fast food than those at lower incomes, the survey found.
Now let's get some details about customer reviews and how it can affect business, can we consider all customer review true and accurate? Below is an example of customer reviews when search on google. At a glance, users can easily see the aggregated star ratting which can have significant impact in shaping the decision of customers.

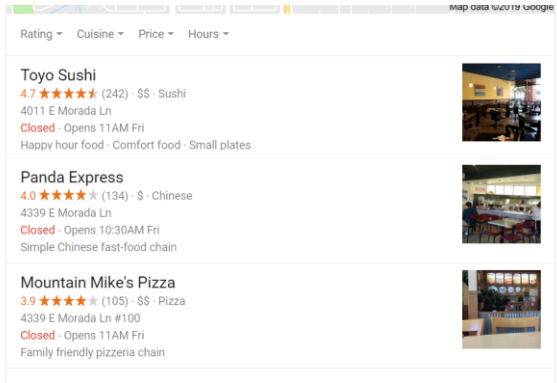

Figure 1: Customer Reviews on google.

Reviews gives customers greater autonomy in their decision-making. They can now easily find information, not only about fine dining venues but also casual dining, ethnic and even fast food restaurants.
In the future, the importance of reviews is unlikely to decline. However, the fragmentation of information sources and the exponential increase in the number of reviews will pose challenges to both restaurateurs and customers.

Our aims and objectives for this project is to see how a machine learning model can be used to predict customer review sentiment and its authenticity using Bernoulli and multinomial naïve Bayes model.

ANALYSIS AND MODELING

This analysis is carried out using python programming language with the following packages

- Pandas: tabular data manipulation
- Numpy: numerical computing
- Matplotlib: visualization
- Wordcloud: word cloud generation
- Nltk: text processing
- Sklearn: modeling
- Seaborn: Visualization

The dataset used for this analysis is from the learning management system of the Syracuse university.it has three columns; lie, sentiment and reviews. Lie and Sentiment are the target labels derived from the customer reviews.
Below is the word cloud generated from the false customer reviews.

Figure 2: Word cloud generated from False Reviews

The most common words from the word clouds are: Food, Restaurant, Services, Price, experience, best, minutes e.tc. as observed from the word cloud.

Let's have a view of the word cloud of true customer reviews.

Figure 3: Word cloud generated from True Review

Observation: from the word cloud of both customer reviews that is consider true and customer reviews that is considered lie, according to the data, the most occurrence word happened to follow a similar pattern. The word, food, restaurant, service, great, ordered, best, friend, occurred frequently in both false reviews and true reviews.
The question now is: Can we used the occurrence of this words to draw a boundary between true and fake reviews?

Let's review the sentiment class of the customer reviews.
Below is the word cloud generated from the positive sentiment of the restaurant reviews.

Figure 4: Word cloud generated from Positive Sentiment.

The word generated from negative sentiment is shown below:

Figure 5: Word cloud generated from Negative Sentiment.

Observation: from the word cloud of reviews sentiments, words like, amazing, best, delicious which are positive words. In negative sentiment reviews, the words like, bad, never, terrible, never, minutes, and time appeared most frequently which might be assumed probably high customer wait time. We need to put this into consideration and careful observation when creating our models, to see if our models is learning the underlying patterns.

Data Distribution

The datasets have two labels, lie and sentiments. The class labels are equally balanced for both sentiment class and lie class as shown in the histogram below:

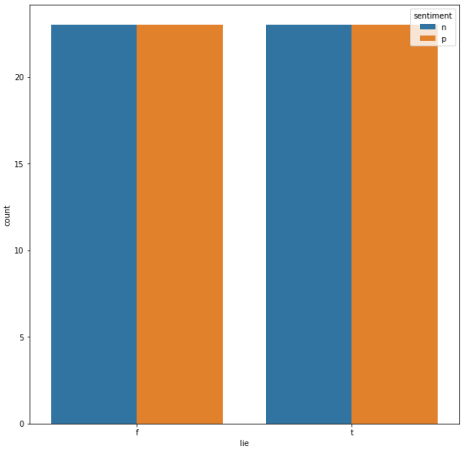

Figure 6: The distribution of the label's variable in the datasets.

Accuracy is used for model evaluation with a baseline of 50%.

Data Preprocessing and Transformation

Before we talk about how the reviews are tokenized or transforms, let's take a look at the first two customer reviews from the raw datasets

'Mike\\'s Pizza High Point NY Service was very slow and the quality was low. You would think they would know at least how to make good pizza not. Stick to pre-made dishes like stuffed pasta or a salad. You should consider dining else where.'

'i really like this buffet restaurant in Marshall street. they have a lot of selection of american japanese and chinese dishes. we also got a free drink and free refill. there are also different kinds of dessert. the staff is very friendly. it is also quite cheap compared with the other restaurant in syracuse area. i will definitely coming back here.'

Special character like '\\', ',', 's','.''\\n' are removed and the resulting clean reviews before tokenization of the first two reviews in shown below:

Mikes Pizza High Point NY Service was very slow and the quality was low You would think they would know at least how to make good pizza not Stick to pre-made dishes like stuffed pasta or a salad You should consider dining else where

i really like this buffet restaurant in Marshall street they have a lot of selection of american japanese and chinese dishes we also got a free drink and free refill there are also different kinds of dessert the staff is very friendly it is also quite cheap compared with the other restaurant in syracuse area i will definitely coming back here

The reviews look a little cleaner, although, we still have stop words. This review is tokenized using Count Vectorizer and Tf*idf Vectorizer.

Let's get little explanation of what vectorization is:

Vectorization is the process of transforming text document in a way computer can understand. The only way a computer can understand text is to represent with number which is call text vectorization. This is an important step in data mining. Each word represented by a number is called a token.

Tokenization rules can vary depending on the business problems and how grouping of words can have significant impact in identifying trend in text document. The tokens can be represented with n-grams or bag of words. Bag of word tokens ignore the context of words while n-grams while n-grams capture the local content of words, the representation can be unigrams, token of individual words, bi-grams, token of two words, tri-grams, token of three words and so on.

There are four major different type of converting text to a vector for computer to understand.

1. Boolean Vectorizer
2. Count or Term Frequency Vectorizer
3. Normalized term Frequency
4. TF*IDF Vectorizer

Boolean is a vectorization method where each token is represented, either if it is present in a document or not. If a token is present, it is represented by 1 and 0 otherwise.

Term Frequency is a vectorization method where each token is represented by the number of occurrences in a document. If you a document and a token appeared in the document 5 times, the token or word will be represented by 5.

Normalize term frequency is a method of vectorization where word frequency is normalized by the document length. If you have a document with a length of 1000, and a word appeared in the document for 20 times, the normalized term frequency will be 20/1000.

Tf*idf is a concept borrowed from information retriever, and it's a blind weighting strategy for text classification. in this method, the term frequency is multiplying by the inverse document frequency. This is aimed to penalize the common word across documents.

Training and Test set

The datasets are divided into train and test sets, with 70/30 propulsion using sklearn train-test-split.

The train set is used for model training while the test set is used to evaluate the model performance.

Modeling and Evaluation

Bernoulli and Multinomial Naïve Bayes is used for modeling. Multinomial naïve Bayes takes count vectorizer as input while Bernoulli Naïve Bayes takes Boolean vectors as input.

Below is the brief explanation of the general concepts of Naïve Bayes.

Naïve Bayes:

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

i. The probability of Y given X is equal to the probability of (X,Y) divide by the probability of X.
ii. The probality of X given Y is equal to probability of (X,Y) divide by the probability of Y

Both are combined to a single equation as shown above, which is the equation the Naïve Bayes uses.

The bayes theorem generally assumed that attributes are independent which make it very suitable for text classification.

Naïve Bayes Classifier

If one of the conditional probability is zero, then the entire expression becomes zero

Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic}+1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes
p: prior probability
m: parameter

Model Evaluation

Accuracy is used to evaluate the model with a baseline accuracy of 50%.
Below is the brief description of metrics for evaluating model using confusion metrix.

Metrics for Performance Evaluation: Confusion Matrix

Confusion Matrix:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	True Positive	False Negative
	Class=No	False Positive	True Negative

The accuracy is calculated as shown in the equation above. The sum of the True prediction (TP,TN) divided by the total number of classes.

Results

Lie Detection

The result for 10-fold cross validation is shown in the table below for both Bernoulli and Multinomial Naïve Bayes Using different Vectorizer.

	Model	Vectorization	10 fold Avg Score
0	BNB	bool	0.5700
1	BNB	count	0.5700
2	BNB	ngram	0.5600
3	BNB	tfidf	0.5700
4	MNB	bool	0.5300
5	MNB	count	0.5600
6	MNB	ngram	0.5600
7	MNB	tfidf	0.5375

Table 1: Lie Classifier results

Bernoulli model did very well in classifying customer reviews authenticity with an accuracy of 57%.

Below are the top 20 feature importance Bernoulli model used as predictor using a boolean vectorizer.

cold: 1.8177349556313156

want: 1.8177349556313156

said: 1.2787384548986285

delicious: 1.1245877750713702

staff: 1.1245877750713702

dine: 0.9422662182774157

high: 0.9422662182774157

meal: 0.9422662182774157

sauce: 0.9422662182774157

waiters: 0.9422662182774157

worth: 0.9422662182774157

bar: 0.8903152454708945

environment: 0.8903152454708945

definitely: 0.8732733467904639

people: 0.8213223739839426

dont: 0.7191226669632056

order: 0.7191226669632056

prices: 0.7191226669632056

table: 0.7191226669632056

great: 0.6137621513053797

Sentiment Detection

The result for 10-fold cross validation is shown both Bernoulli and Multinomial Naïve Bayes Using different Vectorizer is shown in the table below:

	Model	Vectorization	10 fold Avg Score
0	BNB	bool	0.7825
1	BNB	count	0.7825
2	BNB	ngram	0.7825
3	BNB	tfidf	0.7825
4	MNB	bool	0.7675
5	MNB	count	0.8100
6	MNB	ngram	0.8100
7	MNB	tfidf	0.8000

Table 2: Sentiment Classifier results.

Multinomial Naïve Bayes did very well in classifying the customer reviews sentiment with an accuracy of 81%.

Below is the to 20 feature importance the model used as a predictor using ngram vectorizer.

amazing: 2.705095989204778

terrible: 2.2577486410551293

asked: 2.1624384612508045

took: 2.0570779455929777

best: 1.9727280954915511

came: 1.939294909936594

said: 1.939294909936594

fresh: 1.8448947239816667

great: 1.5870656146795667

friendly: 1.5264409928631322

nice: 1.5264409928631322

prices: 1.5264409928631322

minutes: 1.400298409203907

delicious: 1.3929096002386094

need: 1.3929096002386094

bad: 1.3639307650330323

wasnt: 1.3639307650330323

sauce: 1.1209758847549676

ask: 1.0564373636173965

environment: 1.0564373636173965

Looked at the feature importance, we can actually see that our predictor has learn something in identifying the customer sentiment.

Conclusion

We can make the following conclusion from the experimental results and analysis:

- Bernoulli Naïve Bayes perform well in predicting the authenticity of customer reviews using bool vectors as an input with an accuracy of 57%. The model exceeds the baseline accuracy of 50%.
- Multinomial Naïve Bayes perform better in predicting customer reviews sentiment using count vectors as input with an accuracy of 81%. This exceeds the baseline accuracy of 50%.
- Sentiment classification attained more accuracy using machine learning classifier compared to authenticity classification.

[Source Code]:
https://colab.research.google.com/drive/1O9lRTc1GwMvWs4_bSHAKjpS-eG7tesRZ

[Code on Github]:
https://github.com/toraaglobal/CustomerReview

Ref:
https://www.usatoday.com/story/news/nation-now/2018/10/03/americans-eat-fast-food-daily-cdc-survey/1507702002/