Customer Sentiment Analysis: Applied Natural Language Processing Approach

Baibhav Datta
Department of Computer Science
City, University of London

***Abstract*—Sentiment Analysis also known as Opinion Mining refers to the use of natural language processing, text analysis to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. In this project, we aim to perform Sentiment Analysis of product based reviews. Data used in this project are reviews collected from McDonald's stores. We expect to do review-level categorisation of review data with promising outcomes.**

I. INTRODUCTION

The development of computational models of aspects of human language processing is called as Natural Language Processing. The purpose of NLP is to analyse, extract, and present information for better decision-making in businesses. Researchers would use ML models with NLP techniques to process and classify datasets filled with various reviews in their study. Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. Opinions about a product, service or idea are determined and organised using this sentiment analysis. Sentiment analysis is used as it has to deal with opinion, reality, and decisions that has to be taken. Using advanced text mining techniques, we classify the sentiment of the text in the form of positive, negative and neutral. Challenges in word ambiguity, multi-polarization, detection of negation and sarcasm are to be taken into account while building a sentiment classifier.

It is not uncommon for people to seek out the opinions of others when making decisions. According to survey, "what other people think" has more impact to a buyer's decision while purchasing products or services. So, for a successful business providing products or services customer satisfaction plays a major role. An analysis of customer opinions is important to understand what the customer wants in terms of sentiment, and also the betterment of a company to grow overtime. So, sentiment analysis is used in this study to look at the ideas of the customers written in their review for different McDonald's stores.

II. ANALYTICAL QUESTIONS

- What do the sentiments of the majority of customers' reviews reflect?
- Which store is the most popular with the highest number of positive ratings and lowest number of negative ratings?
- Which store is the least popular with the lowest number of positive ratings and highest number of negative ratings?
- The dishonest relationship between reviews and ratings has already become a significant issue. In some cases, there is a pointless meaning in a product review and its rating. For instance, although a product has a good review, a low rating is given to it, or the opposite situation of this. When purchasing a product, looking at its rating first can be easier than reading its reviews. For this reason, a rating should have a similar meaning to a review for a fair shopping experience.

Therefore, the main objective of this research is to identify the effects of unfair ratings by using sentiment analysis and indicate statistically. Then, to reduce his inanity, creating a machine-learning model that analyzes customers' comments as positive or negative. If a business can use a similar model, then customers can be informed that their reviews for a product are not meaningful with the rating. Thus, customers can get a fairer shopping experience by accessing accurate product information.

III. DATA

The dataset selected for this project is- McDonald's Store Reviews (https://www.kaggle.com/datasets/nelgiriyewithana/mcdonalds-store-reviews)

This dataset contains over 33,000 anonymised reviews of McDonald's stores in the United States, scraped from Google reviews. It provides valuable insights into customer experiences and opinions about various McDonald's locations across the country. The dataset includes information such as store names, categories, addresses, geographic coordinates, review ratings, review texts, and timestamps. In addition to the text-based reviews, there is a rating score of 1– 5. A rating score of 4 or 5 is considered to be positive, 3 is considered as neutral and and 2 or 1 is considered to be negative.

The dataset contains reviews and their corresponding customer ratings, which will allow us to compare the customer ratings to the ratings predicted by machine learning models based on the textual review. Thus, the dataset contains sufficient information to answer the analytical questions.

IV. DATA ANALYSIS

A. *Data Preprocessing*

Data Preprocessing is a crucial phase in text data analysis. Due to repetitions and redundancies in customer reviews, text data become more complicated. This work implemented various tasks to achieve the data in the desired format.

- rating- The rating column has string values like "1 star" or "5 star". So it has to be converted to numeric form, which discards the irrelevant part of it, i.e. 'star' and just retain the numeric value. So we'll create another column, namely 'rating_numeric', containing the numeric value of the rating.

• sentiment- A new variable namely 'sentiment' is created which depicts the sentiment inferred from the customer's ratings. We are setting the sentiment metrics as follows-

1: positive for rating_numeric = 4 & 5

-1: negative for rating_numeric = 1 & 2

0: neutral for rating_numeric = 3

• store_address- The feature 'store_address' has 40 different unique values, which are essentially 40 unique McDonald's store locations. Since these values are strings, and are considered to be unstructured data, so we will use Scikit-learn preprocessing function called LabelEncoder() to encode these textual data into numbers, i.e. each unique store_address will be mapped to an unique integer.

• For preprocessing the main focus of our dataset, that is the customer reviews, we'll go through various steps as follows-

1) Tokenization of the Text

Tokenization is a technique for dividing text streams into phrases or tiny chunks of textual material. Tokens are fragmented pieces of text This technique aims to make complex textual contents straightforward to solve.

2) Removal of Stop Words

Many words in text files appear repeatedly. As a result, it is critical to delete the stop words. Indeed, stop words never provide significance to the written substance. These kinds of words often appear in large numbers in the text.

3) Convert the Text to Lowercase

In the customer reviews, consumers enter material without following grammar norms, in that the entered text contains both lower and upper case characters. As a result, the classifier has difficulty determining the polarity of the provided text. Such an issue may be avoided simply by changing the entire text to a standard format.

4) Remove Punctuations, Numbers

Punctuations, Numbers don't help much in processing the given text, if included, they will just increase the size of a bag of words that we will create as the last step and decrease the efficiency of an algorithm.

5) Stemming

Stemming is a technique used to reduce an inflected word down to its word stem. For example, the words "programming," "programmer," and "programs" can all be reduced down to the common word stem "program." In other words, "program" can be used as a synonym for the prior three inflection words.

6) Making the bag of words via sparse matrix

Bag-of-words(BoW) is a statistical language model used to analyse text and documents based on word count. The model does not account for word order within a document. BoW can be implemented as a Python dictionary with each key set to a word and each value set to the number of times that word appears in a text.

- Take all the different words of reviews in the dataset without repeating of words.
- One column for each word, therefore there is going to be many columns.
- Rows are reviews
- If a word is there in the row of a dataset of reviews, then the count of the word will be there in the row of a bag of words under the column of the word.

Data preparation was concluded with a final set of variables according to the Bag of Words model. The data was then partitioned into testing and training sets that were used to train and evaluate all models. The data were split 75:25 between testing and training sets: 75% for training and 25% for testing.

*B. **Methodology***

The initial data analysis phase of our research applied exclusively graphical visualisation and analysis. This allowed for the clear rendering of viable relationships between factors across the dataset.

Four machine learning models are then applied to the prepared and transformed McDonald's Reviews dataset. The machine learning models used for comparing results of accurately predicting the rating from the textual reviews are as follows-

• Random Forest Classification

Random forest is a supervised learning algorithm which is used for both classification as well as regression. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

• Multinomial Logistic Regression

The multinomial logistic regression algorithm is an extension to the logistic regression model that involves changing the loss function to cross-entropy loss and predict probability distribution to a multinomial probability distribution to natively support multi-class classification problems.

The LogisticRegression class can be configured for multinomial logistic regression by setting the "multi_class" argument to "multinomial" and the "solver" argument to a solver that supports multinomial logistic regression, such as "lbfgs".

• Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

• K-Nearest Neighbour

The KNN algorithm classifies or predicts a new data point by comparing it to the labeled data points in the training dataset. The algorithm measures the distance between the new data point and its K nearest neighbors in the feature space. The K nearest neighbors are determined based on a distance metric, commonly the Euclidean distance. Once the neighbours are identified, the algorithm assigns the majority class label for classification or calculates the average value for regression.

V. RESULTS AND DISCUSSION

• From the McDonald's Customer reviews dataset, it was observed that a large proportion of observations occurred to be a positive review from the customers. The pie chart below depicts the proportions for the respective customer sentiments, namely 1:positive; 0:neutral; -1:negative.

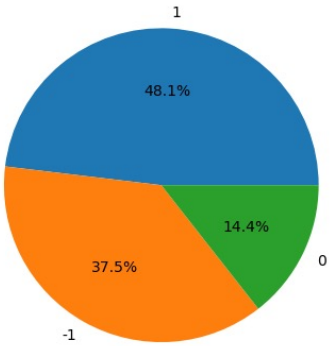


Fig. 1- Pie Chart depicting proportions of unique sentiments

To find the most & least popular McDonald's stores, the dataset was grouped by the store address and the average sentiment was calculated. Since the values of sentiment are among 1,0 & -1, the average value of all sentiments corresponding to a particular store address gives us a score that would clearly depict the store's fame or infamy. Following this method, it was found :

The store which has the maximum positive reviews, with highest number of positive ratings and lowest number of negative ratings is-

429 7th Ave, New York, NY 10001, United States

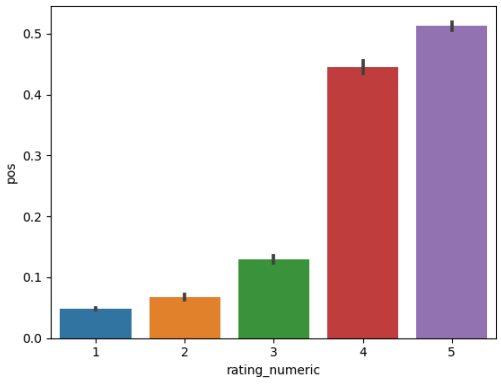
The store which has the maximum positive reviews, with highest number of positive ratings and lowest number of negative ratings is-

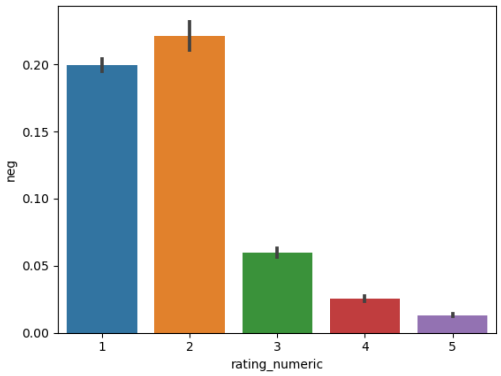
151 West 34th Street (Macy's 7th Floor, New York, NY 10001, United States

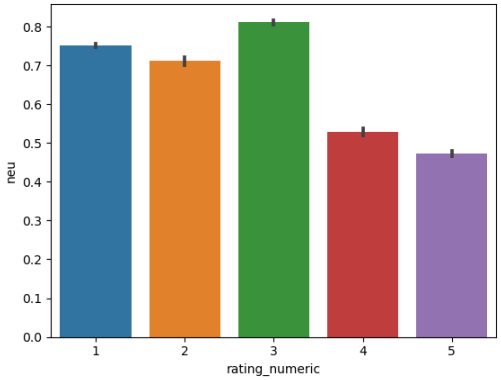
• The Natural language Toolkit (nltk) platform is for building Python programs to work with human language data. A certain function in this platform can be used to generate a sentiment intensity score for sentences. It is the nltk SentimentIntensityAnalyzer, which has a function polarity_scores which takes a sentence as input and returns a float for sentiment strength based on the input text. Positive values are positive valence, negative values are negative valence.

Using this function on the McDonald's reviews dataset, sentiment intensity values are generated, putting the values of polarity scores in three columns, namely 'pos', 'neg', 'neu' depicting the scores for positivity, negativity, neutrality of the sentences respectively.

The bar-graphs shown below signifies how accurately the nltk sentiment analyzer has generated the polarity scores. It shows the ratings along x-axis, and the polarity_scores along the y-axis.







VALIDATION OF RESULTS

The evaluation of the classification models we have used on the dataset will be done with Scikit-learn's accuracy_score function. The accuracy_score function computes the accuracy, either the fraction (default) or the count (normalize=False) of correct predictions.

In multi-label classification, the function returns the subset accuracy. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0.

If ^yi is the predicted value of the i-th sample and yi is the corresponding true value, then the fraction of correct predictions over n samples is defined as-

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(y_i = y_i)$$

Hence, the accuracy scores for predicted sentiments vs true sentiments, after applying the Bag of Words model to the data, for different classification models are as follows-

- Random Forest Classification

Accuracy Score: 83.10%

- Logistic Regression

Accuracy Score: 80.73%

- Support Vector Classifier

Accuracy Score: 81.00%

- K-Nearest Neighbour Classifier

Accuracy Score: 73.59 %

The classification reports for the 4 classification models I have used, which displays the precision, recall, F1 score, and support scores are shown below-

Random Forest Classifier

	precision	recall	f1-score	support
-1	0.83	0.90	0.86	3127
0	0.78	0.42	0.54	1158
1	0.85	0.91	0.88	4064
accuracy			0.84	8349
macro avg	0.82	0.74	0.76	8349
weighted avg	0.83	0.84	0.83	8349

Logistic Regression

	precision	recall	f1-score	support
-1	0.84	0.84	0.84	3127
0	0.63	0.35	0.45	1158
1	0.82	0.92	0.86	4064
accuracy			0.81	8349
macro avg	0.76	0.70	0.72	8349
weighted avg	0.80	0.81	0.80	8349

Support Vector Machine

	precision	recall	f1-score	support
-1	0.83	0.85	0.84	3127
0	0.63	0.35	0.45	1158
1	0.82	0.92	0.87	4064
accuracy			0.81	8349
macro avg	0.76	0.70	0.72	8349
weighted avg	0.80	0.81	0.80	8349

KNN Classifier

	precision	recall	f1-score	support
-1	0.81	0.70	0.75	3127
0	0.48	0.37	0.42	1158
1	0.75	0.88	0.81	4064
accuracy			0.74	8349
macro avg	0.68	0.65	0.66	8349
weighted avg	0.74	0.74	0.73	8349

On comparing the four classification models, the Random Forest Classifier predicted the sentiment from the customer reviews most accurately. Whereas the KNN Classifier has the lowest accuracy score, which means many of the predicted sentiments by the KNN Classifier model are incorrect. Also, it is worth noting that Logistic Regression and Support Vector Classifier also predicted the sentiments quite accurately, and has accuracy scores almost near to the best model.

Also, the high accuracy score of the model predicting the customer sentiment confirms an honest relationship between the customer's reviews and corresponding ratings.

WORD COUNTS

Section	Word Count
Abstract	94
Introduction	255
Analytical Questions	213
Data	146
Analysis	996
Results and Discussion	598
Total	**2250**

REFERENCES

1. Xing Fang & Justin Zhan on "Sentiment Analysis using product review data", 2015
2. Arwa A. Al Shamsi, Reem Bayari, Said A. Salloum on "Sentiment Analysis in English Texts", 2020
3. Ibrahim Lazrig, Sean L. Humpherys on "Using Machine Learning Sentiment Analysis to Evaluate Learning Impact"
4. Jiawei Yao 2019 on "Automated Sentiment Analysis of Text Data with NLTK"
5. C.J Hutto & Eric Gilbert on "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text"
6. Mayur Wankhade, Annavarapu Chandra Sekhara Rao & Chaitanya Kulkarni on "A survey on sentiment Analysis methods, applications and challenged" 2022.
7. Aman Yadav & Abhishek Vichare on "Natural Language Processing Through Transfer Learning: A Case Study on Sentiment Analysis", 2023
8. Wei Yen Chong, Bhawani Selvaretnam & Lay-Ki Soon on "Natural Language Processing for Sentiment Analysis: An Exploratory Analysis on Tweets"
9. Medisetty Madhuri, International Journal of Research Publications and Reviews on Sentiment Analysis on Customer Reviews