Sentiment Analysis on Amazon Fine Food Reviews

Asmita Singh, Amrita Jose, Prateek Chitpur

Dr. Liao

Natural Language Processing – AIT 590 – 001

George Mason University

December 3, 2020

Abstract

Customer Experience is an essential metric for any industry to gauge its performance in the market. Customer perception can be measured from the feedback and reviews obtained on various online or social media platforms. Amazon Fine Food Reviews consists of customer reviews for products in multiple categories. These reviews serve as an indicator of product quality for customers who purchase the products online. Amazon can obtain meaningful insights from its fine food reviews, which have real customers' opinions about their commodities and services. However, manually analyzing such reviews is an expensive and challenging task in huge datasets. This project implements a machine that can capture and learn the user reviews' sentiments and predict the Sentiment of an unseen customer review. The system interprets the language complexities such as context and Sentiment in reviews and would help customers make informed decisions regarding the purchase of products. Three machine learning models – Logistic Regression, Naïve Bayes, and K Nearest Neighbor are applied to the user reviews to compare the best performing model for sentiment prediction. Two vectorization techniques – Bag of Words and TF-IDF are used to determine word similarities and semantics in the prediction. It is observed that Logistic Regression has the best classification accuracy while K Nearest Neighbor performed relatively lesser in prediction. TF-IDF word embeddings gave a higher accuracy for sentiment analysis due to its consideration for the high frequency or rarity of words. Rare words hold more weight as they could determine if a review is positive or negative, while frequently occurring words are given less importance. Summarization of reviews was done using NLTK and spaCy to reduce the reading time for customers. It was observed that spaCy took lesser computation time in generating summaries of the reviews.

Introduction

Amazon.com, Inc. is an American multinational technology company based in Seattle, Washington, which focuses on e-commerce, cloud computing, digital streaming, and artificial intelligence. It is considered one of the Big Four technology companies, along with Google, Apple, and Facebook. The company started as an online marketplace for books but expanded to sell electronics, software, video games, apparel, furniture, food, toys, and jewelry [1]. Due to their influential presence in the e-commerce platform, there are significant risks that their review system can be misused by the seller or the customer (giving fake reviews). Product reviews are one of the users generated content, which is considered a goldmine of consumer sentiment. It allows businesses like Amazon.com an unparalleled acumen in real customers' opinions about their commodities and services. Examining the reviews manually would be challenging and expensive on massive datasets. It needs a human intervention factor in the study to help interpret language complexities such as context, ambiguity, sarcasm, or irony. However, this method will have high accuracy, but it is expensive and time-consuming. Analyzing the reviews of large datasets and assigning sentiments within a stipulated timeframe makes this problem technically challenging and interesting.

Sentiment analysis using NLP is one of the critical solutions to solve the problem stated above. It is cost-effective and provides a scalable solution to analyze data coming from various channels. Sentiment analysis would help Amazon in making decisions faster and more informed than ever before. Our project aims to perform sentiment analysis on the Amazon Fine Food Reviews dataset hosted on Kaggle to solve the above-stated problem. The objective of doing sentiment analysis is to ascertain the customer's attitude based on their language concerning a product. The Sentiment could be positive, negative, or neutral. The dataset consists of reviews of more than ten years. It contains more than 500K records with many upvotes & total votes to the comments. The reviews include product, user information, rating, time, summary, and direct text review. The dataset consists of other category reviews as well [2].

Our project focuses on performing sentiment analysis to mine customers' opinions on products and determine the polarity of the reviews. While this is the project's main objective, we are extending the scope of the project by including additional features. One feature is to provide a summary of the user review (restricted to less than 20 words) intended for customers who do not wish to spend time reading the full review of the product. Also, as part of analyzing the Sentiment of the reviews, this project aims to predict or determine the Sentiment of the review based on the rating and review of the product to provide accurate and meaningful insights about the product the customers can make informed choices.

Source of the dataset - https://www.kaggle.com/snap/amazon-fine-food-reviews

Related Work

Opinions, emotions, comments, feedback, and attitudes are central to various human activities and are significant influencers of one's behavior. Studying customer opinions help determine people's emotion of a product and its acceptance in the market. Sasikala P et al. [3] proposed a model that classifies sentiment polarity, a fundamental sentiment classification problem. The study involves the collection of the latest 3000 food reviews from Amazon. Sentiment polarity is identified from a rating scale of 1 to 5. The sentence subjectivity is detected by collecting and counting the frequency of nouns, verbs, and adjectives using parts-of-speech (POS) tagger. The opinion words are extracted with score value and visually represented as a word cloud.

Further, prediction techniques such as Naïve Bayes and Regression are used to test the data. This proposed model yielded good results with score ratings. However, it works better only for open sentiments such as scores or ratings. In the future, many features can be extracted for hidden sentiments, and prediction models are implemented.

According to the research conducted on sentiment analysis by Suci [4], the Analysis was performed on collected comments data from YouTube and Twitter social media. Naïve Bayes is chosen to perform research as it has greater accuracy in different fields, and their work is split into two phases, before and after endorsement. Although the result showed accuracies of 82% and 85.95%, the increase in accuracy is due to typos, abbreviations, and slang words. The negative Sentiment increased by 12.51%. Further research can be conducted with the inclusion of slang languages, which may increase prediction accuracy.

Pakawan [5] conducted a study on comment analysis for food recipes with 7,222 comments to help users pick the preferred ones from a variety of food recipes. The comments are classified as positive, negative, and neutral by adding up each sentence's polarity scores in the comments. The results show an accuracy of more than 90%. The accuracy of negative classification is less as there are fewer negative comments. The system cannot detect comments on a small data size. Recipe authors can use this study to improve food recipes.

Such models focus majorly on provide accurate user ratings based on written user reviews. While this is an appreciable approach to making informed decisions about their purchases, our project's highlight includes generating summaries from the user reviews. This initiative is intended for customers who are not satisfied by just viewing the ratings but also do not have enough time to read through an entire review. The generated summary will give the gist of the review and enhance user experience and reading time considerably. Also, sentiment prediction will enable customers to understand whether a product has

positive or negative feedback before reading through the review, thus helping them filter out products with positive feedback efficiently, which immensely enhances customer experience.

Objectives

The objectives of this project are listed below: -

- Exploratory Analysis of the Amazon Food Reviews dataset to identify the top 10 products with the maximum number of positive reviews and top 10 products with the maximum number of negative reviews.
- Exploratory Analysis of the Amazon Food Reviews dataset to identify the top 10 products with the most helpful reviews and top 10 products with the least helpful reviews.
- Analyze the subjectivity and polarity of comments (and make visualizations to assess polarity) for the respective Amazon products in the dataset and mine the customers' opinions, feedback, and emotions.
- Predict the Sentiment of reviews (using various predictive/machine learning models) based on rating and customer review to provide correct Sentiment and rating feedback for future customers to make informed purchase decisions based on predicted Sentiment and ratings.
- Comparison of various machine learning models on Sentiment capturing.
- Summarizing user reviews to improve customers' reading time who do not wish to spend a lot of time reading long reviews.

Dataset

The selected dataset contains reviews of fine foods from Amazon. The dataset has ten attributes and 0.57 million reviews spanning more than ten years (October 1999 – October 2012). The dataset contains user reviews of around 0.26 million customers and about 74,258 products. The attributes of the dataset are as described below: -

- Id – Row Id
- ProductId – Unique identifier for the product
- UserId – Unique identifier for the user
- ProfileName – Profile name of the user
- HelpfulnessNumerator – Count of users who found the review helpful
- HelpfulnessDenominator – Count of users who indicated whether they found review helpful or not
- Score – User entered Rating score for the product that ranges between 1 to 5
- Time - TimeStamp for Review
- Summary – User entered summary for the product
- Text – User entered product review

The 'Text' field is analyzed to determine the polarity and subjectivity of reviews and mine customers' opinions. The ProductId and the polarity counts are used to determine the top 10 products with the highest positive and negative reviews. The UserId and Helpfulness Numerator and Helpfulness Denominator fields are analyzed to determine the top 10 products with the most helpful and least helpful reviews. The 'Text' and 'Score' field will be analyzed to predict/determine the Sentiment of the review and summarize the reviews to a few words to enhance readability.

Dataset schema

The Amazon Fine Food Reviews dataset is a CSV file. The schema for the dataset with the data types and sample values are as given below.

Attribute/Field/Column Name	Data Type	Sample Value
Id	Integer	1, 2, 3
ProductId	Varchar	B001E4KFG0, B00813GRG4 B000LQOCH0
UserId	Varchar	A3SGXH7AUHU8GW, A1D87F6ZCVE5NK, ABXLMWJIXXAIN
ProfileName	Varchar	Delmartian, dll pa, Natalia Corres
HelpfulnessNumerator	Integer	1, 0, 1
HelpfulnessDenominator	Integer	1, 0, 1
Score	Integer	5, 1, 4
Time	BigInt (Data type is a big integer, as Time Stamp is stored as a big integer and not as a timestamp)	1303862400, 1346976000, 1219017600
Summary	Varchar	Good Quality Dog Food, Not as Advertised, "Delight," says it all
Text	Varchar	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than processed meat, and it smells better. My Labrador is finicky, and she appreciates this product better than most.

Data Preprocessing

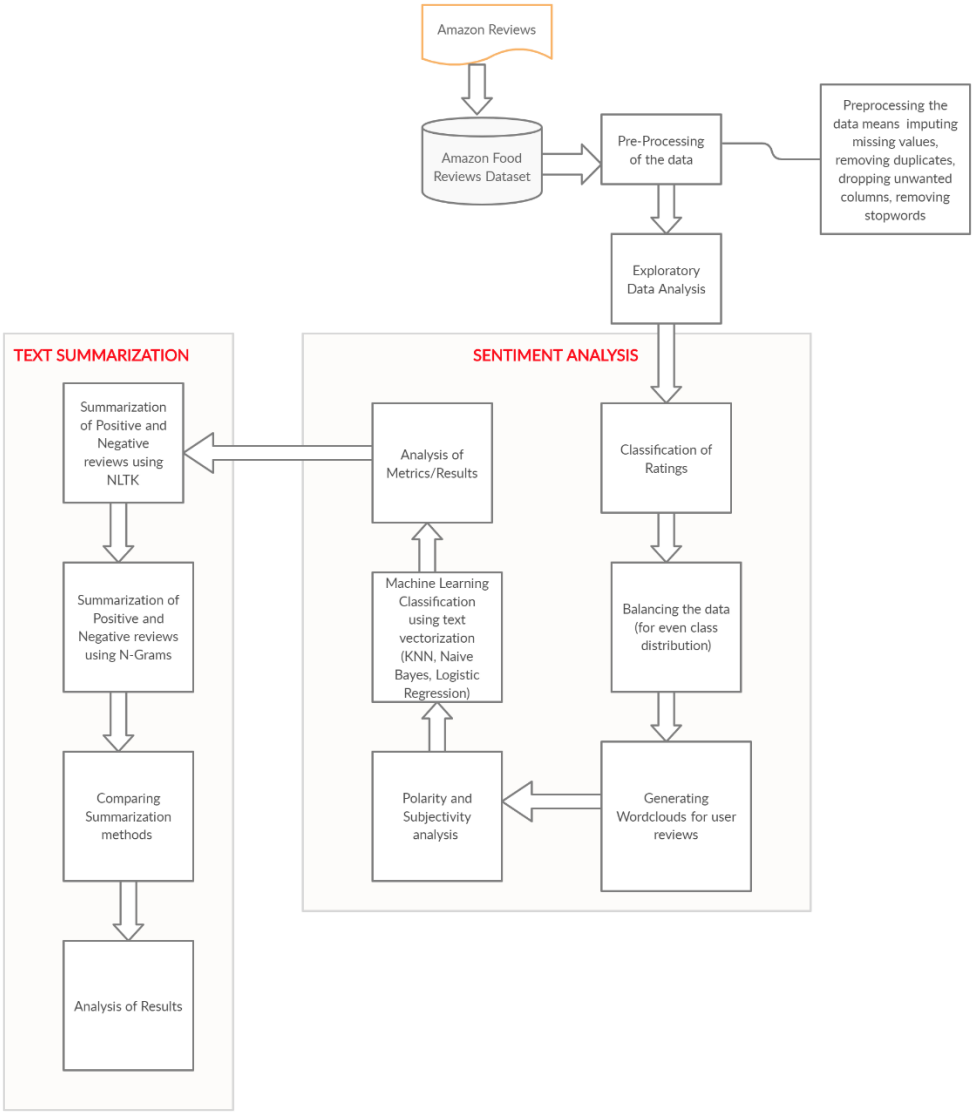
- Handling Missing Values: Missing values were handled in the dataset by imputing or replacing the missing fields.
- Handling Duplicate/Redundant Data: 281 duplicate records were found and dropped from the dataset to eliminate redundancies.
- Handling Neutral Reviews: Reviews with a rating score of 3 do not add value to the Analysis due to their non-evident polarity and are hence dropped from the dataset.
- New Variable Definitions: New variables defined as part of preprocessing and for Analysis are as follows: -

- Sentiment: This field is used to classify positive and negative reviews based on score/rating. Reviews with scores 1 and 2 are aggregated and identified with the value 0 in this field (Negative Sentiment), and the reviews with scores 4 and 5 are aggregated and identified with value 1(Positive Sentiment).
- Text_Clean: This field consists of cleaned text reviews. The customer reviews are converted to lower case and cleaned of punctuations and stopwords to ensure its availability for processing and Analysis.

The System

Architecture/Framework

Our initial system architecture implementation was an ideal scenario where the dataset was balanced, and POS tagging implementation could have been achieved easily. However, due to the volume and uneven distribution of data, we tweaked our system architecture to fit the expected outcome. The updated system architecture is as described below.



As shown in the figure, we distinguished thirteen stages:

1) Data Collection from Amazon Fine Food Reviews
   We have a dataset of product reviews and pertinent information about each product (for example, the productid, score, summary, and text for each product). Our project analyzes these data and discovers new information to help the organization and users make decisions concerning the products.

2) Preprocessing the Data
   In this stage, the data is preprocessed, and the non-textual information is removed. The data cleaning step involves removing irrelevant information like HTML tags, removing punctuation, imputing/removing missing values, removing duplicates, dropping unwanted columns, and removing stopwords.

3) Exploratory Data Analysis
   The exploratory data analysis refers to the dataset's initial investigation to discover new patterns and spot anomalies. The graphs are plotted on columns like score ratings, sentiments ("1" - Positive Sentiment, "0" - Negative Sentiment), and plotting top 10 products with positive and negative sentiments to find out the patterns in the dataset.

4) Classification of Ratings
   It is essential to aggregate the ratings to identify the Sentiment as positive or negative. A rating of 3 is considered neutral and does not serve as a useful indicator to determine Sentiment, due to which it is not included in the sample. Reviews with ratings 4 and 5 are aggregated and identified as Sentiment 1(Positive Sentiment), and ratings 1 and 2 are aggregated and identified as Sentiment 0(Negative Sentiment).

5) Balancing the data
   Data balancing (equal number of rows for positive and negative class) is done once the binarizing of review scores is completed; otherwise, the data will be imbalanced.

6) Generating Wordclouds
   Word Cloud Analysis is done to identify important words based on textual polarity. Five words/text clouds are generated to identify the texts' pertinent words – two-word clouds are based on Sentiment (positive and negative Sentiment), and the other two word clouds are based on review ratings (best and worst ratings).

7) Polarity and Subjectivity analysis
   Polarity is a float value that ranges from -1 to 1. 1 represents the statement is positive, whereas -1 means negative, and 0 represents neutral. Polarity is calculated from the sentiment function of the text blob. This function returns two properties subjectivity and polarity. Subjectivity refers to whether the statement in public opinion or factual information.

8) Machine Learning Classifiers
   Sentiment Analysis is done by predicting the Sentiment of the text using various models. Sentiment analysis is performed using three models - Logistic Regression, Naive Bayes, and K Nearest Neighbor (KNN). We split our balanced data into train and test for applying machine

learning algorithms. For each of the models, the data is subjected to two vectorization techniques - Bag of Words and TF IDF to determine which would give the best performance.

9) Analyzing the results
   The metrics are analyzed, obtained from three machine learning models to determine the best classifier for the sentiment prediction. Metrics like Accuracy, AUC, and F1 scores are used here for comparison.

10) Summarization of Positive and Negative reviews using NLTK
    Summarization of reviews is done using NLTK on the product level by aggregating the data at the product level. Once the summaries are generated, the Sentiment of the summary is captured to identify positive and negative reviews for a product. The summaries are also classified as Sentiment 0(negative Sentiment) and Sentiment 1(positive Sentiment) to compare them with the Sentiment captured from the original summary

11) Summarization of Positive and Negative reviews using N-Grams
    Summarization of reviews is done using N-Grams on the product level by aggregating the data at the product level. The trigram relative frequencies and word weights are calculated for summarization. Once the summaries are generated, the Sentiment of the summary is captured to identify positive and negative reviews for a product.

12) Comparing the Summarization Methods
    The summaries generated from NLTK and N-Grams are compared with the Sentiment captured from the original summary, and the accuracy is determined for each method.

13) Analysis of Text Summarization Methods Results
    The Text Summarization accuracy's for NLTK and N-Grams are compared to determine the best method for summarization.

NLP and Data Analytics Approaches

Sentiment Analysis

A sentiment analysis task is usually modeled as a classification problem, whereby a classifier is fed a text and returns a category, e.g., positive or negative [6]. The balanced data is divided into train and test sets. In the training process, the classification model learns to correlate a particular input (i.e., a review) to the corresponding output (i.e., if the review is positive or negative) based on the test samples used for training.
The first step in a machine learning text classifier is text vectorization, and we used bag-of-words and TFIDF approach for it. To represent the input dataset as Bag of words, we will use CountVectorizer, a transform method. CountVectorizer is a transformer that converts the input documents into a sparse matrix of features [6]. The next step is using the Classification Algorithms. In classification, we used Naïve Bayes, Logistic Regression, and KNN statistical models.

Machine Learning Classifiers (Algorithms used):

1. Logistic Regression

Logistic Regression is a supervised learning classification algorithm used to predict a target variable's probability. The nature of the target variable is dichotomous. Logistic Regression is a discriminative classifier that learns what features from the input are most useful for discriminating between the distinct likely classes. The discriminative systems are often more accurate as compared to generative systems in sentiment analysis. Hence, we choose Logistic Regression to be our baseline model to evaluate. We began by creating a Logistic Regression sentiment model, split the data to train and test, followed by text vectorization (BOW and TF-IDF). The predictions were made over our test data using the trained model.

As said by Jason Brownlee, "A classifier is only as good as the metric used to evaluate it." Here the accuracy might be the appropriate metric as the class is balanced and evenly distributed. The output of prediction for BOW and TF-IDF in Logistic Regression shows an accuracy of 93% and 92%, respectively, over test data. This means that the prediction over test data using BOW and TF-IDF is 93% and 92% accurate. Here we calculated other evaluation metrics as well.

Model Name	AUC Score	F-Score
Logistic Regression - Bag of Words	0.9270537258490439	0.9270464378117974
Logistic Regression - TF-IDF	0.9150785360980027	0.9150473265852821

2. Naïve Bayes

The Naive Bayes Classifier is a well-known machine learning classifier with applications in NLP. It can achieve above-average performance in different tasks like sentiment analysis. A naïve Bayes classifier is a generative classifier that builds a model of how a class could generate input data. It returns the class, which is most likely to have developed for a given observation, and uses probabilistic algorithms (Bayes Theorem) to predict a text. Hence, we choose naïve Bayes as our other Algorithm. For it, we split the data to train and test, followed by text vectorization (using BOW and TF-IDF) in the Naive Bayes model. We have tried multinomial Naive Bayes on BOW features and TF-IDF features.

Here the accuracy might be the appropriate metric as the class is balanced and evenly distributed. The output of prediction for BOW and TF-IDF in Naïve Bayes shows an accuracy of 89.03% and 91%, respectively, over test data. This means that Sentiment's prediction over test data using BOW and TF-IDF is 89.03% and 91% accurate. Here we calculated other evaluation metrics as well.

Model Name	AUC Score	F-Score
Naive Bayes - Bag of Words	0.8900771233203372	0.8903573565214484
Naive Bayes - TF-IDF	0.9139306130389728	0.913718681380323

New Algorithm(s)

Text Summarization

Summarization is intended for customers who are not satisfied by just viewing the ratings but also do not have enough time to read through an entire review. The generated summary will give the gist of the review and enhance the user experience.

Two summarization techniques are used to generate summaries of user reviews. These include summarization using NLTK and summarization using Ngrams(trigrams). Summarization of reviews is done on the product level by aggregating the data at the product level. Once the summaries are generated, the Sentiment of the summary is captured using the TextBlob polarity function to identify positive and negative reviews for a product. The summaries are also classified as Sentiment 0(negative Sentiment) and Sentiment 1(positive Sentiment) in order to compare them with the Sentiment captured from the original summary and determine the accuracy of the Sentiment captured by summaries generated using these two summarization techniques

We begin with a sampling of the data because the data size is enormous, and the total execution time for summarization was 30mins. A random sample of 2000 rows is taken to generate the summaries. The data frame is aggregated based on 'ProductId,' 'Sentiment,' and 'Text_Clean' columns to generate positive and negative summaries for each product.

Summarization using NLTK:
Sentence summaries are generated after removing stop words, tokenizing, lemmatizing, and calculating word weighted frequencies. Based on the sentence summary of the cleaned text, the polarity of the sentences is generated (Positive Review/Neutral Review/Negative Review). The 'cleaned text' polarity and 'generated summary' polarity are compared for calculating the Sentiment's accuracy. The accuracy of the sentiment capture by summaries generated using NLTK is 65.25%.

Summarizing using N-Grams:
The N-Grams and relative frequencies are generated for the cleaned text, followed by calculating weights of the N-Grams for Summarization. We used trigram for generating the summaries. Based on the sentence summary of the cleaned text generated using trigrams, the sentences' polarity is generated (Positive Review/Neutral Review/Negative Review). The 'cleaned text' polarity and 'generated summary' polarity are compared for calculating the Sentiment's accuracy. The accuracy of the sentiment capture by summaries generated using N-Grams is 64.8%.

The accuracy will change based on the random 2000 samples selected from the dataset. However, it will lie in the range of 60%-70% for both techniques. Apart from the above two summarization techniques, Gensim was also intended to summarize and compare Sentiment. However, due to its constraint of requiring more than one sentence as input, it could not be implemented as a major proportion of the user reviews comprises only one sentence.

Summarization Technique	Sentiment Capture Accuracy
NLTK	65.25
Ngrams(Trigrams)	64.8

Other Algorithm(s)

K-Nearest Neighbour

The K-Nearest Neighbor (KNN) is one of the simplest lazy machine learning algorithms. The k-Nearest Neighbor classifier assumes that the classification of an instance is most like the classification of some other cases that are nearby in the vector space. It does not depend on prior probabilities than the Naive Bayes classifier and Logistic regression classifier and is computationally efficient. The preliminary calculation is ordering training documents to obtain the *k* nearest neighbors for the test document. We created a KNN model, split the data to train and test, followed by text vectorization (BOW and TF-IDF). The predictions were made over our test data using the trained model.

The prediction output for BOW and TF-IDF in KNN shows 73% and 76% accuracy, respectively, over test data. This means that the prediction over test data using BOW and TF-IDF is 73% and 76% accurate, which is very low compared to Naïve Bayes and Logistic Regression. Here we calculated other evaluation metrics as well.

Model Name	AUC Score	F-Score
K-Neighbors Classifier - Bag of Words	0.719763447739878	0.7184013630540727
K-Neighbors Classifier - TF-IDF	0.7561277557723972	0.7553788964569014

SW/HW Development platforms

The hardware and software platforms used for the project are as described below: -

Hardware Development Platform
RAM: 8 GB/16 GB
Processor: 64-bit quad-core

Software Development Platform
Operating System: Windows/Mac/Linux
Tools: Jupyter Notebook, Tableau (for exploratory Analysis)
Language: Python

Experimental Results and Analysis

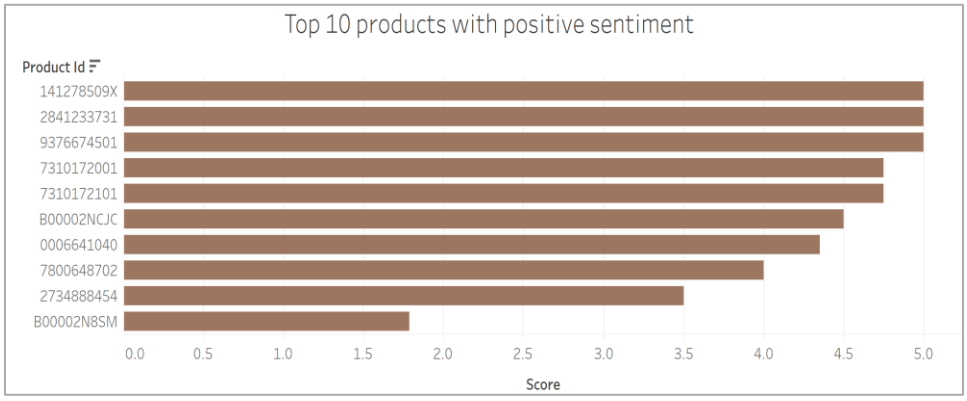
The newly created fields, such as Sentiment and Text_Clean, are used in building machine learning models. Text_Clean is a predictor variable, whereas Sentiment is the target variable. Since Sentiment is a categorical variable, Logistic Regression, KNN, and Naïve Bayes machine learning classifiers are built to make performance comparisons and obtain the best sentiment prediction model. Bag of Words and TF-IDF vectorization techniques are used in determining word similarities and semantics in sentiment prediction. The below table illustrates the different results obtained from three machine learning models.

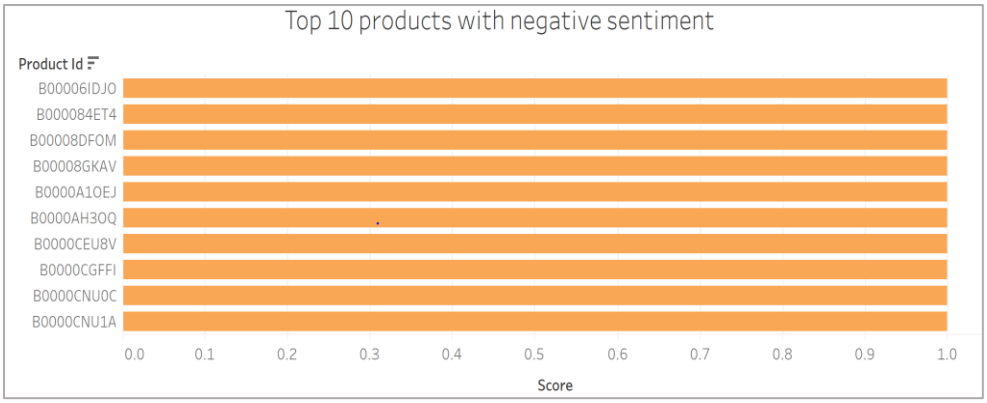
Vectorization Methods	Models	AUC	F1-Score	Accuracy
	Logistic Regression	0.9270	0.9270	**93%**
Bag-of-Words	Naïve Bayes	0.8900	0.8903	89%
	K-Nearest Neighbors	0.719	0.718	73%
	Logistic Regression	0.9150	0.9150	**92%**
TF-IDF	Naïve Bayes	0.9139	0.9137	91%
	K-Nearest Neighbors	0.756	0.755	76%

The table depicts that the Logistic Regression performed well in sentiment prediction compared to the other two models' performances. Logistic Regression (LR) with Bag-of-Words vectorization technique gave an accuracy of 93%. LR with TF-IDF gave relatively better accuracy of 92%. Naïve Bayes follows this with an accuracy of 91% using the TF-IDF method. It is observable that TF-IDF word embeddings resulted in better accuracy in all models except LR due to its consideration for high frequency or rarity of words. Frequently appearing words hold less weight than rare words in determining whether the review is positive or negative.

One of this project's objectives is to list the top ten products with a positive and negative statement and the top ten products with the most and least helpful reviews. To achieve this, Tableau software is used to represent the product list graphically. This software provides a more significant interactive data visualization platform, which is a revolutionary approach in business intelligence.

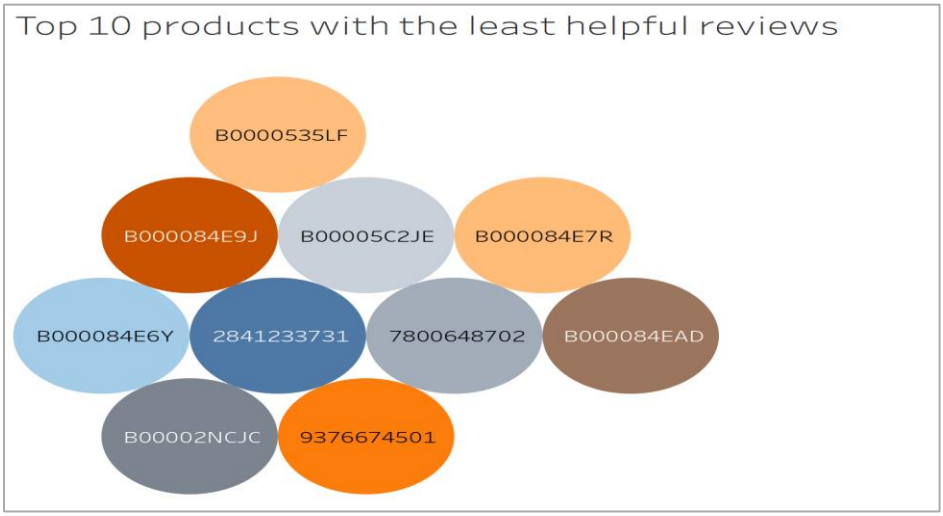
The below figure represents the top ten product Ids that have positive Sentiment, based on the customers' rating score to a product.





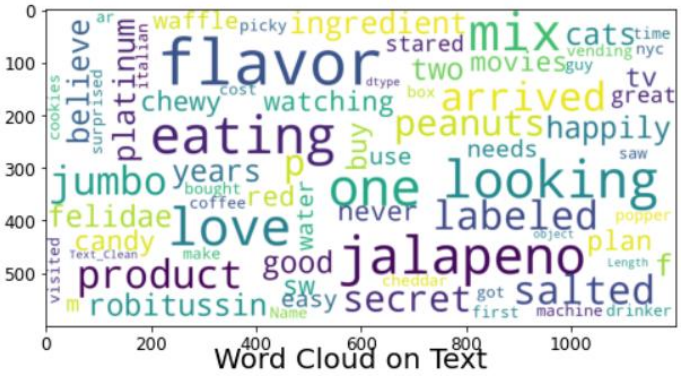
The above figure shows the top ten product Ids with negative Sentiment, given the least rating score by the customers.

The top ten product Ids are clearly shown that are having the most helpful reviews from the above figure.

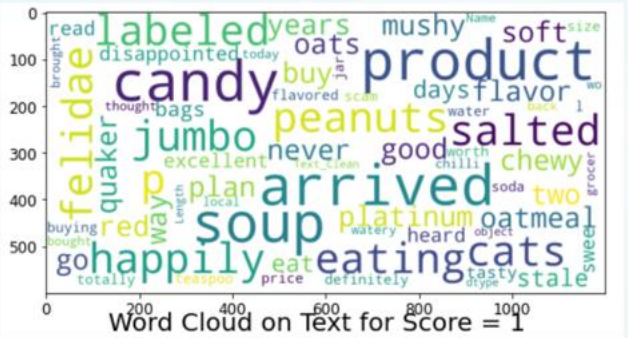


The above diagram depicts the top ten product Ids that received the least helpful reviews.

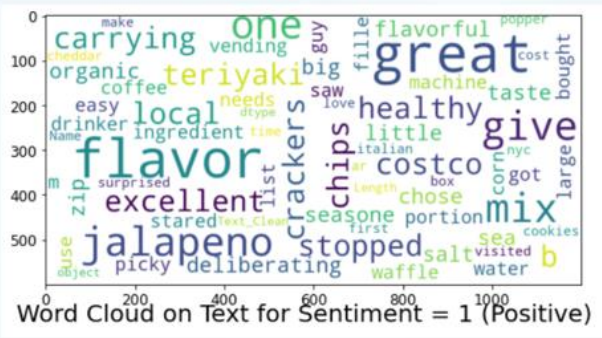
In addition, the data is analyzed by plotting word clouds. The below plot shows word cloud of cleaned user reviews.



The best rated reviews with rating score of 5 and the worst rated reviews with rating score of 1 are plotted respectively as shown below.



Moreover, negative reviews with a sentiment score of 0 and the positive reviews with a sentiment score of 1 are graphically visualized with the help of word cloud, respectively as shown in the below diagram.

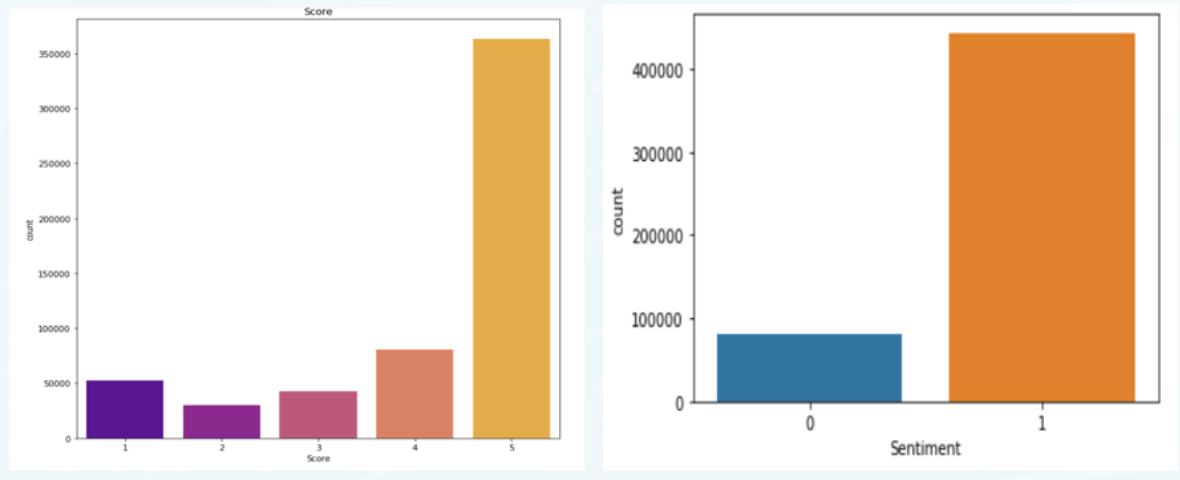


Another objective is to generate summaries of user text reviews. So, NLTK is used in generating summaries. The generated summaries will give the gist of the reviews presented by customers to a product. Hence, summaries will enhance the user experience for customers who are satisfied by viewing the ratings and do not have enough time to read through an entire review.

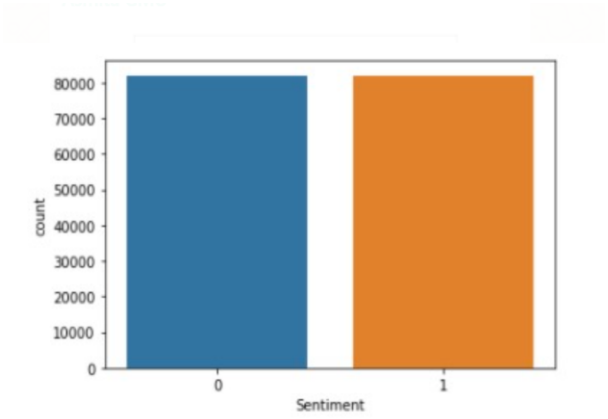
For instance, "The package shown on the web page showed milk chocolate, and had a pink/white wrapper. What came was brown/white wrapper and dark chocolate. It is okay, but not what I ordered!" is a review given by a customer to a product.

The proposed system generates a summary of this review as follows: "chocolate white wrapper package has shown web page showed milk pink came brown dark okay ordered."

After binarizing the review scores, the dataset was appeared to be considerably imbalanced as shown in below diagram. 84% of the reviews were marked as sentiment 1 (for score = 4 and score = 5) and only 16% were 0 (for score = 1 and score = 2). As a result, there are high chances that it would miss classification due to high positive reviews on training the Algorithm with this data. This implies that if we have a classifier that would always predict a review as positive, it will easily achieve 84% accuracy.



Hence, the proposed solution balances uneven class distribution. Resampling the data would help to even class distribution by prompting it to distinguish classes from analyzing underlying patterns in the data. Random Under sampling technique was used to balance class distribution by randomly eliminating majority class examples. A binary case ratio was used to randomly select instances from the majority and minority classes in a 50/50 ratio. It will reduce the data with 4 and 5 scores (majority class) to match the data with 1 and 2 scores (minority class). The below diagram shows the balanced class distribution of Sentiment 0 and 1.



From this data, the model (Logistic Regression with Bag of Words) can attain the highest accuracy of 93% in predicting the Sentiment.

Conclusion

Customer experience is a significant metric for any industry in measuring its performance in the market. Amazon Fine Food Reviews consists of customer reviews for products about various categories. These reviews serve as an indicator of product quality for customers who purchase the products online. Hence, a machine learning technique is implemented to capture and learn the sentiments in the user reviews and predict the Sentiment of an unseen customer review. Machine Learning models such as Logistic Regression, Naïve Bayes, and K-Nearest Neighbor are applied to the user reviews to compare the best performing model for sentiment prediction using Bag-of-Words TF-IDF vectorization techniques. TF-IDF word embeddings gave a higher accuracy for sentiment analysis due to its consideration for the high frequency or rarity of words. The logistic regression model resulted in the highest accuracy of 93% with the Bag-of-Words method. Sentiment analysis using NLTK and N-grams (Trigrams) summarization techniques resulted in capturing Sentiment with an accuracy of 65.25% and 64.8%, respectively. Summarization of reviews is done using NLTK to reduce the reading time for customers. We used Bag of words and TF-IDF vectorization approaches in our models to predict Sentiment. In the future, we would like to explore the classification accuracy obtained when using Word2Vec vectorization for further comparison of vectorization word embeddings in determining Sentiment. Sentiment capture accuracy lies between 60-70% for NLTK and N-Grams currently. We would like to explore other approaches, such as skip grams' potential application in summarizing text. Skip grams is a concept that has been covered in one of the papers provided for paper presentation, and we are keen on exploring the concept's capabilities for summarizing reviews and predicting Sentiment.

References

[1] Amazon. (n.d.). In Wikipedia. Retrieved October 18, 2020, from https://en.wikipedia.org/wiki/Amazon_(company)

[2] Stanford Network Analysis Project. (2016, January). Amazon Fine Food Reviews, Version 2. Retrieved October 18, 2020, from https://www.kaggle.com/snap/amazon-fine-food-reviews/metadata.

[3] Sasikala P et al. (2018). Sentiment Analysis of Online Food Reviews using Customer Ratings. International Journal of Pure and Applied Mathematics. Retrieved October 18, 2020, from https://www.researchgate.net/profile/P_Sasikala3/publication/325957749_Sentiment_Analysis_of_Online_Food_Reviews_using_Customer_Ratings/links/5b2f58850f7e9b0df5c33db0/Sentiment-Analysis-of-Online-Food-Reviews-using-Customer-Ratings.pdf

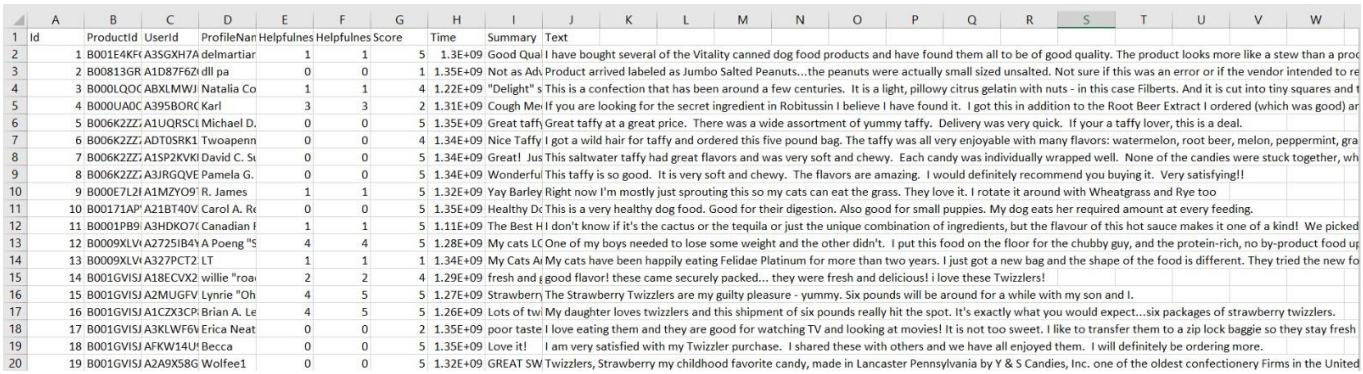
[4] Suci et al. (2018). Sentiment Analysis of Product Reviews using Naïve Bayes Algorithm: A Case Study. IEEE. Retrieved October 18, 2020, from https://ieeexplore-ieee-org.mutex.gmu.edu/document/8878528

[5] Pakawan et al. (2015). Comment Analysis of food recipe preferences. IEEE. Retrieved October 18, 2020, from https://ieeexplore-ieee-org.mutex.gmu.edu/document/7207

[6] MonkeyLearn. Sentiment Analysis: A Definitive Guide. IEEE. Retrieved November 28, 2020, from https://monkeylearn.com/sentiment-analysis/?ref=hackernoon.com

Appendix

A screenshot of the dataset



	A	B	C	D	E	F	G	H	I	J
1	Id	ProductId	UserId	ProfileNan	Helpfulnes	Helpfulnes	Score	Time	Summary	Text
2	1	B001E4KF	A3SGXH7A	delmartian	1	1	5	1.3E+09	Good Qual	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a proc
3	2	B00813GR	A1D87F6Z	dll pa	0	0	1	1.35E+09	Not as Adv	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was an error or if the vendor intended to re
4	3	B000LQOC	ABXLMWJ	Natalia Co	1	1	4	1.22E+09	"Delight" s	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut into tiny squares and t
5	4	B000UA0C	A395BORC	Karl	3	3	2	1.31E+09	Cough Me	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer Extract I ordered (which was good) ar
6	5	B006K2ZZ	A1UQRSCL	Michael D.	0	0	5	1.35E+09	Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.
7	6	B006K2ZZ	ADT0SRK1	Twoapenn	0	0	4	1.34E+09	Nice Taffy	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelon, root beer, melon, peppermint, gra
8	7	B006K2ZZ	A1SP2KVK	David C. St	0	0	5	1.34E+09	Great! Jus	This saltwater taffy had great flavors and was very soft and chewy. Each candy was individually wrapped well. None of the candies were stuck together, wh
9	8	B006K2ZZ	A3JRGQVE	Pamela G.	0	0	5	1.34E+09	Wonderful	This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. Very satisfying!!
10	9	B000E7L2I	A1MZYO9T	R. James	1	1	5	1.32E+09	Yay Barley	Right now I'm mostly just sprouting this so my cats can eat the grass. They love it. I rotate it around with Wheatgrass and Rye too
11	10	B00171AP	A21BT40V	Carol A. Re	0	0	5	1.35E+09	Healthy Dc	This is a very healthy dog food. Good for their digestion. Also good for small puppies. My dog eats her required amount at every feeding.
12	11	B0001PB9	A3HDKO7(	Canadian F	1	1	5	1.11E+09	The Best H	I don't know if it's the cactus or the tequila or just the unique combination of ingredients, but the flavour of this hot sauce makes it one of a kind! We picked
13	12	B0009XLV	A2725IB4Y	A Poeng "S	4	4	5	1.28E+09	My cats LC	One of my boys needed to lose some weight and the other didn't. I put this food on the floor for the chubby guy, and the protein-rich, no by-product food up
14	13	B0009XLV	A327PCT2	LT	1	1	1	1.34E+09	My Cats Ai	My cats have been happily eating Felidae Platinum for more than two years. I just got a new bag and the shape of the food is different. They tried the new fo
15	14	B001GVISJ	A18ECVX2	willie "roa	2	2	4	1.29E+09	fresh and g	good flavor! these came securely packed... they were fresh and delicious! i love these Twizzlers!
16	15	B001GVISJ	A2MUGFV	Lynrie "Oh	4	5	5	1.27E+09	Strawberry	The Strawberry Twizzlers are my guilty pleasure - yummy. Six pounds will be around for a while with my son and I.
17	16	B001GVISJ	A1CZX3CP	Brian A. Le	4	5	5	1.26E+09	Lots of tw	My daughter loves twizzlers and this shipment of six pounds really hit the spot. It's exactly what you would expect...six packages of strawberry twizzlers.
18	17	B001GVISJ	A3KLWF6V	Erica Neat	0	0	2	1.35E+09	poor taste	I love eating them and they are good for watching TV and looking at movies! It is not too sweet. I like to transfer them to a zip lock baggie so they stay fresh
19	18	B001GVISJ	AFKW14U	Becca	0	0	5	1.35E+09	Love it!	I am very satisfied with my Twizzler purchase. I shared these with others and we have all enjoyed them. I will definitely be ordering more.
20	19	B001GVISJ	A2A9X58G	Wolfee1	0	0	5	1.32E+09	GREAT SW	Twizzlers, Strawberry my childhood favorite candy, made in Lancaster Pennsylvania by Y & S Candies, Inc. one of the oldest confectionery Firms in the United