BINARY CLASSIFICATION OF AMAZON FINE FOOD REVIEWS

Group number: 27

Submitted by:
Jatin Lamba (A20362075)
Sharwari Alat (A20359670)

Title: Binary Classification of Amazon Fine Food Reviews

Group Number: 27

First Name	Last Name	Online Students? (Y or N)	Monday or Tuesday	Shared with ITMD 527? (Y or N)
Jatin	Lamba	N	Tuesday	N
Sharwari	Alat	N	Tuesday	N

Table of Contents

1. Introduction and Motivations

In this era there are thousands of websites which provide online services. Evaluation of products or services provided is done by its users in the form of reviews. Reviews are key factor for the success of Organizations such as Amazon, eBay etc which have huge impact on market because of availability of useful reviews provided by them.

Generally users of the service post their reviews about the respective product or services in the form of rating and comments. However, sometimes reviews are provided in the form of comments without rating the product (Example: reviews provided on social media like Facebook, Twitter etc.). ,

Rating of reviews generally helps to determine the polarity (positive or negative) of the review but in case of reviews without rating classification can be done using various sentiment classification techniques.
In this project we are going to determine the polarity of reviews using sentiment classification and also evaluate the different classifiers such Decision Tree Classifier, Logistic Regression and Naive Bayes for accuracy.

The purpose of our project is analyze the large dataset for sentiment polarity and also evaluate the classifiers for the same. For this purpose we used Amazon Fine Food Reviews where it has data set of around 568,454 food reviews Amazon users. This classification will improve the selection of helpful reviews and also guide the users for writing the helpful reviews.

2. Data Description

We will be working on the Amazon Fine Food Reviews. We have data set of around **568,454** food reviews Amazon users. It is available at: https://snap.stanford.edu/data/web-FineFoods.html

Each review has the following 10 features:

Name of Field	Description
Id	
ProductId	Unique identifier for the product
UserId	Unique identifier for the user
ProfileName	
HelpfulnessNumerator	Number of users who found the review helpful
HelpfulnessDenominator	Number of users who indicated whether they found the review helpful
Score	Rating between 1 and 5
Time -	Timestamp for the review
Summary	Brief summary of the review
Text	Text of the review

So out of these 10 features for the reviews ‘score’, ‘summary’ and ‘text’ are the ones having some kind of predictive value. Also ‘text’ is kind of redundant as summary is sufficient to extract.

```
1 SELECT * FROM Reviews;
```

All Fields

Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat...	Score	Time	Summary	Text
1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several o...
2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled ...
3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia C...	1	1	4	1219017600	"Delight" says it all	This is a confection that...
4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for th...
5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. W...	0	0	5	1350777600	Great taffy	Great taffy at a great pr...
6	B006K2ZZ7K	ADT0SRK1MGOEU	Twoapennything	0	0	4	1342051200	Nice Taffy	I got a wild hair for taffy...
7	B006K2ZZ7K	A1SP2KVKFXXRU1	David C. Sullivan	0	0	5	1340150400	Great! Just as good as t...	This saltwater taffy had ...
8	B006K2ZZ7K	A3JRGQVEQN31IQ	Pamela G. Williams	0	0	5	1336003200	Wonderful, tasty taffy	This taffy is so good. It ...
9	B000E7L2R4	A1MZYO9TZK0BBI	R. James	1	1	5	1322006400	Yay Barley	Right now I'm mostly ju...
10	B00171APVA	A21BT40VZCCYT4	Carol A. Reed	0	0	5	1351209600	Healthy Dog Food	This is a very healthy do...
11	B0001PB9FE	A3HDKO7OW0QNK4	Canadian Fan	1	1	5	1107820800	The Best Hot Sauce in t...	I don't know if it's the c...
12	B0009XLVG0	A27251B4YY9JEB	A Poeng "SparkyGoHom...	4	4	5	1282867200	My cats LOVE this "diet"...	One of my boys needed...
13	B0009XLVG0	A327PCT23YH90	LT	1	1	1	1339545600	My Cats Are Not Fans of...	My cats have been happ...
14	B001GVISJM	A18ECVX2RJ7HUE	willie "roadie"	2	2	4	1288915200	fresh and greasy!	good flavor! these came...
15	B001GVISJM	A2MUGFV2TDQ47K	Lynrie "Oh HELL no"	4	5	5	1268352000	Strawberry Twizzlers - Y...	The Strawberry Twizzler...
16	B001GVISJM	A1CZX3CP8IKQIJ	Brian A. Lee	4	5	5	1262044800	Lots of twizzlers, just w...	My daughter loves twizz...
17	B001GVISJM	A3KLWF6WQ5BNYO	Erica Neathery	0	0	2	1348099200	poor taste	I love eating them and t...
18	B001GVISJM	AFKW14U97Z6QO	Becca	0	0	5	1345075200	Love it!	I am very satisfied with ...
19	B001GVISJM	A2A9X58G2GTBLP	Wolfee1	0	0	5	1324598400	GREAT SWEET CANDY!	Twizzlers, Strawberry m...
20	B001GVISJM	A3IV7CL2C13K2U	Greg	0	0	5	1318032000	Home delivered twizlers	Candy was delivered ver...
21	B001GVISJM	A1WO0KGLPR5PV6	mom2emma	0	0	5	1313452800	Always fresh	My husband is a Twizzle...
22	B001GVISJM	AZOF9E17RGZH8	Tammy Anderson	0	0	5	1308960000	TWIZZLERS	I bought these for my h...
23	B001GVISJM	ARYVQL4N737A1	Charles Brown	0	0	5	1304899200	Delicious product!	I can remember buying ...
24	B001GVISJM	AJ613OLZZUG7V	Mare's	0	0	5	1304467200	Twizzlers	I love this candy. After w...
25	B001GVISJM	A22P2J09NJ9HKE	S. Cabanaugh "jilly pep...	0	0	5	1295481600	Please sell these in Mexi...	I have lived out of the U...
26	B001GVISJM	A3FONPR03H3PJS	Deborah S. Linzer "Cat ...	0	0	5	1288310400	Twizzlers - Strawberry	Product received is as a...
27	B001GVISJM	A3RXAU2N8KV45G	lady21	0	1	1	1332633600	Nasty No flavor	The candy is just red . N...

3. Research Problems and Solutions

Observing the data it is can be seen that score has value in the range of 1 to 5. For the primary analysis of data we used binary classification method for label distribution. We choose score as a label and distributed it into positive and negative labels. Labels which are equal to or less than three are distributed as negative and reviews with score 4 and 5 are distributed as positive.

3.1. Research Problems:

- We will be using **score** as our **label** for differentiating the reviews as positive and negative.
- In this the **challenge** for us is to filter the raw text data of reviews column of the dataset in the form of frequency of the words.
- We will be **implementing** the classification algorithms that can be applied efficiently to obtain the best classifier for the dataset.

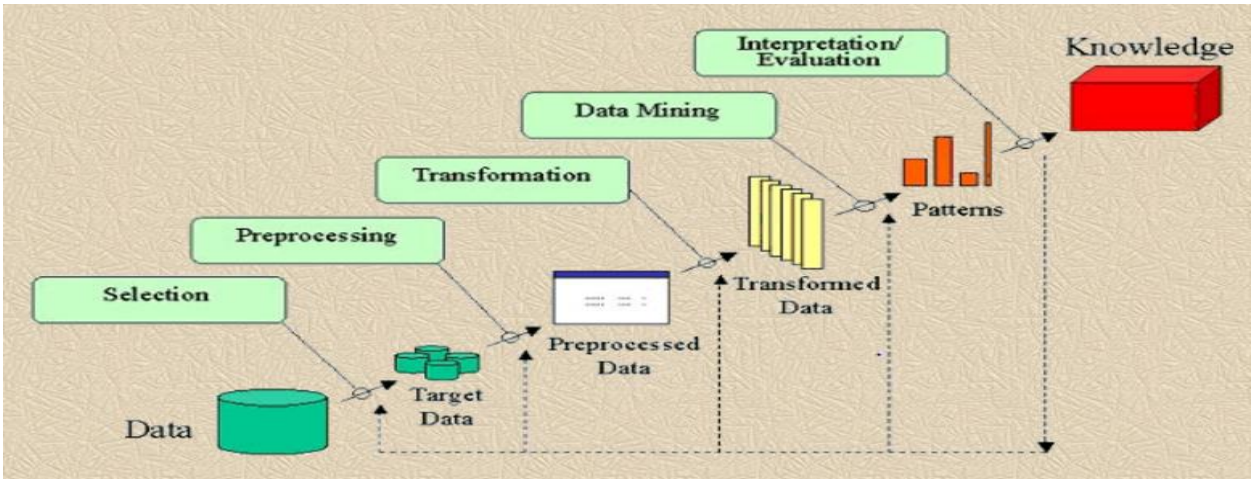
3.2. Potential Solutions:

- Our plan is to work on the dataset by fetching the score and predict the positive or negative with the use of **Binary Classification**.
- We will be using **feature reduction or selection** to reduce the size of feature set.
- We will be making use of **Decision Tree Classifier, Naïve Bayes Classifier and Logistic Regression Classifier** models on the training data set and will evaluate it against the test data set.

4. KDD

4.1. Data Processing

In general KDD involves following steps:



4.1.1. Creating the target dataset:

- **Selecting a data set:** For the purpose of project is to perform sentiment classification and evaluating the classifiers.
- **Focusing on a subset of variables or data samples:** In our dataset fields such as “score”, “text”, “summary” have some predictive value hence we considered those 3 fields out of 10 on which discovery is to be performed.

4.1.2. Data cleaning and preprocessing:

To make data more useful following preprocessing techniques are applied

- **Stop words removal:** commonly used words in language which doesn't have any predictive value associated with it are removed.
- **Lemmatization:** reduced inflectional and some derivationally related forms of a word to a common base for further prediction of sentiments.
- **Punctuation removal:** common punctuation marks are removed for clarification and to avoid the redundancy of words which are taken into account.
- **Upper-to-lower:** converts all uppercase characters to lowercase

4.1.3. Splitting data into Train and Test dataset:

From the label distribution it can be seen that distribution of reviews is skewed as it has 78.07% of the reviews are positive hence we split the data into train and test dataset.

4.1.4. Vectorization:

Size of "text" column in dataset is not fixed and most of the classification algorithms perform inefficiently on variable size input. This problem was solved by implementing Bag-of-words strategy which converts variable length columns into vectors with numerical values. It involves following steps:

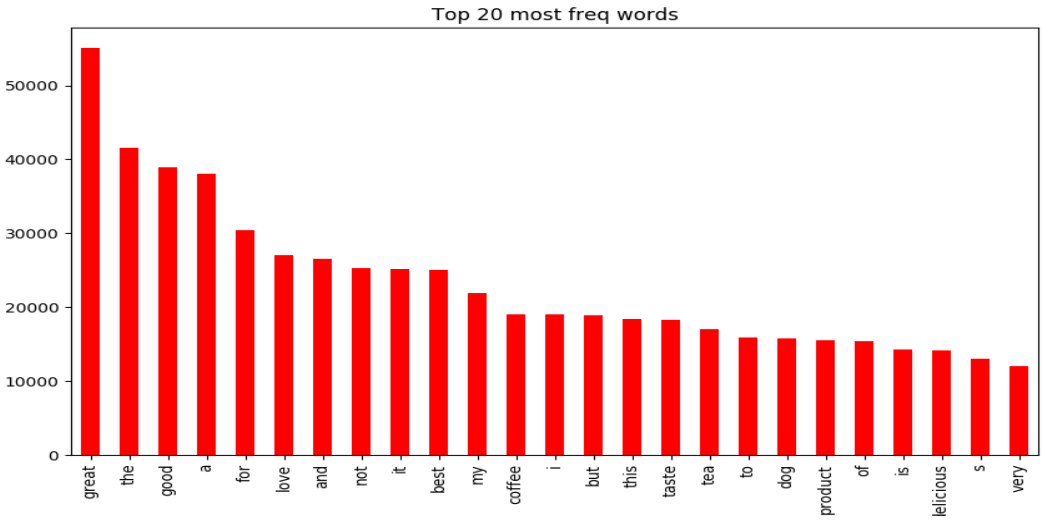
- **Tokenization:** splitting input into tokens representing single word
- **Counting:** count of frequency of each word
- **Normalization:** reduction of importance of most occurring words

4.1.5. Data reduction and projection:

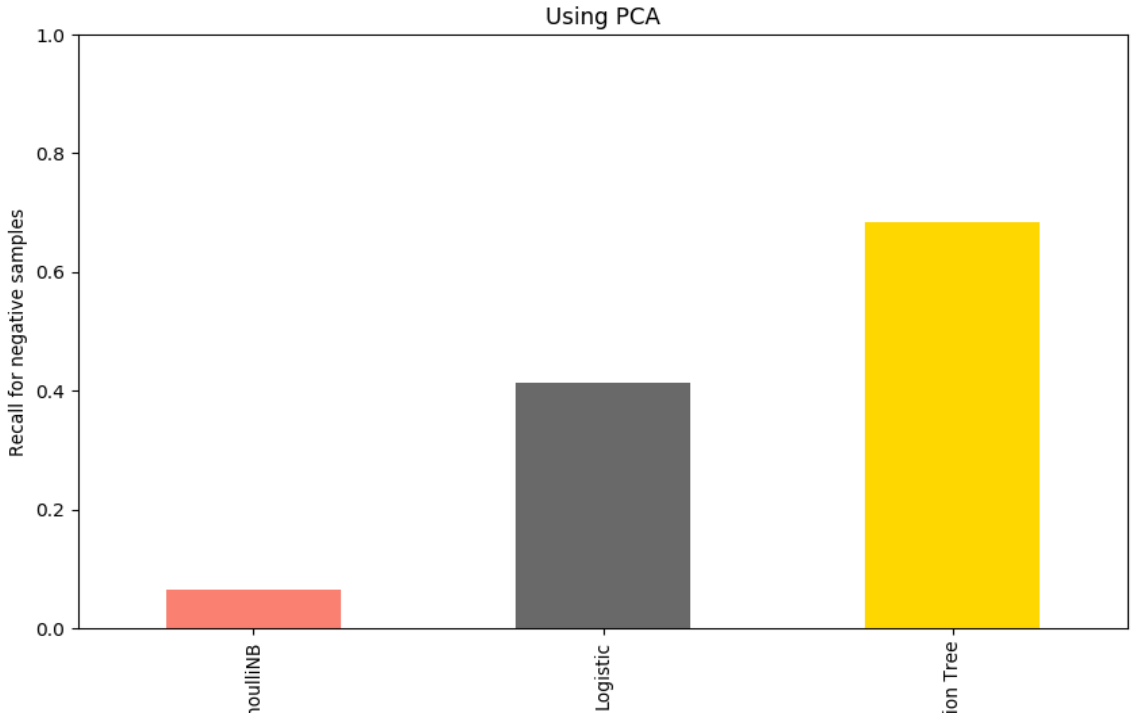
After implementation of above 3 preprocessing steps 27048 unique words were emitted. These words form a feature set for next processing steps. This was done in two steps one with feature reduction/selection process and one without the feature reduction/selection process by running the classifiers on the whole dataset to obtain the best model with accurate performance.

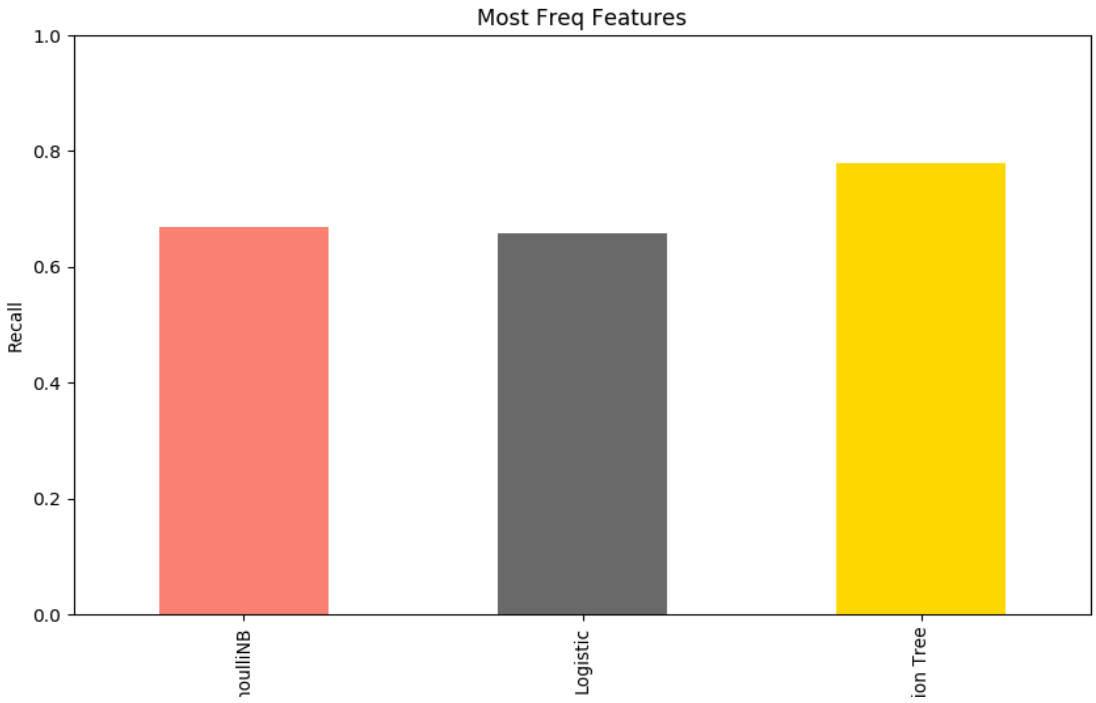
Feature set of 27048 is too big to handle and hence there was need to reduce the feature set for future processing. This can be done by Feature reduction and Feature selection using many different ways.

- **PCA (Principle Component Analysis):** as PCA converts the set of n-dimensional space into smaller dimensional space. Using PCA we reduced feature set up to 200 components with the help of Truncated SVD.
- **Most Frequent Words:** in this case we took subset of most frequently occurred words in the feature set. Using this technique feature set was reduced to 5000 components that is 1/5th of original feature set are vectorized using **Tf-idf transformer**.



From the above results even though value of PCA is lesser, we plotted the performance of classifiers against the PCA and Most Frequent Words.





From the above results it is clear that performance of a model is better using Most Frequent Words than PCA. Therefore **Most frequent words** are considered for the further analysis.

To check the performance of classifiers on the dataset without using feature reduction/ selection process. This was performed with the help of following 3 sequences:

- **Unigram:** To get better results further we made use of sequence of words as sometimes these sequence have an effect in the better prediction of model. For example “not bad” or “not good” can mean different when use of these words are made individually.

- **Bigram:** In bigram all these sequence of adjacent words are also considered as features apart from Unigrams. Hence, the word phrases like “not good”, “not bad”, “pretty bad”, “pretty well”, etc will also have a predictive value which wasn’t present during the use of Unigrams.

- **Trigram:** Now in trigram all the sequences of 3 adjacent words are brought into consideration as a separate feature apart from above two sequences.

4.2. Data Mining Tasks and Processes

4.2.1. Choosing Data mining task:

Among several tasks such as classification, clustering, regression etc. we choose classification task for mining the data because we want to classify the reviews on the basis of their usefulness.

4.2.2. Choosing Data mining algorithms:

In this step of KDD process we choose appropriate algorithm for the dataset on the basis of accuracy.

- **Selection of classification algorithm:** We selected following classification algorithms for the evaluation:
    - Decision Tree classifier
    - Naive Bayes classifier
    - Logistic Regression classifier
- **Deciding appropriate model and parameters:** trained above 3 classifiers against training dataset and evaluated for test set. Size of dataset is huge hence classifiers such as K-Nearest Neighbors or Random Forests etc. does not perform efficiently.

4.2.3. Data mining:

- **Searching patterns of interest:** implemented evaluation matrix for searching the patterns and evaluate respective classifier.

4.2.4. Interpreting mined patterns:

Interpreted patterns found in the process of data mining using:

- Word Cloud
- Charts

4.2.5. Consolidating discovered knowledge:

Came up with optimal prediction model for the current dataset.

5. Evaluations and Results

5.1. Evaluation Methods

Hold-out evaluation method was used for splitting the dataset into training and test dataset. Three variants were used for train and test data i.e.

- Train (70%) test (30%)
- Train (75%) test (25%)
- Train (80%) test (20%)

The best accuracy was obtained on train (75%) test (25%)

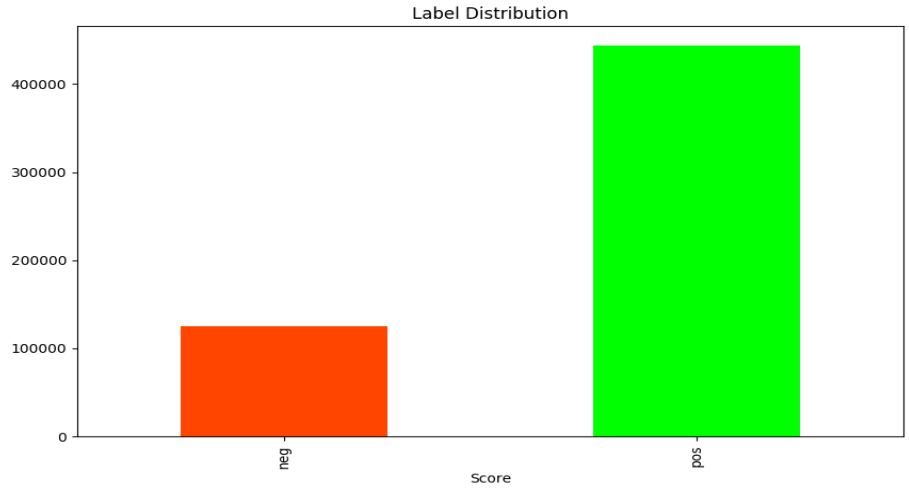
Sparse matrix was used for the representation of the evaluations of the classifiers and confusion matrix plots the True labels against predicted labels for the visualizing the classification report.

Our first solution is to fetch the score and predict positive or negative and this is performed with the help of binary classification and this is the perfect approach as we are working with only two attributes of the label i.e positive and negative.

After implementing Binary Classification we found below results:

Positive reviews: 443777 (78.07 %)

Negative Reviews: 124677 (21.93 %)



Secondly, as the data is big enough so we need to reduce the size of data so that it can work more accurately with the classifiers. This will be obtained by the process of feature reduction and feature selection with the use of PCA and most frequent features respectively.

5.2. Results and Findings:

5.2.1. Finding Top 5 useful reviews

Top 5 review were obtained with the help of most frequent positive occurring words in the review by the bag-of-words methods.

Home

```
1 SELECT Summary
2 FROM Reviews
3 WHERE Score = 5
4 ORDER BY (HelpfulnessNumerator / HelpfulnessDenominator) desc
5 LIMIT 5;
```

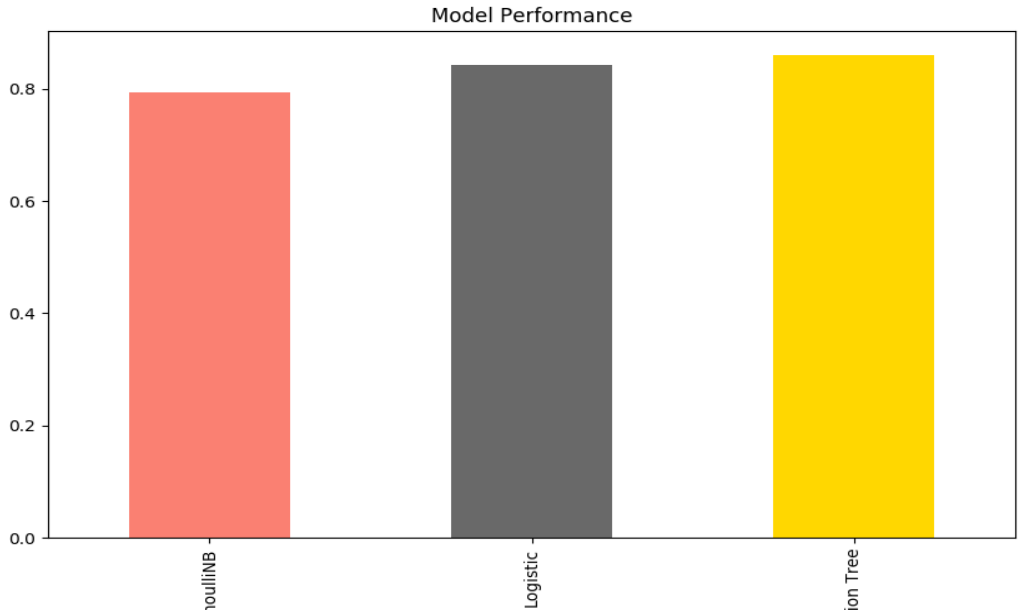
All Fields

Summary
Bought This for My Son at College
Good Quality Dog Food
Yay Barley
The Best Hot Sauce in the World
My cats LOVE this "diet" food better than their regular food

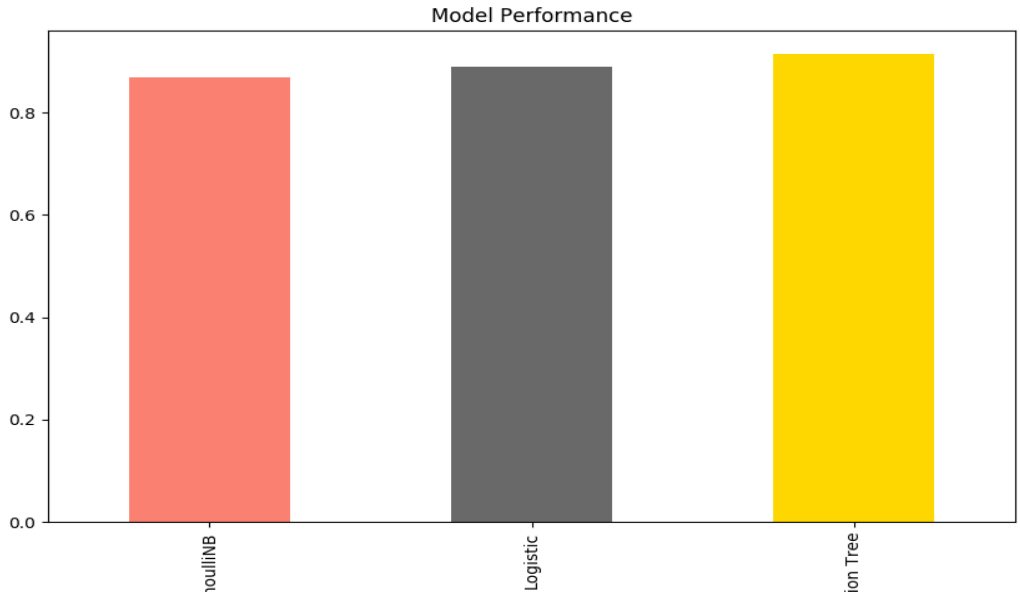
5.2.2. Comparison of model performance using PCA and Most Frequent Features

A significant improvement is notices in all the models after the feature reduction/ selection as shown below:

Model Performance (with PCA):



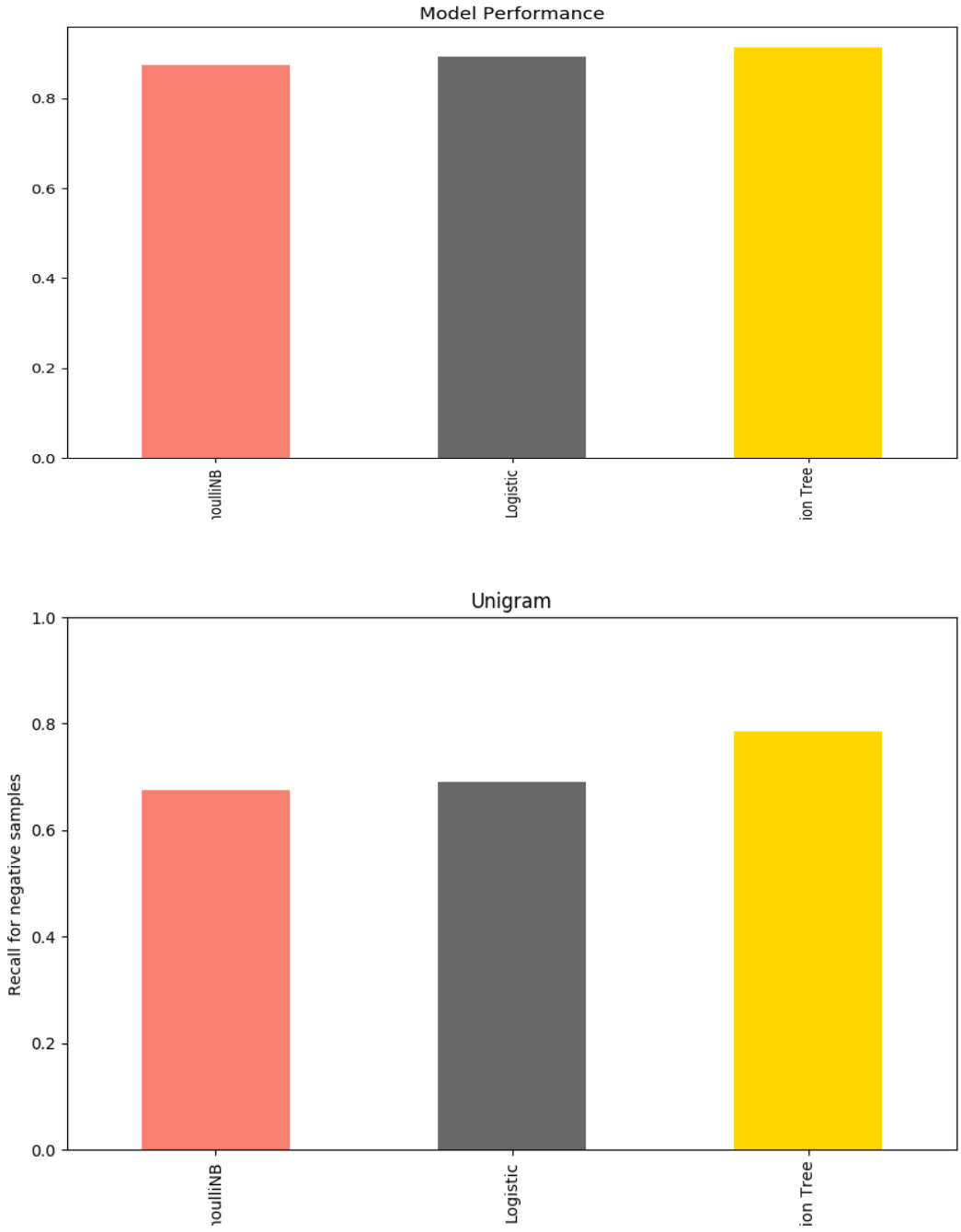
Model Performance (with most frequent features):



5.2.3. Improvement in classification by considering sequence of words technique:

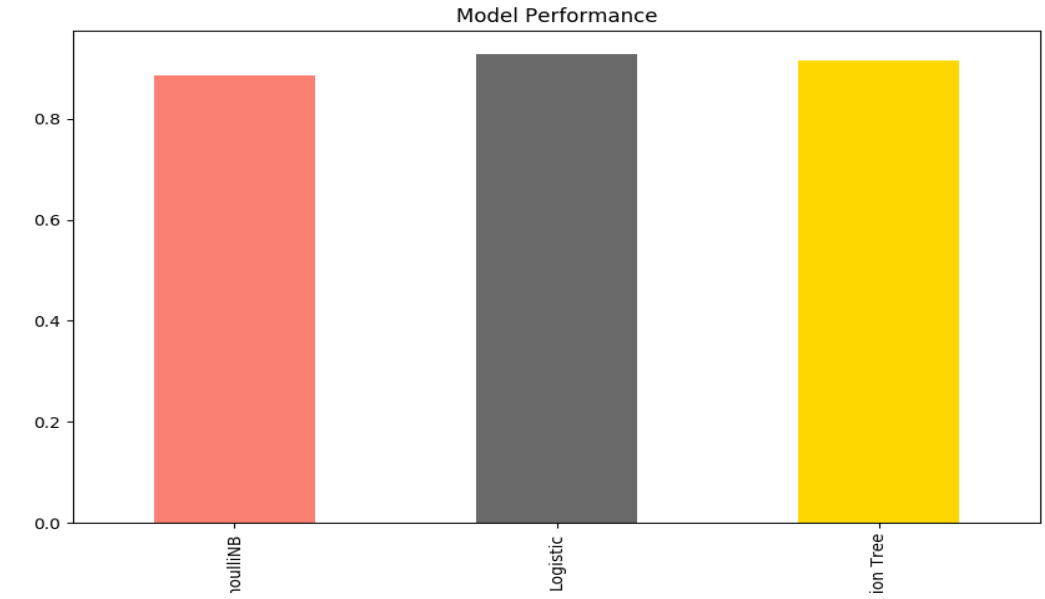
We considered the models which gives the higher performance by using the sequence of words technique.

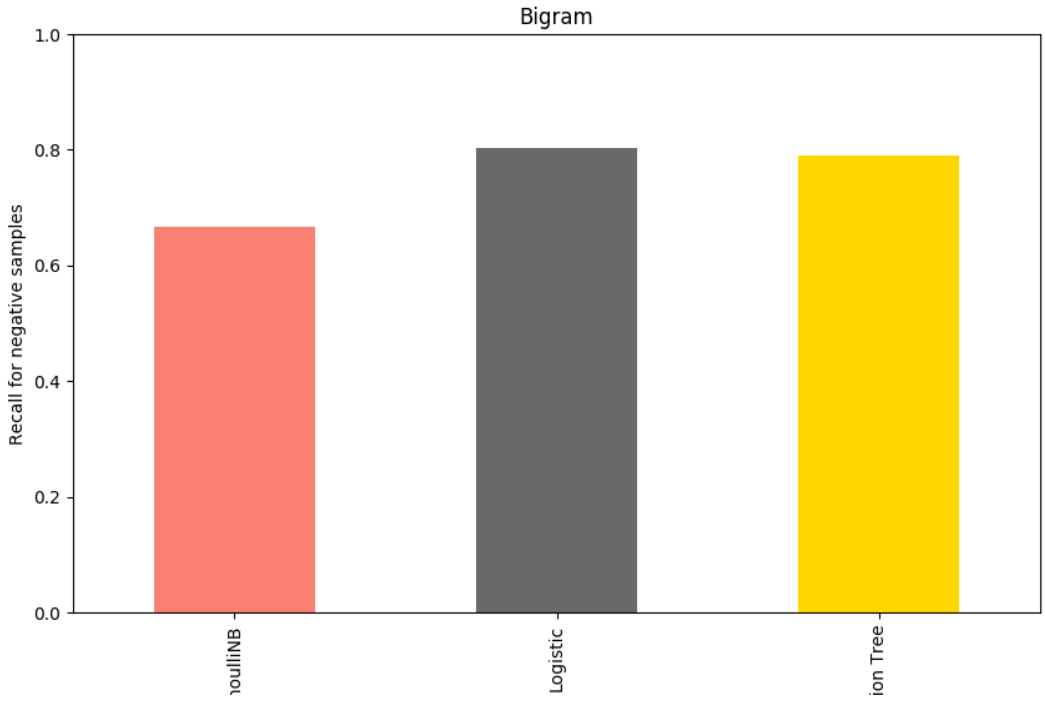
Firstly for Unigram below is a visual comparison of recall values for negative samples accuracy:



Secondly, Now a significant improvement on the recall of negative instances is noticed which might refer to the number of reviewers that would have used 2 word phrases sequence like “not good” or “not great” to imply a negative review.

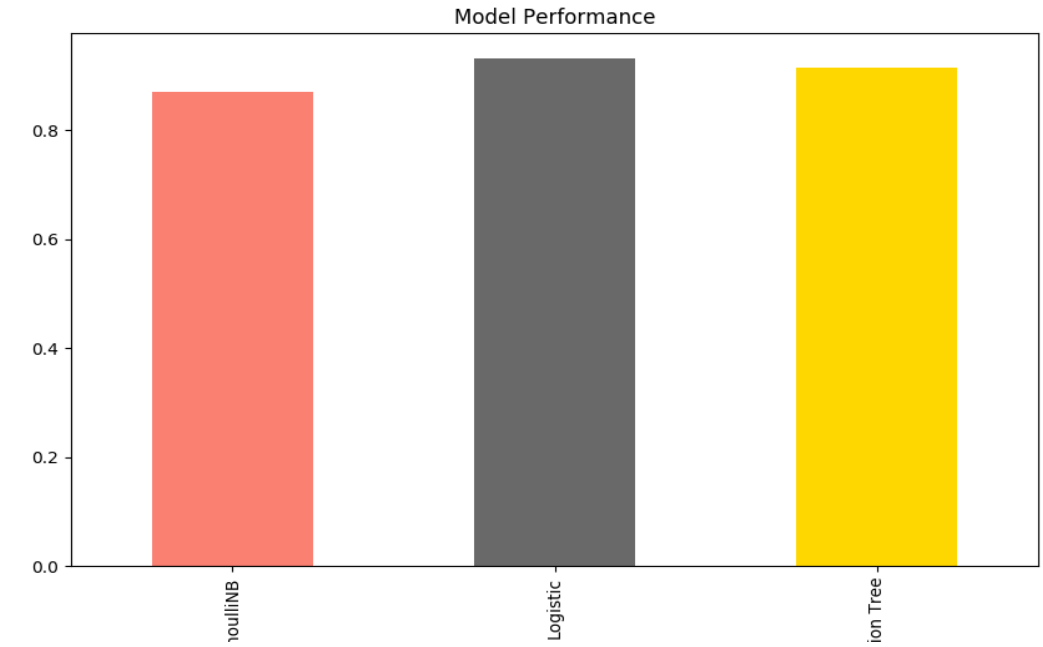
Now for Bigram below is the visual representation of recall values of negative samples accuracy:

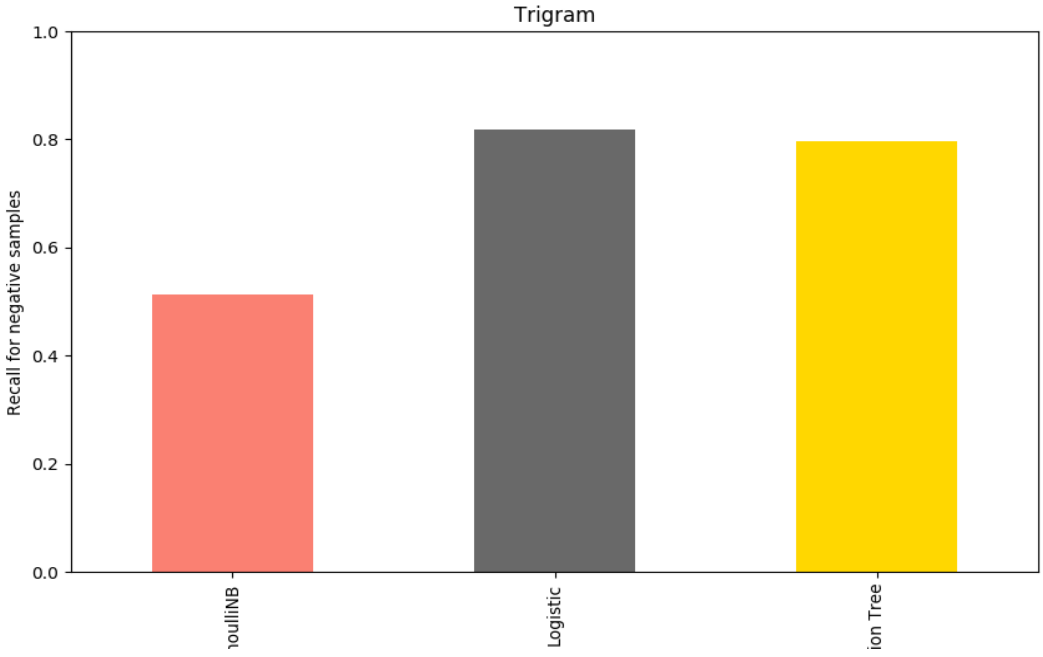




Thirdly, Trigrams give the best accuracy results in all the 3 sequences.

Below is a visual comparison of recall values for negative samples accuracy:

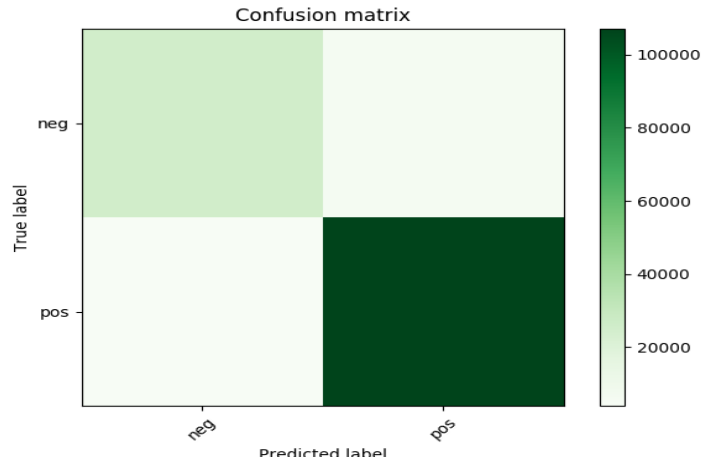


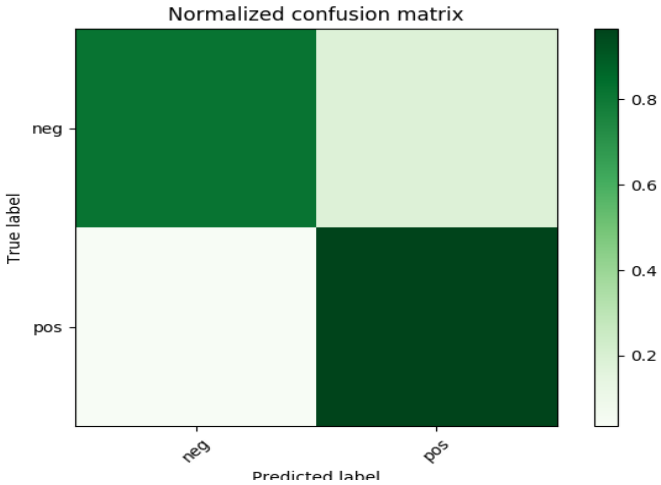


5.2.3.a. Evaluation and Interpretation using confusion matrix

Logistic Regression gives accuracy as high as 93.2%. Hence, recall/precision values for negative samples are higher than ever in this case. As the logistic regression gives the best model performance accuracy we can further analyze this with a use of confusion matrix which is being plotted between the true and predicted labels.

Now from the matrix below we can easily make out that a large number of samples predicted to be positive are actually positive in their actual label. Whereas a few number of samples predicted to be negative are actually negative in their actual label. To get a clear visual of this matrix for better understanding of the performance we will look into the normalized confusion matrix generated. Now with the light color of negative samples proves that the logistic regression predicted negative samples accurately too.





5.2.3.b. Word Cloud representation

To get a better understanding here is a visual representation on word cloud of the most frequent helpful words occurring in the reviews.

6. Conclusions and Future Work

6.1. Conclusions

As a conclusion we can conclude that bag-of-words is a pretty efficient method if one can compromise a little with accuracy. Also for such a large datasets it is advisable to make use of algorithms that can run in linear time (like Naïve Bayes, although they might not give a very high accuracy). Finally, utilization of sequence of words is a good approach when the main goal is to improve accuracy of the prediction model.

6.2. Limitations

- As for getting the most helpful reviews we are working with the most frequently occurring words and group of words but we must also take care of the words which are considered as positive words but are not most occurring and are not considered by us in prediction of helpful review.
- Many of the common words consist of food words and amazon words. For example Coffee, Torreon, etc.

6.3. Potential Improvements or Future Work

- We can make use of POS tagging mechanism to tag words in the training data and extract the important words based on the tags. For sentiment classification adjectives are the critical tags. So we must take care of other tags too which might also have some predictive value.
- We would explore curating a domain-specific dictionary for this project to avoid common food words and Amazon words in reviews.
- Price, Flavor, and Great are some top indicator words for a helpful review. So more use of these words will indicate a possible bias among customers to mark a review as helpful when the review is positive.
- We would explore more using these findings as a guide for reviewers. For example, when writing a review, Amazon could show "Tips for writing a helpful review" something like Describe the flavor of this product ("Flavor" is the most highly correlated parameter with "helpfulness"), etc.