**Final Project Report**

**INFO6105 Data Science Methods and Tools**

**Topic: Amazon Fine Food Review**

**Team Members:**

**Deepank Khurana (001352458)**

**Anindita Baishya (001387422)**

**Sindhura Kolli (001387687)**

**Pratiksha Lavhatre (001388250)**

**INTRODUCTION/OVERVIEW**

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.  
We are using Score/Rating. A rating of 4 or 5 can be considered as a positive review. A rating of 1 or 2 can be considered as a negative one. A review of rating 3 is considered neutral and such reviews are ignored from our analysis. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

**KEYWORDS**

* Amazon
* Food
* Reviews
* Rating
* Algorithms
* Customers
* Analysis

**BACKGROUND**

The development of the internet changed the way people eat and respond to food. Amazon is one of the biggest website where users can easily purchase all kind of food they need with only a mouse-click or with many supportive ways for decision making. Most of the food have the ratings and reviews from other customers who have different backgrounds, cultures, eating habits and personalities. We aim to create a predictive model, which is designed for predicting the rating of different food based on the data given. In order to achieve the goal, we utilize the Amazon Fine Food Reviews dataset, which has some necessary features for us to analyze. We analyze the data using the models, such as KNN algorithm, Naïve Bayes, SVM, all of which we have learned from our course. Eventually we want to build an acceptable model which helps us better understand how customers rate the food they purchase.

**OBJECTIVE**

In this project, we will be applying various machine learning algorithms to learn if a review provided by user is positive or negative based on the rating given by user.

In short, we are precisely trying to analyze on the following points:

1. Is there any trend in review score from 1999 to 2012?
2. Is it possible to predict the polarity of the review (positive/negative) using supervised methods?
3. Understand which algorithm has the highest accuracy score for the given data set.
4. Analyze the speed performance of the algorithms and confusion matrix.

**METHODOLOGY**

In this project we will be applying various machine learning algorithms such as :

**KNN ALGORITHM:**

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output

2. Calculation time

3. Predictive Power

**NAÏVE BAYES:**

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

**LOGISTIC REGRESSION:**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analysis, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

**SPECIFICATIONS/DESCRIPTIONS**

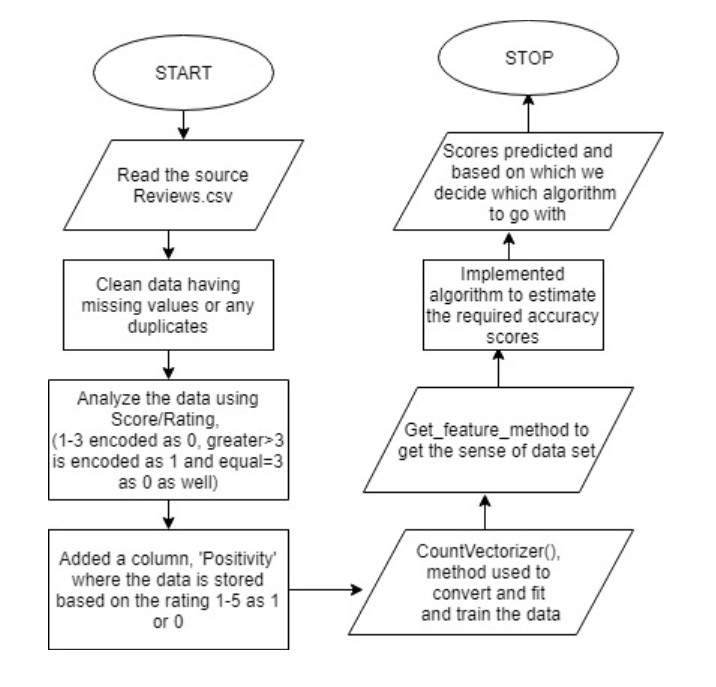
This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

* Number of reviews: 568,454
* Number of users: 256,059
* Number of products: 74,258
* Timespan: 13 Years
* Number of Attributes/Columns in data: 10

Column names and Description:

1. Id
2. ProductId - unique identifier for the product
3. UserId - unqiue identifier for the user
4. ProfileName
5. HelpfulnessNumerator - number of users who found the review helpful
6. HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

**DATA PREPROCESSING FLOW**

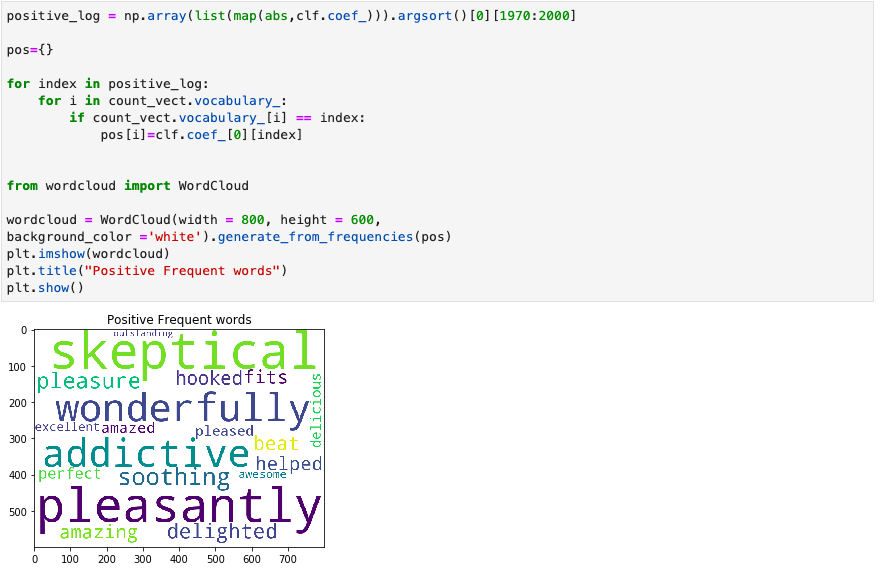
****

**SUMMARY AND OBSERVATIONS OF EACH ALGORITHM**

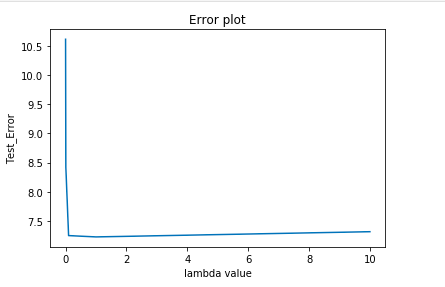
**LOGISTIC REGRESSION:**

* Considered 100k datapoints for running this model
* Used Pretty table to show all the data in a table format so that user can understand what is going on with every model
* Done Perturbation test and found out that features are collinear
* Both BoW and TF-IDF models gave best accuracies compared to AVG W2V and TF-IDF W2V
* Also proved that as C value decreases sparsity also decreases
* Plotted Error plots, Heat maps for confusion matrix, Pretty tables for all models
* The highest accuracy observed for this model is 92.35%

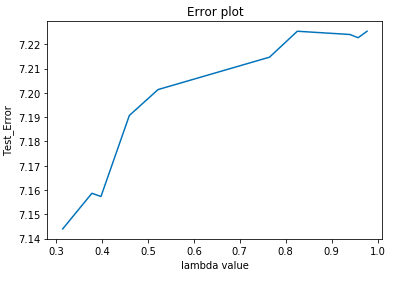
Word Cloud :



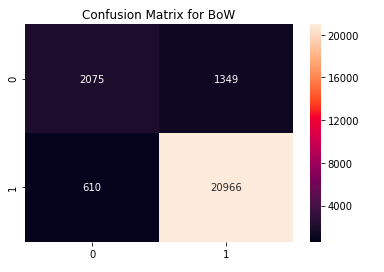
Error Plots using Grid Search :



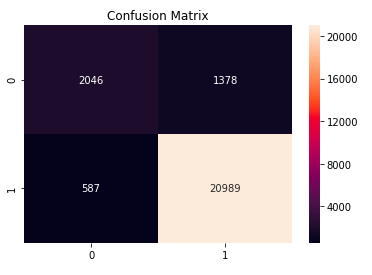
Error Plots using Randomised Search:



Confusion matrix for GridSearchCV:

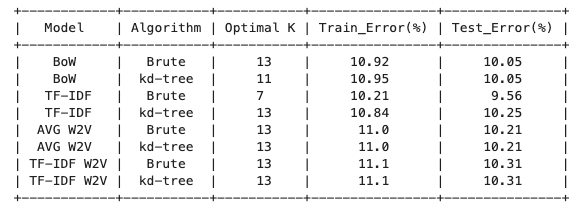


Confusion matrix for Randomised Search:



**KNN ALGORITHM:**

* Considered a sample of 30k datapoints. (Box Running times)
* Plotted Error plots based on K values
* Shown Confusion Matrix using sns heatmap, majorly we observed most of the points are classified as positive
* Used PrettyTable to show all the data as a table manner
* From the table we can see that TF-IDF on brute force gave less test error compared to other models
* Interestingly noted that, on AVG W2V and TF-IDF W2V both algorithms gave exactly same results
* Highest accuracy observed for this model is 88.9%



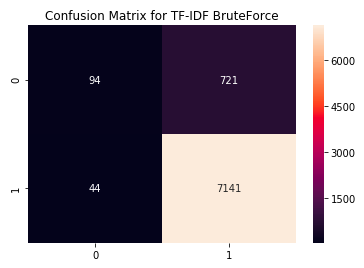
Error Plot using Brute Force:

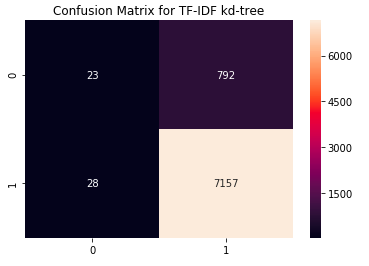


Error Plot using KD Tree :



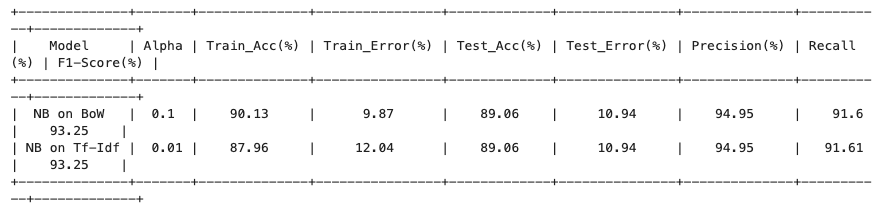
Confusion Matrix:





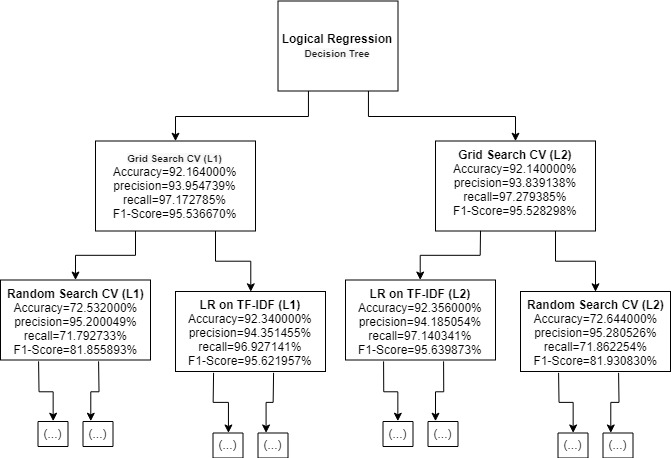
**NAÏVE BAYES:**

* Considered 250k data used TIME BASED SPLITIING
* Time based splitting done with the ratio as 70:30 = train:test
* In BoW model , got alpha= 0.1 and in TF-IDF model alpha = 0.01.
* Used PrettyTable to make a table of all metrics for both models
* But if we see both positive and negative words that impacts our model, both seems semantically similar, because those are the words that repeats the most in both the classes
* Highest accuracy observed for this model is 89.06%



**Decision Tree Model:**

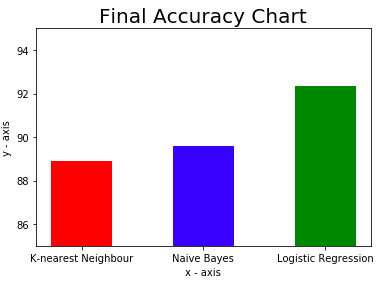
Below is the decision tree for logistic regression. We followed the similar approach for other two algorithms as well.



**CONCLUSION:**

Among the three algorithms, Logistic regression has the highest accuracy as it the best algorithm to conduct predictive analysis when the dependent variable is binary.

Below is the graph showing the accuracies of the three algorithms:



**DATA SOURCE**

* <https://www.kaggle.com/snap/amazon-fine-food-reviews>

**REFERENCES**

* <https://towardsdatascience.com/logistic-regression-python-7c451928efee>
* <https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/>
* <https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>