PROJECT REPORT ON MACHINE LEARNING

TOPIC::::AMAZON FINE FOOD REVIEW ANALYSIS

TAPAS KUMAR PANIGRAHI
M.Sc. in Big Data Analytics, Department of Computer Science
Ramakrishna Mission Vivekananda Educational And Research Institute
Belur Math, Howrah
Pin-711202, West Bengal

June 27, 2020

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my faculty guide Dr./Br. Dripta Mj. as well as our H.O.D Swathy Prabhu Mj. who gave me the golden opportunity to do this wonderful project on the topic AMAZON FINE FOOD REVIEW ANALYSIS, which also helped me in doing a lot of Research and i came to know about so many new things I am really thankful to them.Secondly i would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

TAPAS KUMAR PANIGRAHI
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, West Bengal
18 JULY, 2020

Contents

1 PROJECT BACKGROUND

Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.

Amazon is the largest Internet company by revenue in the world. Amazon was founded by **Jeff Bezos** in **Bellevue, Washington**, on **July 5, 1994**. The company started as an online marketplace for books but expanded to sell electronics, software, video games, apparel, furniture, **food**, toys, and jewelry.This project report, **AMAZON FINE FOOD REVIEWS ANALYSIS**,is based on model created to predict about the food product quality based on the reviews given by customer.

Suppose a company wants to introduce a new food product in **Amazon** website or a customer want to buy food product from **Amaozn** website,they will focus on the reviews,which contain what previous customers like or dislike.

Thus, this project report basically includes how to classify a review to be positive(**i.e** good review) or not(**i.e** bad review using **Machine learning** algorithms and **NLP**(natural languagae processing) algorithm.

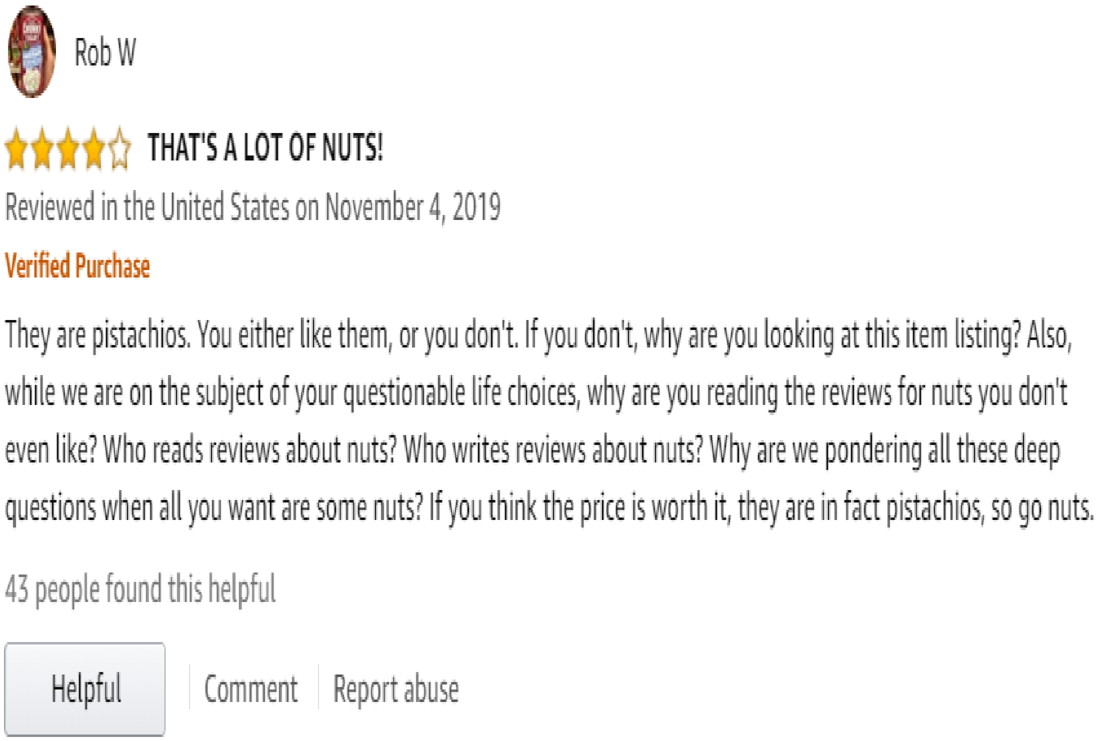
2 INTRODUCTION

With the upcoming surge of purchasing everything online, more and more people are prepared to shop for food online since they are provided with all the convenience.

In this project,i have worked on **AMAZON FINE FOOD REVIEWS** dataset.i worked on text data.So at first,i convert the text data into vector to use machine learning model.To convert into vector,i perform **BOW**(Bag of words),**bigram**, **tfidf**(term frequency–inverse document frequency).And then i apply **Logistic regression** along with some visualistaion and **TSNE**.

3 PROJECT DATASET

The Dataset that I have used here for my Project is called **reviews.csv**. 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



Rob W

★★★★☆ **THAT'S A LOT OF NUTS!**

Reviewed in the United States on November 4, 2019

Verified Purchase

They are pistachios. You either like them, or you don't. If you don't, why are you looking at this item listing? Also, while we are on the subject of your questionable life choices, why are you reading the reviews for nuts you don't even like? Who reads reviews about nuts? Who writes reviews about nuts? Why are we pondering all these deep questions when all you want are some nuts? If you think the price is worth it, they are in fact pistachios, so go nuts.

43 people found this helpful

[This is a **REVIEW** page from **AMAZON** website.]

Data includes:

1. Reviews from Oct 1999 - Oct 2012
2. 568,454 reviews
3. 256,059 users
4. 74,258 products
5. 260 users with ¿ 50 reviews

3.1 Dataset Description

The dataset has fooloeing attributes:

1. PRODUCT ID – Unique identifier for the product
2. USER ID – Unique identifier for the user
3. PROFILE NAME – Profile name of the user
4. HELPFULNESS NUMERATOR – Number of users who found the review helpful
5. HELPFULNESSDENOMINATOR – Number of users who indicated ID - Row id whether they found the review helpful or not
6. SCORE – Rating between 1 and 5
7. TIME – Timestamp for the review
8. SUMMARY – Summary of the review
9. TEXT – Text of the review

4 EXPLORATORY DTAA ANALYSIS

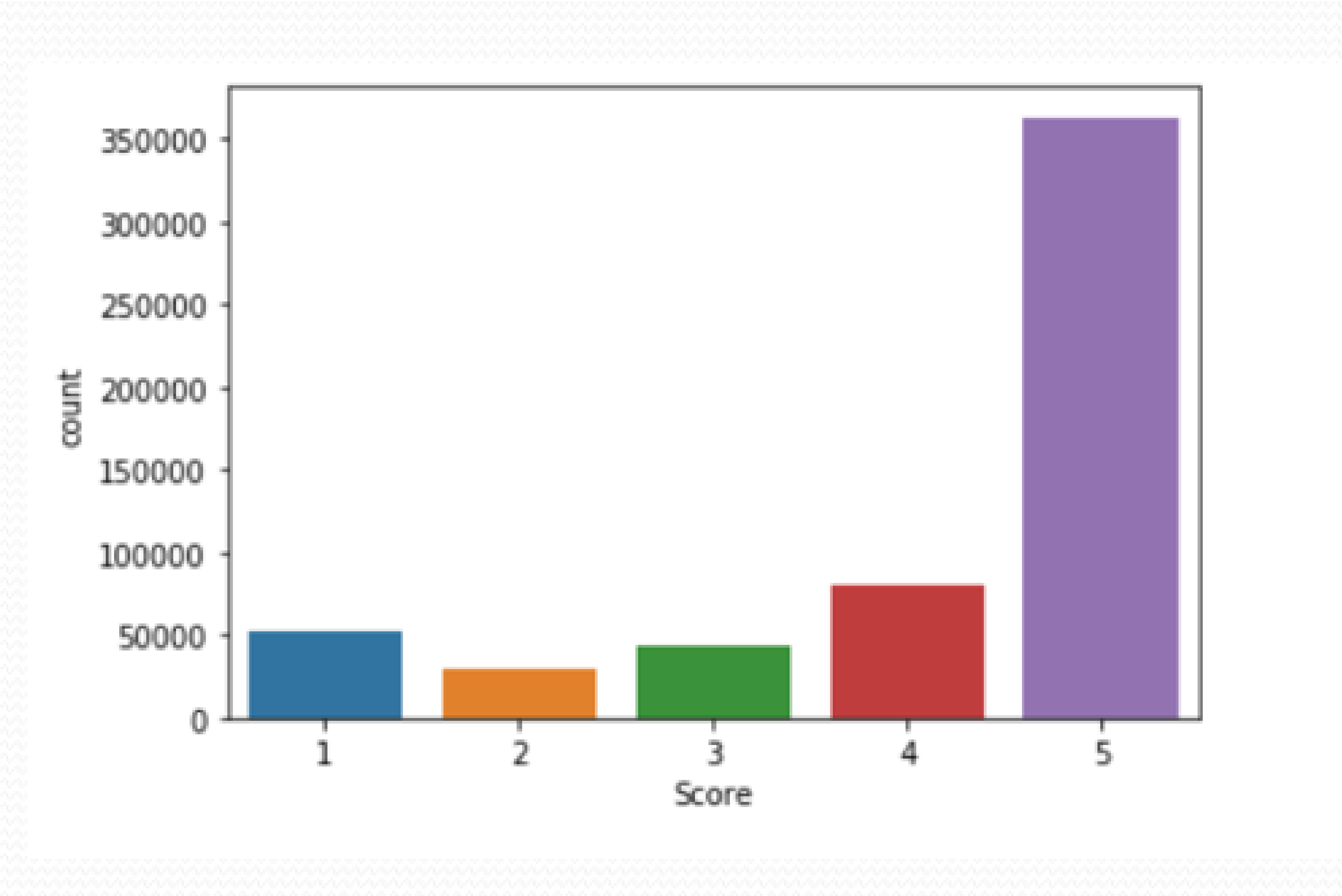


Figure 1: Number of reviews of different score

From the above figure:

1. Number of reviews having score (1):: 50000
2. Number of reviews having score (2):: 25000
3. Number of reviews having score (3):: 40000
4. Number of reviews having score (4):: 65000
5. Number of reviews having score (5):: 350000

Now i create a column positivity,where reviews score ¿3 i take it as 1 (**i.e** Positive review) and where reviews score ¡3 i take it as 0 (**i.e** Negative review and i ignore the review having score 3 .

5 TSNE representation

T-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm for visualization developed by **Laurens van der Maaten** and **Geoffrey Hinton**.
It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

5.1 TSNE representation with BAG OF WORDS

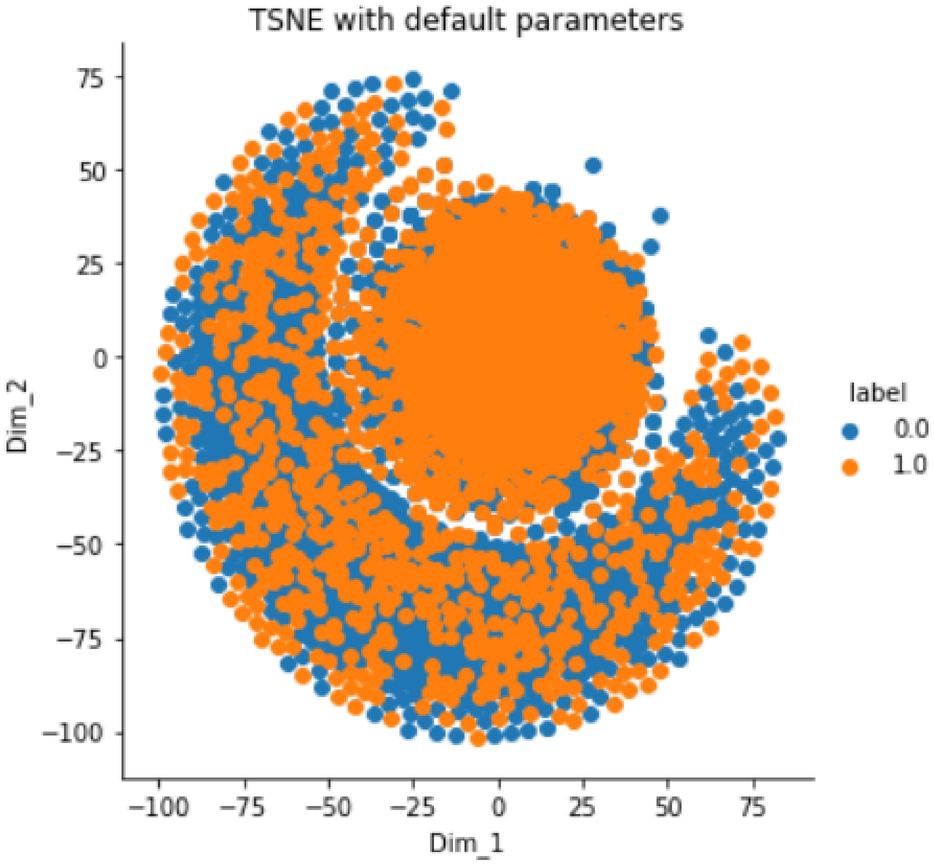


Figure 2: TSNE REPRESENTATION WITH BAG OF WORDS

5.2 TSNE representation with TFIDF

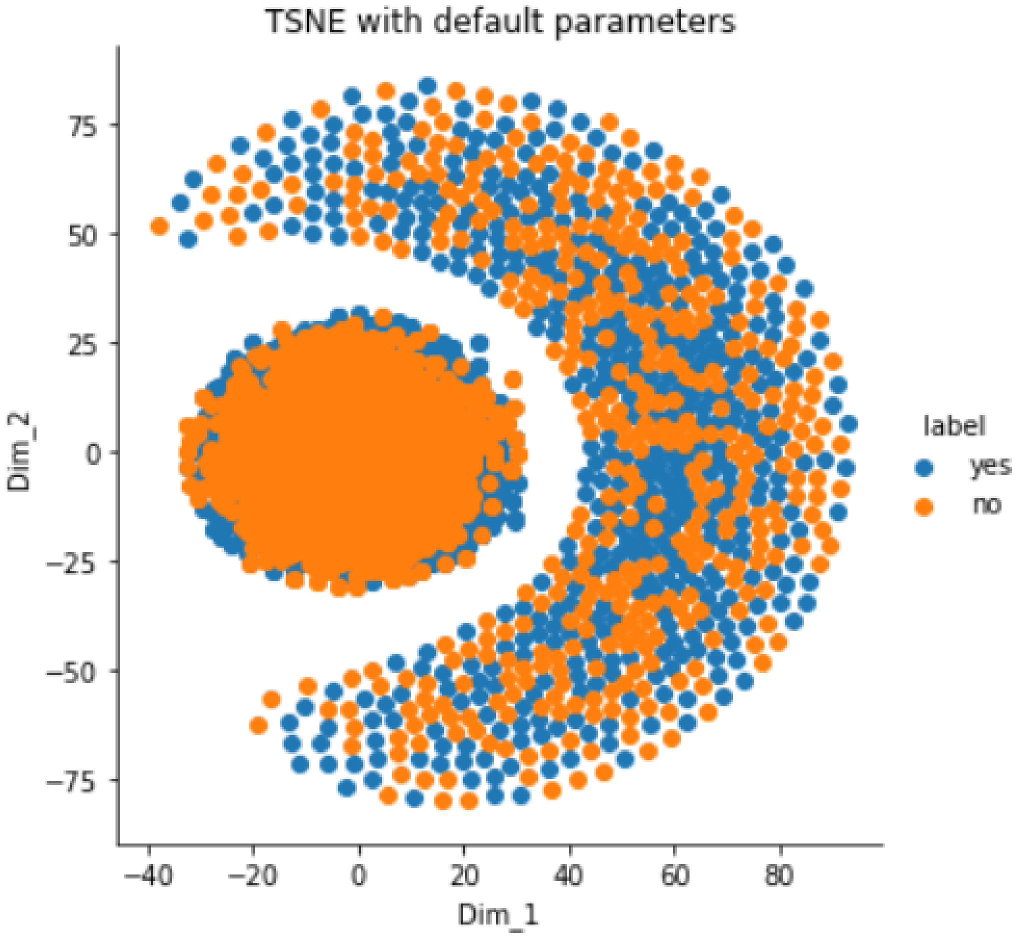


Figure 3: TSNE REPRESENTATION WITH TFIDF

6 Text to vector conversion

To analyze text and run algorithms on it, we need to represent the text as a vector. The notion of embedding simply means that we'll convert the input text into a set of numerical vectors that can be used into algorithms.

6.1 BOW(BAG OF WORDS)

When we use **Bag-Of-Words** approaches, we apply a simple word embedding technique. Technically speaking, we take our whole corpus that has been preprocessed, and create a giant matrix :

1. The columns correspond to all the vocabulary that has ever been used with all the documents we have at our disposal
2. The lines correspond to each of the document
3. The value at each position corresponds to the number of occurrence of a given token within a given document

6.2 BIGRAM

An n-gram is a contiguous sequence of n items from a given sequence of text. Given a sentence, s, we can construct a list of n-grams from s by finding pairs of words that occur next to each other. For example, given the sentence "I am Sam" you can construct bigrams (n-grams of length 2) by finding consecutive pairs of words.

When n =2 it is called bigram.

6.3 TFIDF

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

7 LOGISTIC REGRESSION

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

It uses a black box function to understand the relation between the categorical dependent variable and the independent variables.

The dependent variable is the target class variable we are going to predict.

8 PROJECT RESULTS

Table 1: RESULTS AFTER APPLYING LOGISTIC REGRESSION

LOGISTIC REGRESSION	MODEL NAME	ACCURACY
1	BAG OF WORDS	0.819587797615879
2	BIGRAM	0.8991346890822678
3	TFIDF	0.8338351323376996

9 CONCLUSION

From the used models, it is seen that **Logistic Regression** performs the best when used with **BIGRAM**, but it increases the dimension.so i conclude that,textbfLogistic Regression performs the best when used with **TFIDF**.

10 REFERENCE

1. **Google** ’www.google.com’
2. **Quora** ’www.Quora.com’
3. **Wikipedia** ’www.wikipedia.org’