USC Viterbi
School of Engineering

SENTIMENT ANALYSIS ON AMAZON REVIEWS.

- Raj Thaker.

rpthaker@usc.edu

CSCI 561: FOUNDATIONS OF ARTIFICIAL INTELLIGENCE.
HOMEWORK – 2

DATE : 07/26/2018

➢ Dataset

- The given dataset is an extract from the Amazon Reviews Kaggle competition.
- The dataset of 400,000 reviews is in the form of CSV file delimited by '|'.
- Data contains 2 fields – Classification of Review - (Positive/Negative) and corresponding Review.

➢ Goal

- The goal is to perform sentiment analysis to determine whether a review is positive or negative.

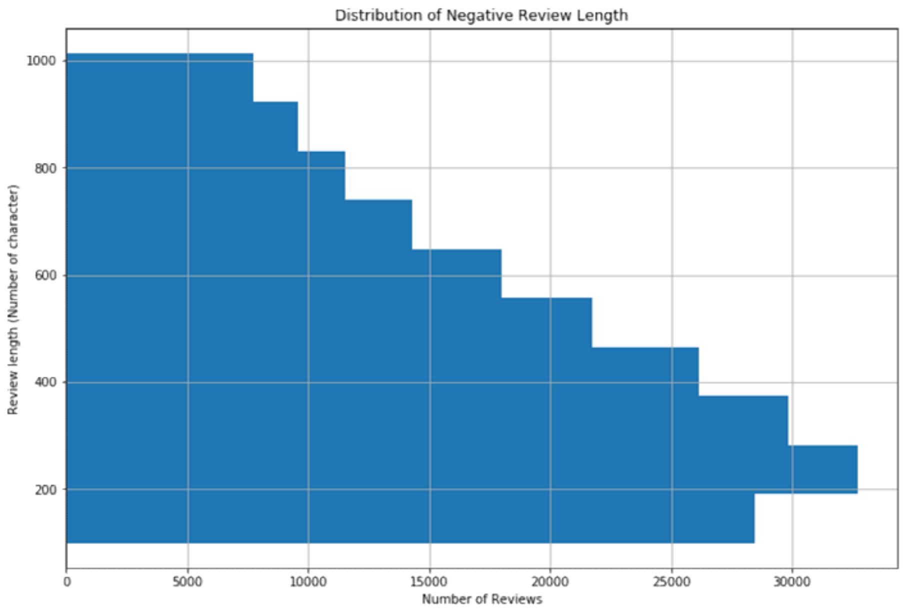
Reference for Kaggle Competition - https://www.kaggle.com/bittlingmayer/amazonreviews

➢ Exploratory Data Analysis.

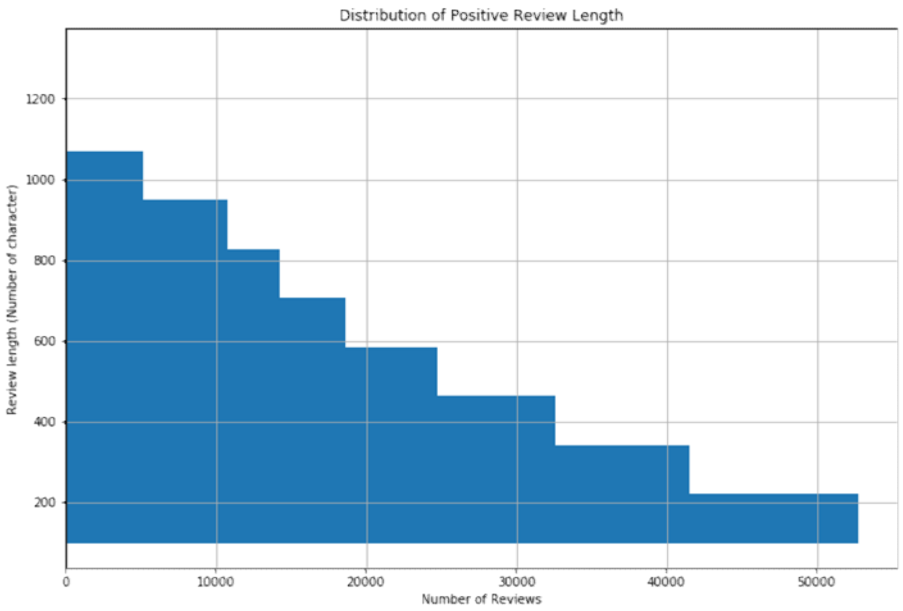
- The given data contains ***equal number*** of positive and negative reviews.

- Reviews Length Distribution:

Length Distribution of Negative Reviews.



Length Distribution of Positive Reviews.

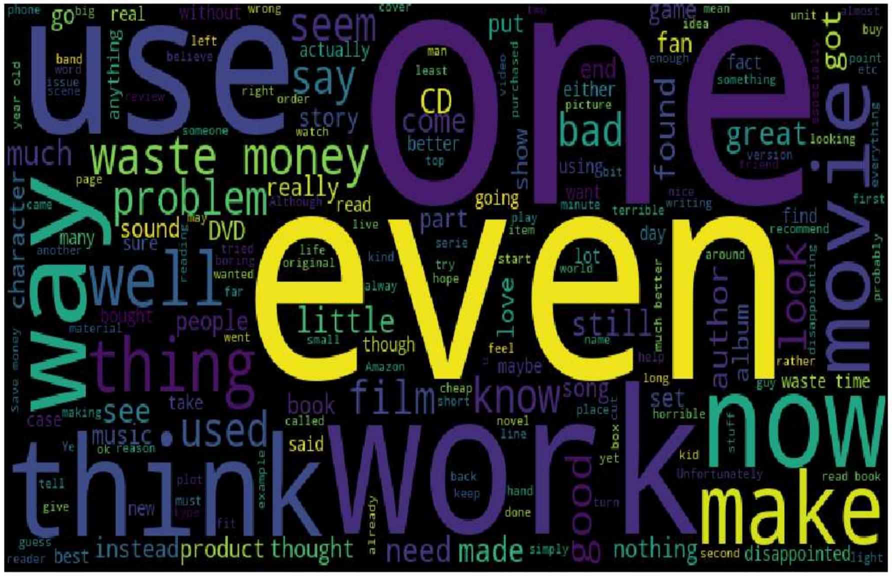


➢ Exploratory Data Analysis.

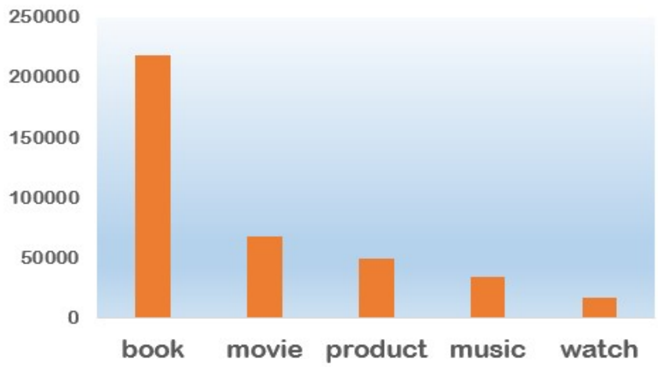
Positive Words Cloud.

love
one
use
even
work
make
think
used
now
time
got
well
good
little
need
much
know
way
music
show
found
come
still
film
play
song
look
say
set
problem
excellent
nice
made
author
life
great
movie
better
album
sound
fan
character
people
world
find
best
read
book
put
see
seem
perfect
many
story
really
lot
new

Negative Words Cloud



- Most reviews spoken about:



➢ Data Preprocessing.

- Removal of Punctuations, Stop words, Stemming.
- Count Vectorizer.

➢ How data was setup for training the model.

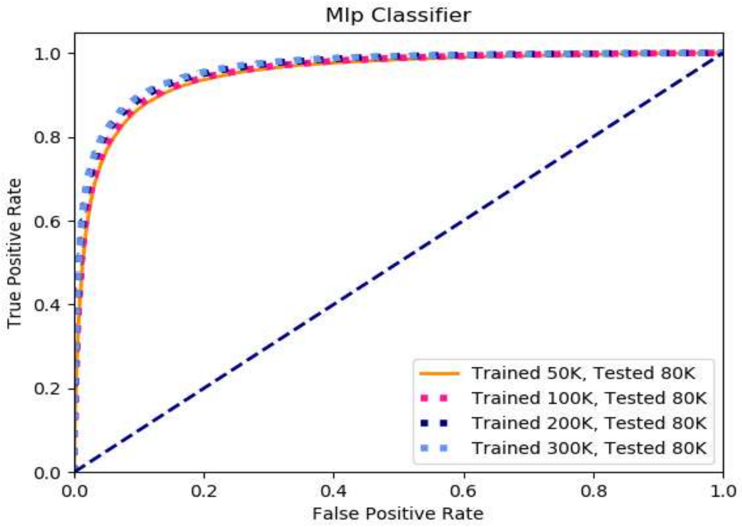
- To understand different ML Classification algorithms, the given dataset was divided in following ways –
  I. 50K records were trained | 80K records tested.
  II. 100K records were trained | 80K records tested.
  III. 200K records were trained | 80K records tested.
  IV. 310K records were trained | 80K records tested.

➢ 3 Machine Learning Algorithms Used:

1. Neural Network – MLPClassifier.
2. Decision Tree
3. Logistic Regression

1. Neural Network – MLPClassifier.

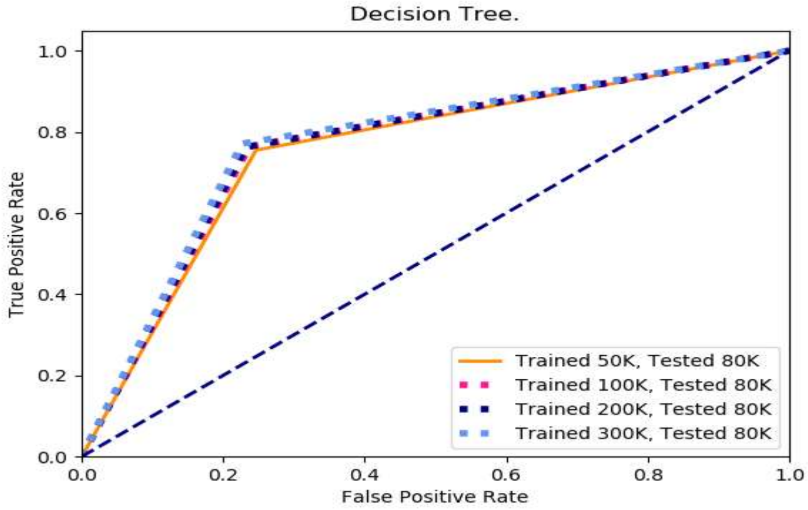
	50K Train | 80K Test	100K Train | 80K Test	200K Train | 80K Test	310K Train | 80K Test
Accuracy	88.40	88.85	89.49	89.85
Precision	0.87	0.88	0.89	0.89
Recall	0.89	0.89	0.88	0.90



➢ 3 Machine Learning Algorithms Used:

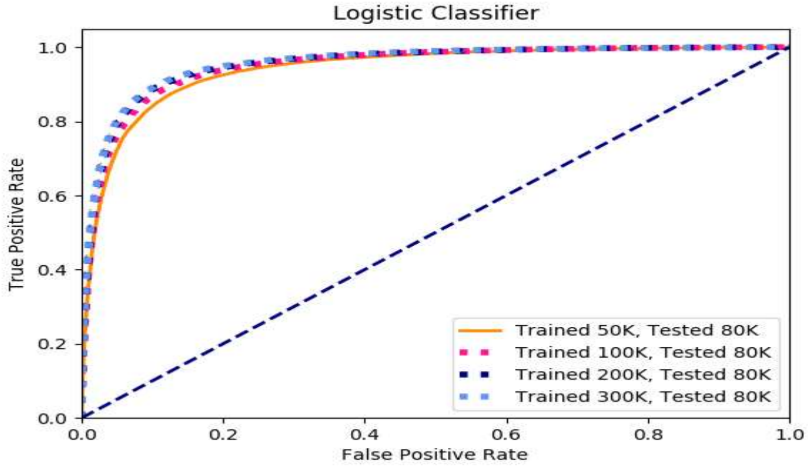
2. Decision Tree

	50K Train | 80K Test	100K Train | 80K Test	200K Train | 80K Test	310K Train | 80K Test
Accuracy	75.41	76.33	76.51	77.19
Precision	0.75	0.76	0.76	0.77
Recall	0.75	0.76	0.76	0.76

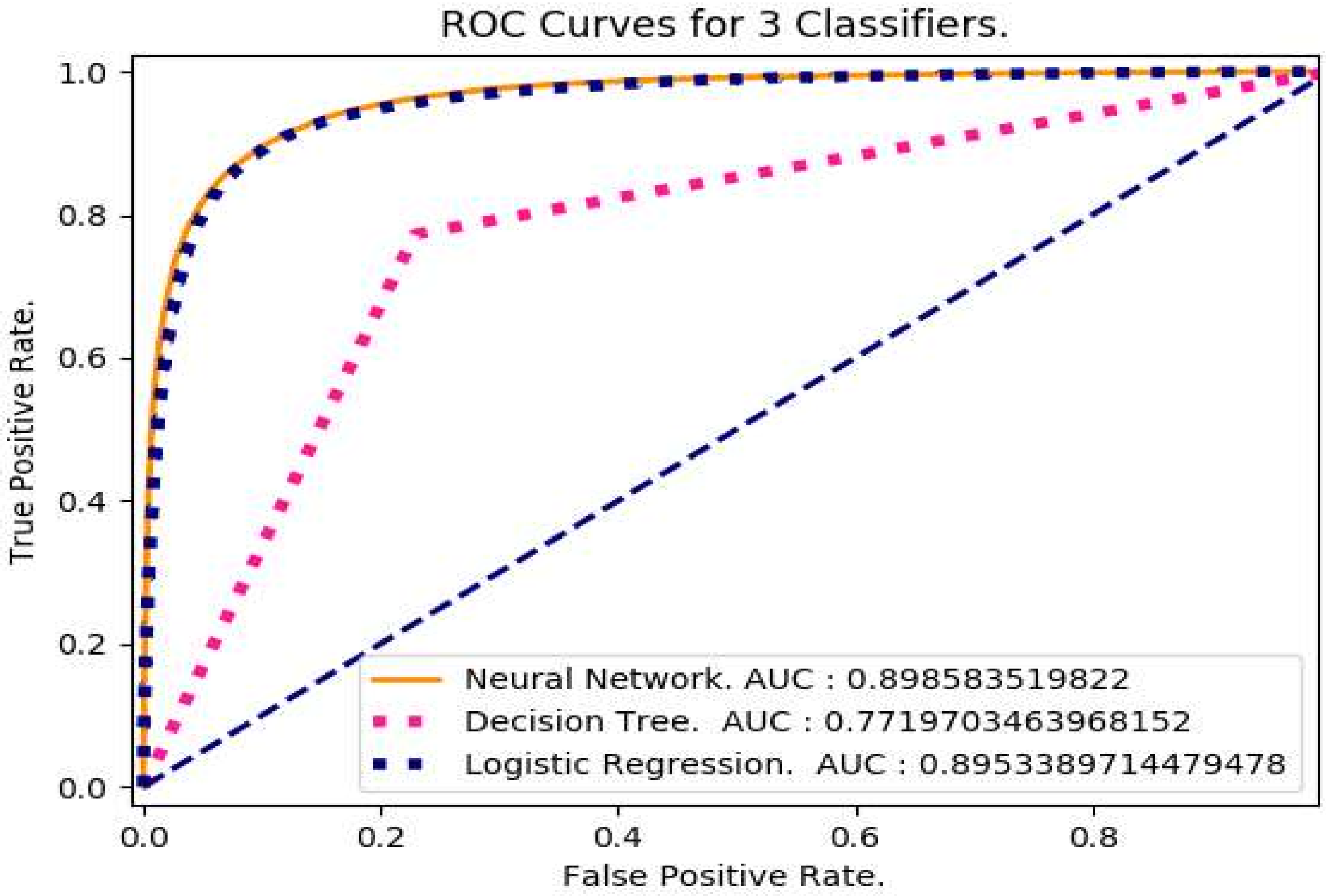


3. Logistic Regression

	50K Train | 80K Test	100K Train | 80K Test	200K Train | 80K Test	310K Train | 80K Test
Accuracy	87.40	88.52	89.23	77.19
Precision	0.87	0.88	0.89	0.89
Recall	0.87	0.88	0.88	0.89



➢ Final Verdict – Neural Networks!!


ROC Curves for 3 Classifiers.
True Positive Rate.
False Positive Rate.
Neural Network. AUC : 0.898583519822
Decision Tree. AUC : 0.7719703463968152
Logistic Regression. AUC : 0.8953389714479478