**Amazon Cell Phone Reviews**

BANA 7047 – Data Mining II

Group 3

By

Abhijith Antony (M13433446)

Hridhay Mehta (M13474723)

Sanjay Jayakumar (M13419548)

Vishnu Vijayakumar (M13255870)

MS, University of Cincinnati

# Abstract

< Can be put in the first page only >

**Contents**

[Abstract 2](#_heading=h.gjdgxs)

[Chapter 1 Introduction 4](#_heading=h.1fob9te)

[1.1 Background 4](#_heading=h.3znysh7)

[1.2 Problem Description 4](#_heading=h.2et92p0)

[1.3 Goal of the Thesis 4](#_heading=h.tyjcwt)

[1.4 Executive summary 4](#_heading=h.3dy6vkm)

[Chapter 2 Exploratory Data Analysis 4](#_heading=h.1t3h5sf)

[2.1 Data Gathering 4](#_heading=h.4d34og8)

[2.2 Data Cleaning 4](#_heading=h.2s8eyo1)

[2.3 Visualization 5](#_heading=h.17dp8vu)

[Chapter 3 Sentiment Analysis 6](#_heading=h.26in1rg)

[3.1 Sentiment Analysis Methods 6](#_heading=h.lnxbz9)

[3.2 Interpreting Results 6](#_heading=h.35nkun2)

[Chapter 4 Topic Modeling 7](#_heading=h.1ksv4uv)

[4.1 Topic Modeling Methods 7](#_heading=h.44sinio)

[4.2 Interpreting Results 7](#_heading=h.2jxsxqh)

[Chapter 5 User Rating Prediction 8](#_heading=h.z337ya)

[5.1 User Rating Prediction Methods 8](#_heading=h.3j2qqm3)

[5.2 Model Performance Evaluation 8](#_heading=h.1y810tw)

[Chapter 6 Conclusion 8](#_heading=h.4i7ojhp)

[6.1 Conclusion 8](#_heading=h.2xcytpi)

[6.2 Future work 8](#_heading=h.1ci93xb)

[References 8](#_heading=h.3whwml4)

# Chapter 1 Introduction

## Background

Natural language processing

## 1.2 Problem Description

## 1.3 Goal of the Thesis

## 1.4 Executive summary

# Chapter 2 Exploratory Data Analysis

## 2.1 Data Gathering

In this project we are working with 2 files:

* reviews.csv: This file has around 68K reviews with 8 feature variables for all brands of phones. There are 2 numerical variables rating and helpfulVotes.
* items.csv: Pre-scraped data for 720 phone items from amazon with 10 feature variables. There are 4 numerical variables: rating, totalReviews, price and original price.

We merged both these files with column ‘asin’.

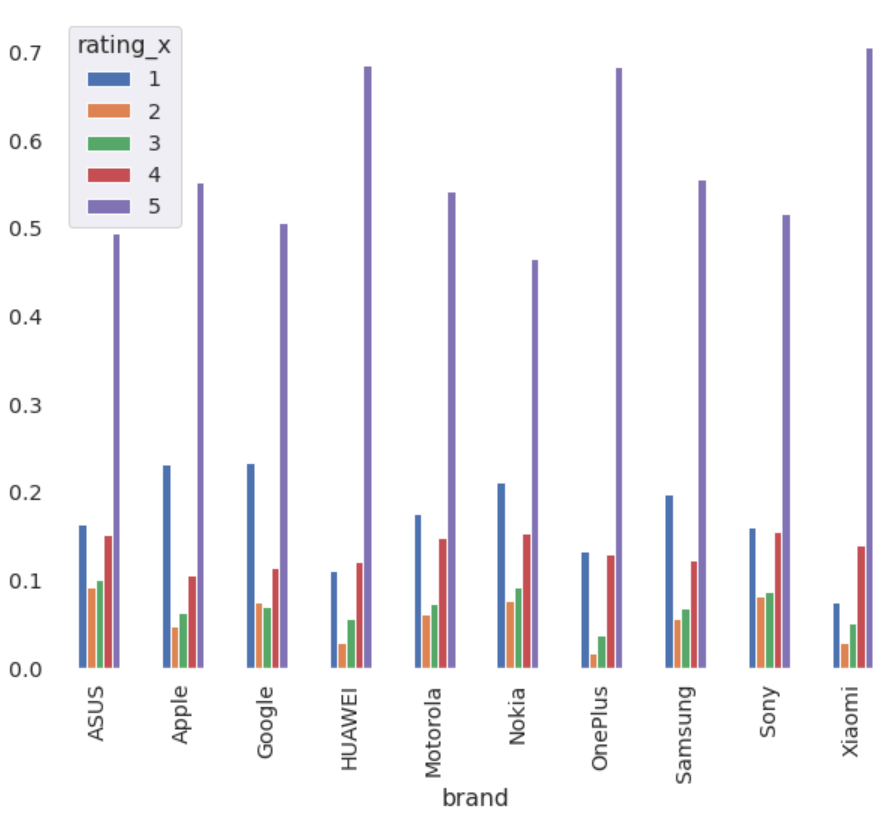
## 2.2 Data Cleaning

It is important that we have the data in clean and standard format before proceeding to further analysis. Data cleaning steps followed in the project:

* First step is to do the basic cleaning processes like converting text into lowercase, removing punctuations. This was done using regular expressions (re) in python.
* Next, we checked for null values in the dataset. For reviews.csv we saw a significant amount of NA’s for column helpfulVotes which was then removed.
* We also removed stop words; these are usually the most used words in English. Removing such words will help us focus on other important words in the text.
* There is a need to tokenize data, which essentially means to split the text into smaller pieces.
* We also performed stemming and lemmatization, which are text normalization techniques. We used the NLTK package in Python to perform the same. Stemming is the process of reducing the words to their root forms, like mapping a group of words to the same step. Although lemmatization ensures that the root word is a valid word in English language unlike stemming.

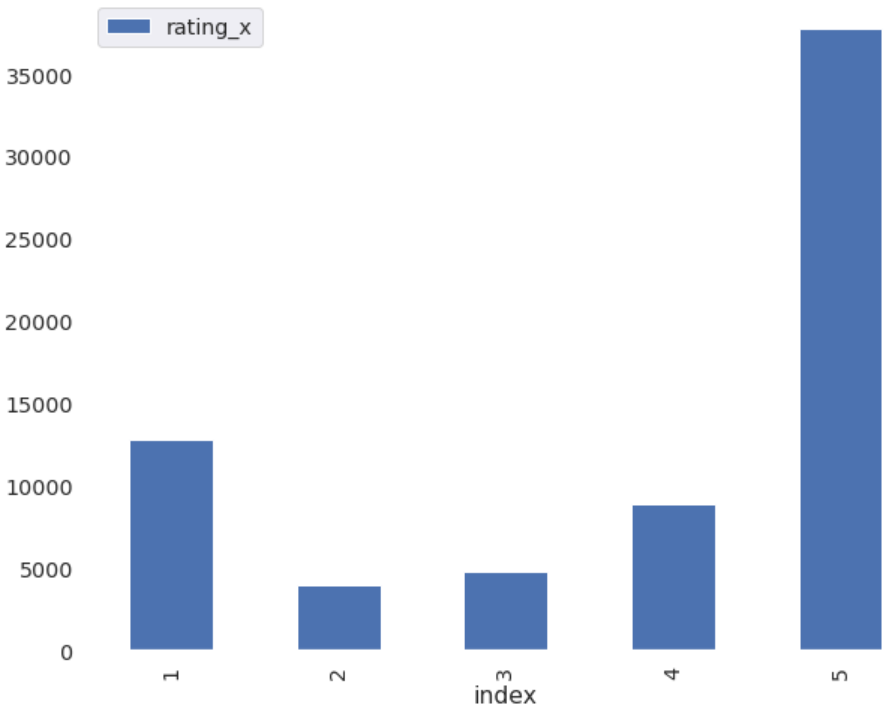
## 2.3 Visualization

Visualization is a very strong medium to understand the data. We generated multiple plots to understand the nature of overall and brand wise user ratings before venturing into Sentiment analysis and topic modeling. *Figure 1* gives a brand wise outlook on the user ratings. The highest deviation in ratings due to missing reviews is for one-plus phones at 0.63. Samsung, Nokia, Motorola and Apple have less than 0.2 rating deviation.



*Figure 1: Average user ratings per Brand*

We can see from *figure 2* that out of all the 68K reviews, more than 70% of them have been rated as either 4 or 5. Hence we see a bi-modal distribution of ratings for all the brands in the figure.



*Figure 2: Overall user ratings*

# Chapter 3 Sentiment Analysis

## 3.1 Sentiment Analysis Methods

Sentiment Analysis is contextual mining of text which identifies and extracts subjective information in source material and helps a business understand the social sentiment of a brand. Here we are dealing with user reviews of multiple brands and hence this technique is very essential to understand the customer sentiment.

Input to a Sentiment analysis technique is a corpus, which is basically a collection of words where order matters.

We are using the TextBlob part of NLTK library in python, and the Vader Analyzer part of vaderSentiment library for applying sentiment analysis techniques.

The output of sentiment analysis is a sentiment score ranging from -1 to 1 (which is polarity) indicating how positive or negative they are and a subjectivity score of 0 to 1 where 0 indicates a fact and 1 indicates an opinion.

## 3.2 Interpreting Results

# Chapter 4 Topic Modeling

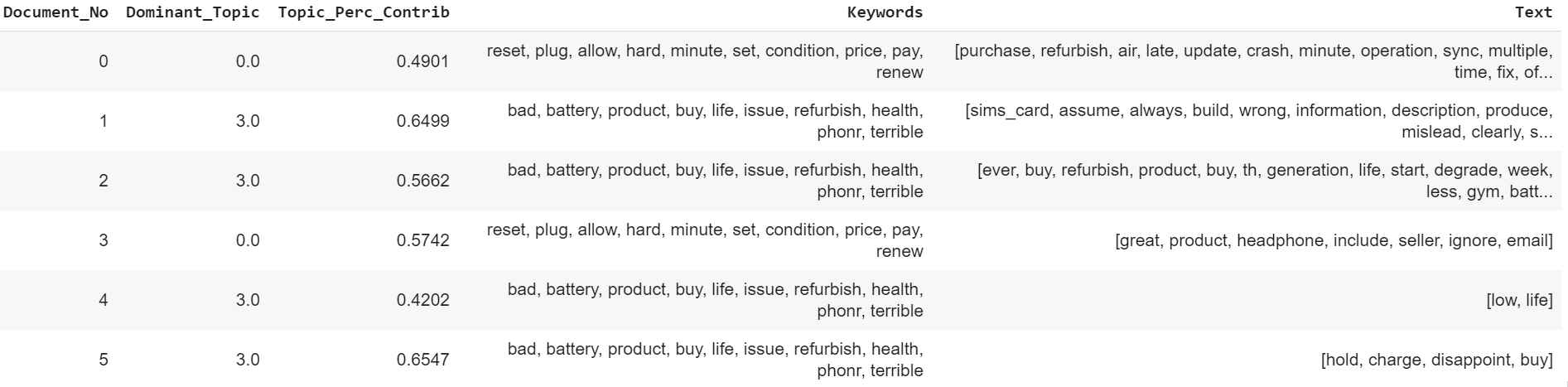
## 4.1 Topic Modeling Methods

Topic modeling is a technique used to extract meaningful information from vast amounts of data. In this project we are trying to label different topics among all the cell phone reviews in order to provide business recommendations to sellers.

Input to the topic modelling technique is a document-term Matrix: It is a matrix where the rows are different documents and columns are different terms and values in the matrix are the word counts. Each topic will consist of a bag of words not necessarily ordered.

We are going to implement topic modeling by using a python library called Gensim. This package utilizes a topic modeling technique called Latent Dirichlet Allocation (LDA). This technique aims to find the hidden probability distributions where every document is a probability distribution of topics and every topic is a distribution of words. We give the document-term matrix, number of topics and number of iterations as input to the Gensim LDA process. Gensim will go through the process of finding the best word distribution for each topic and best topic distribution for each document.

Output of topic modeling will be top words in each topic highlighting themes across various user reviews. From the output we can interpret results and see if the bad words in each topic make sense.



*Figure 3: Formatted LDA output from Gensim package*

## 4.2 Interpreting Results

# Chapter 5 User Rating Prediction

## 5.1 User Rating Prediction Methods

The ability to predict customer sentiment based on their raw text reviews are of increasing significance. With increasing quantity of text data generated applying text mining techniques to id

## 5.2 Model Performance Evaluation

# Chapter 6 Conclusion

## 6.1 Conclusion

## 6.2 Future work

# References

<https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>

<https://towardsdatascience.com/review-rating-prediction-a-combined-approach-538c617c495c>