**K. L. E. SOCIETY’S**

**B. V. BHOOMARADDI COLLEGE OF ENGINEERING & TECHNOLOGY, HUBLI – 580031**

**(An Autonomous Institution)**



**DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

**2016 – 2017**

BDA Report on

Analysis of Amazon Food Reviews

**Under the guidance of**

**Ms. Sunita Hiremath**

Submitted by

**Mr. V T Vishwanath 2BV14IS114**

**Mr. Raghavendra P 2BV14IS131**

**Mr. Vijay Kumar 2BV14IS118**

**Mr. Surendar Singh 2BV14IS112**

**TABLE OF CONTENTS**

1. About ………………………………….……….………………………….. 3
2. Motivation …………………………………………………...……………. 3
3. Problem Statement ………………………………………………………... 3
4. Objectives ………………………….…………………………………….... 3
5. Description of the data ….…………...……………………………………. 3
6. Dataset ……………………………….……………………………………. 4
7. Data Statistics ……………………………………………………………... 4
8. Implementation and analysis …………………………………………….... 5

**ABOUT**

The dataset is taken from [www.kaggle.org](http://www.kaggle.org/). The dataset is provided by the Amazon Company and all the products are food items. The dataset contains the details of the reviews about the products over time. The dataset contains ProductId for which review details are given; and the same product can have reviews over time**.**

**MOTIVATION**

Amazon fine food dataset contains the review of the food items sold by amazon. The buyers review food items. It is very important as it will indicate the user’s reaction on each food item. User will give rating and write the summary and detailed review. This review on food item is important as it will indicate the food item is good or bad. From packaging to quality, quantity everything is important. So indirectly these reviews will increase/decrease the marketing shares and number of food items sold on amazon. Now a days text mining is trending and it will indicate the user’s reaction by giving it as positive, negative or neutral. This motivates us to analyze the each review and give the sentimental result. And to do the time series analysis which will shows the current rating of particular product. This will help amazon to sell that particular product and this will also increase the user’s experience.

**PROBLEM STATEMENT**

To analyse the review for each food item and provide the status of it in the market over time.

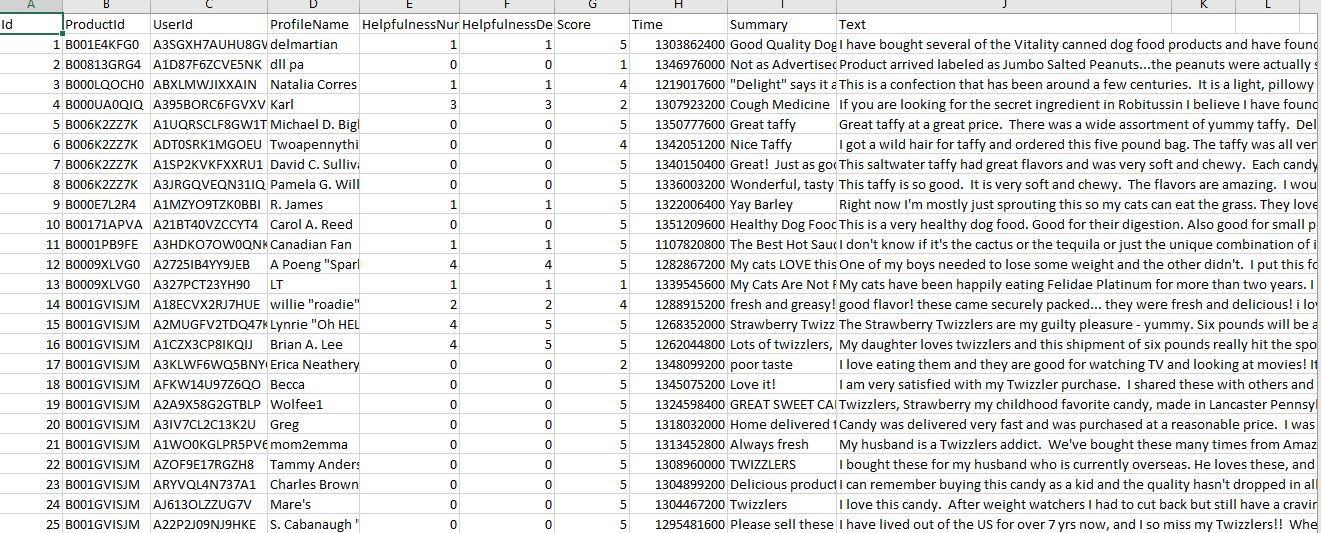
**OBJECTIVES**

* Providing average sentiment for each product based on the text present in each review.
* Taking most helpful review from each product.
* Providing the details of the scores for each food item (Saying how many 1 star, 2 star reviews till 5 stars).
* Providing the time series analysis for the product.

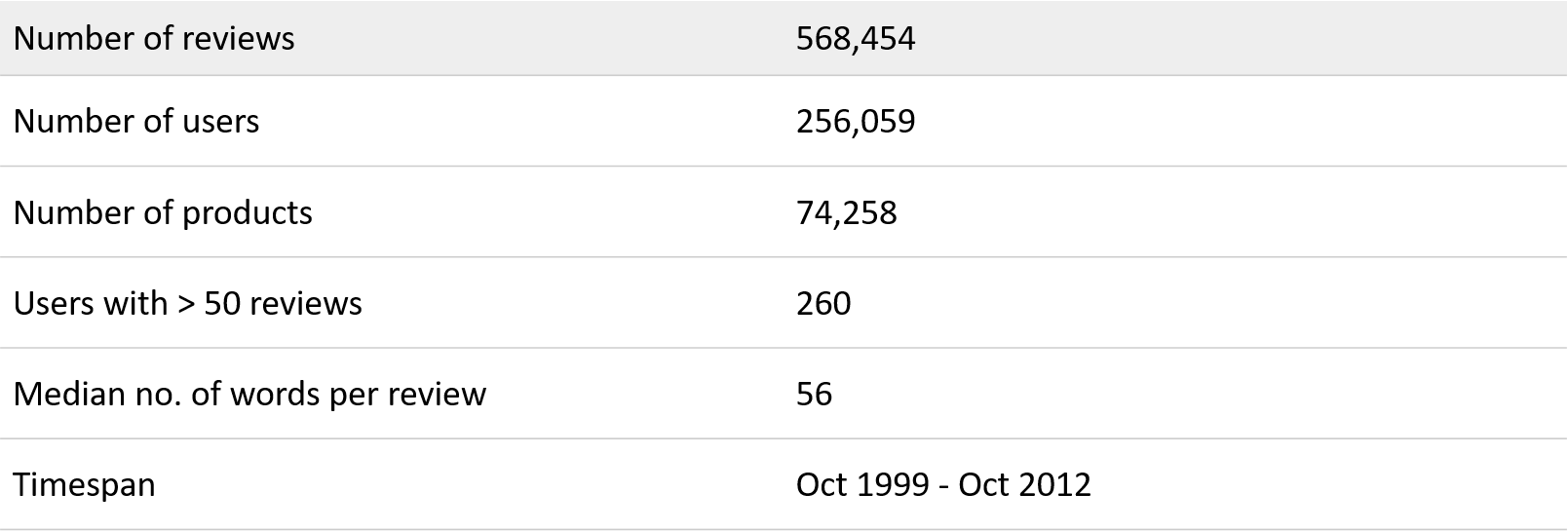
**DESCRIPTION OF THE DATA**

* The Amazon Fine Food Reviews dataset consists of 568,454 food reviews Amazon users left up to October 2012.
* Volume of the data – 289MB
* Velocity – Amazon review dataset consists of real time reviews regarding food items by the users.
* Data must be categorized based on the rating provided. Thus the retrieval of the data becomes faster as the data becomes organized.
* Each day more than 5000 food items will be sold and this number increases gradually with increase in company’s performance.
* The columns in the dataset are:
* **Id** – column number
* **ProductId** - unique identifier for the product
* **UserId** - unique identifier for the user
* **Profile Name** – User Profile name
* **Helpfulness Numerator** - number of users who found the review helpful
* **Helpfulness Denominator** - 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

**DATASET**

****

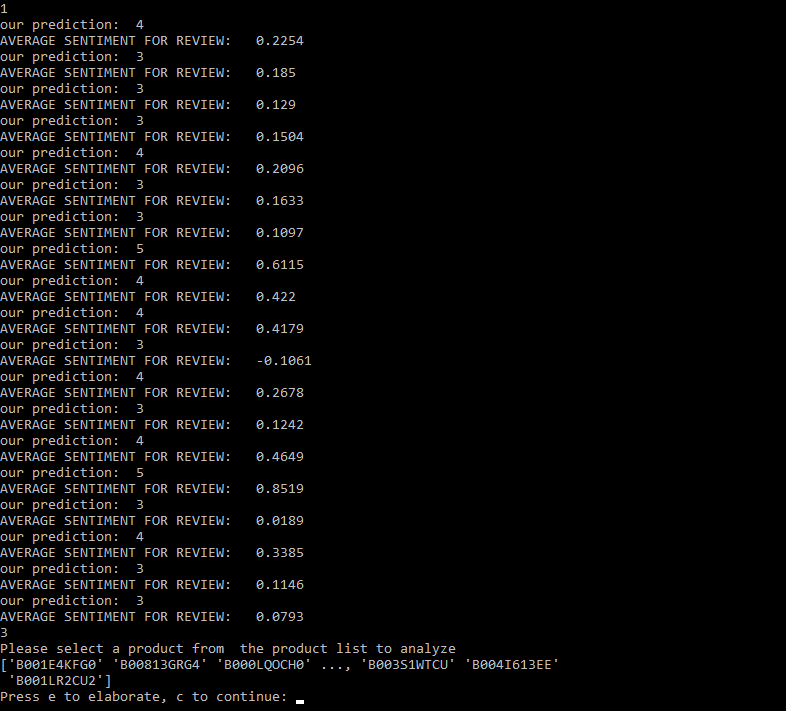
**DATASET STATISTICS**

****

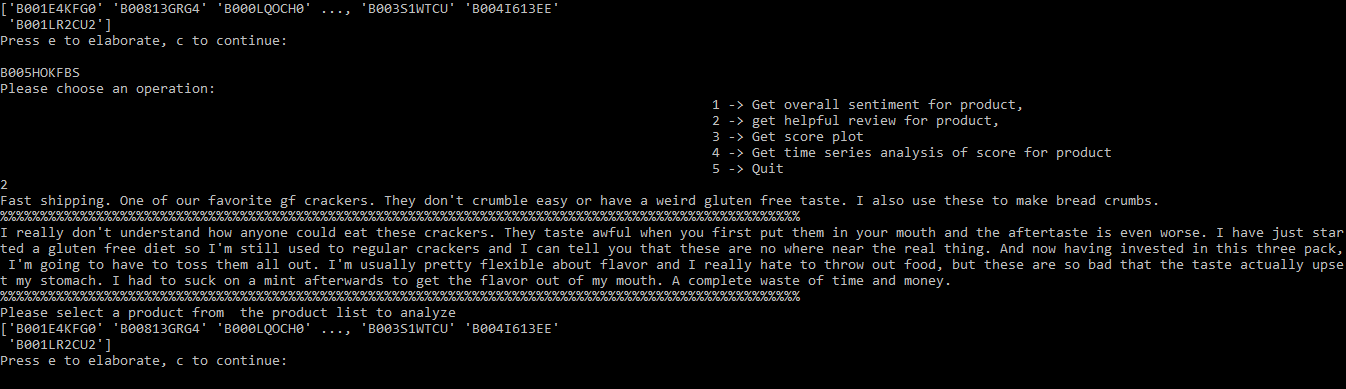
**IMPLEMENTATION AND ANALYSIS**

Python script using panda’s package is used for the implementation of the objectives

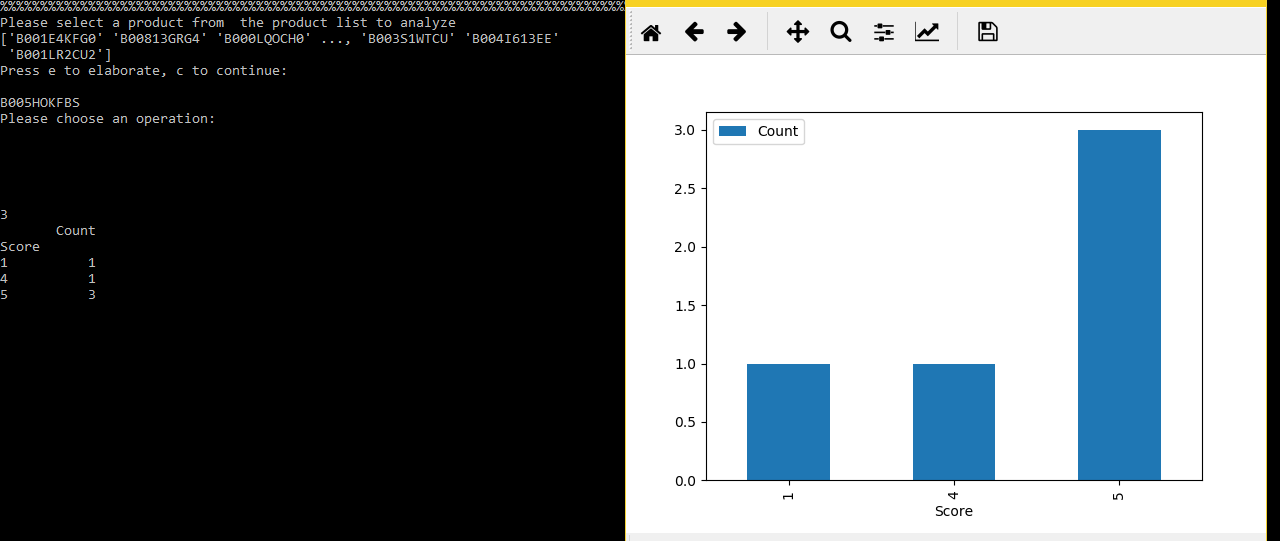
1. **Providing average sentiment for each product based on the text present in each review:** “vaderSentiment” library is used for the sentimental analysis of the comments by the user. Providing the time series analysis for the product based on the previous history. Sentiment provides the information of each product saying how good or bad the product is based on the comments given by the users. It signifies the company (Amazon) to improve the products which are not gaining appreciation from the users.



1. **Taking most helpful review from each product:** Most helpful reviews for a particular product is done by considering Helpfulness Numerator and Denominator. Helpfulness Numerator signifies that the review is most viewed by the user. Helpfulness Denominator signifies that the comment for that review is found good.



1. **Providing the details of the scores for each food item (Saying how many 1 star, 2 star reviews till 5 stars):** This will show the reaction of users for a particular product.



1. **Providing the time series analysis for the product:** The time series analysis is provided for each product based on the previous year’s records. This indicates how the product is seen by the people over time.

