Predictive Analysis of Amazon Products

A project report submitted to

Visvesvaraya Technological University. Belgaum, Karnataka

in the partial fulfillment of the requirements for the award of degree of

Bachelor of Engineering

in

Computer Science and Engineering

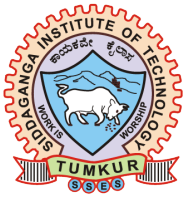
by

Ashish Gupta	**1SI13CS020**
Jagatjyoti G Tuladhar	**1SI13CS040**
Sachit Shrestha	**1SI13CS104**
Ujjen Man Bania	**1SI13CS127**

under the guidance of

Prof. K Bhargavi

Assistant Professor



Department of Computer Science and Engineering

Siddaganga Institute of Technology, Tumakuru

Jan, 2017

Department of Computer Science and Engineering
Siddaganga Institute of Technology
Tumakuru - 572103

CERTIFICATE

Certified that the Project Report entitled **"Predictive Analysis of Amazon Products"** is a bonafide work carried out by **Ashish Gupta (1SI13CS020), Jagatjyoti G Tuladhar (1SI13CS040), Sachit Shrestha (1SI13CS104) and Ujjen Man Bania (1SI13CS127)** in the partial fulfillment of the requirement for the award of the degree of Bachelor of Engineering in Computer Science and Engineering , Visvesvaraya Technological University, Belagavi during the year 2016-17. It is certified that all corrections/suggestions indicated for the internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

.....	.....
Guide	**Group Convener**
Prof. K Bhargavi	**Dr. B Sathish Babu**
Asst. Professor	**Professor**
Dept of CSE, SIT	**Dept of CSE, SIT**
.....	.....
Dr. N R Sunitha	**Dr. Shivakumaraiah**
Professor and Head	**Principal**
Dept of CSE, SIT	**SIT, Tumakuru**

Name of the Examiners — Signature with Date

1. Prof.
2. Prof.

Acknowledgements

We consider this as a privilege to express a few words of gratitude to all those who guided us for the successful completion of our project work.
We express our deep sense of gratitude, indebtedness and sincere salutations to His Holiness **Dr. Sree Sree Sree Shivakumara Swamigalu**, *President, Sree Siddaganga Education Society*, for being a constant source of inspiration in the course of study.
We express our gratitude and will remain indebted to **Dr. Shivakumaraiah**, our beloved *Principal, SIT, Tumakuru*, for fostering an excellent academic environment in this institution and also providing excellent lab facilities, which made our endeavor possible.
We would like to express our sincere gratitude to **Dr. N. R. Sunitha**, *Professor and Head, Dept of CSE, SIT, Tumakuru* for her encouragement and valuable suggestion.
It is a genuine pleasure to express our deep sense of thanks and gratitude to our convener **Dr. B. Satish Babu**, *Professor, Dept of CSE, SIT, Tumakuru*. His dedication and keen interest above all his overwhelming attitude to help his students had been solely and mainly responsible for the completion of our project.
With profound sense of gratitude, we acknowledge the guidance and support extended by **Prof K. Bhargavi**, *Assistant Professor, Dept of CSE, SIT, Tumakuru*. Her guidance gave us the environment to enhance our knowledge, skills and to reach the pinnacle with sheer determination, dedication and hard work.
Finally, we would like to express thanks to our **parents**, friends and all those who have directly or indirectly helped us in the successful completion of our project.

Abstract

In the recent years, it has become a trend for people to buy products online. E-Commerce sites has provided a platform for the customers to purchase goods online in a user friendly way. Unfortunately, there is no guarantee that the product purchased will satisfy the customer. The customers usually tend to purchase the product online by viewing the reviews given by the other customers. Customers like to get opinion from other customers about the product before buying it online.

Glancing through all the reviews would be time consuming and the customer may have difficulty in determining the quality of the product because of the mixture of good and bad reviews given by the other customers and also the customer would want to know the future trend of that product. In this project a predictive analysis system detects hidden sentiments to rates the product accordingly. The customer draws inferences about the quality of the product through various graphs and charts that is generated using this analysis. The polarity of the reviews, i.e. whether the product is good, bad or neutral is also provided. This system will also provide a prediction about the future success of the product by reviewing all the comments according to the timestamp.

The objective of this project is to generate a rating for the product by analysing the reviews given by the customers using different algorithms. Along with the rating, a chart or graph is also generated in order to show the customers the ratio of good, bad and neutral comments. We also predict the future success of the product.

The problem statement that we mentioned above can be handled by creating a framework on Python. For this project, the reviews are gathered

from Amazon E-Commerce website. The review collected from the Amazon e-commerce site using scrapping algorithm are subjected to Python for organization of data in tabular and well fashioned manner. This data is subjected to preprocessing model developed in python for analysis in terms resulting as input for sentimental analysis. The rating of different time periods are stored which will be helpful in predicting the future trend of the product by using a time series algorithm.

We believe this project is going to help the customer, since they don 't have to read all the reviews of a particular product they only need to view the rating of the product. Also the customers would want to know the future trend of the product they are planning to purchase so that the product that they buy won't degrade or go out of fashion after a certain period of time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background Study

Predictive analysis is the advanced analysis technique that are used to predict about unspecific future events. It uses techniques like, machine learning, statistical algorithm, artificial intelligence and data mining in order to analyze present-day data so that the predictions on future outcome based on historical data. The goal of the predictive analysis is to do depth analysis and to go further to know what would happen in order to provide the best assessment of what may happen in the future.

The main idea of predictive analysis was to crawl all the textual content i.e. reviews of Amazon websites of a particular product at regular interval of time, process it, store the extracted reviews into a database, build a model and finally predict the future outcome of the product.

In this project, the reviews given by the customers over a particular product in the Amazon website is extracted. These reviews are the main source of information that are needed to analyse the product. The reviews given by the customers are filtered by using the NLTK (Natural Language Tool Kit) API (Application Programming Interface). The reviews are first tokenized into separate words using a tokenizer. Each token of words are then paired with their part of speech. After pairing them, the base words are extracted. The words which doesn 't play any significant role in the analysis such as is, are, am, etc. are discarded. Then finally we filter the words based

on the parts of speech. This completes the pre-processing of reviews. The database called "Bag of Words" created which contains words like good, bad, worst and better along with sentiment values assigned to it. Next, compare token of words of each review with the bag of words that is stored in the database. If the token matches with the words in the bag of word, then the value if assigned to the token. Each words in the review now consists of a sentiment value. If the token doesn't match with any words in the bag of words, then discard that word. All the sentiment values in the review are aggregated to produce a final sentiment value for the review.

The final sentiment value of the review rated the review as good, bad or neutral. A timestamp is also recorded for the reviews. All the sentiment values of each reviews of the timestamp are then calculated to produce the final rating of the product for that period of time. Now using predictive analysis compare the rating of the product for each timestamp. If the rating of the product decreases after each contiguous period of time in the timestamp, then can predict that the product may not succeed in the future. Constant rating indicates that the product may succeed in the future as well.

1.1.1 Scope of Project

The importance of this project relate mostly to the websites that deal with e-commerce like Amazon. Through this project we are attempting to boost the e-commerce industry by making it more user centric in nature. The e-commerce industry as of today focuses mainly on providing very generic products to customers. E-commerce websites operating in this rapidly evolving ecosystem need new collaborative tools to attract customers and prioritize the optimal use of the websites. Through our project we are going to make this more user-centric rather than general in nature by analysing the customer reviews and predicting the trends of the product. This project helps us standardize systems and procedures to better manage and deliver e-commerce services in a much more effective way.

1.2 Related Works

Title of the work: *Sentiment Analysis and Opinion Mining*
Authors: *Liu Bing*
Publication details: *Synthesis Lectures on Human Language Technologies, May 2012, Vol. 5, No. 1, Pages 1-67*
Description: Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative (or neutral). Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level.The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions.

Title of the work: *Mining and summarizing customer reviews*
Authors: *Hu M and Liu B*
Publication details: *Proceedings of the tenth ACM SIGKDD, International conference on Knowledge discovery and data mining. ACM, New York, NY, USA,August 22-25, 2004, pages 168-177*
Description:A list of positive words and a list of negative words, respectively, based on customer reviews. The positive list contains 2006 words and the negative list has 4783 words. Both lists also include some misspelled words that are frequently present in social media content. Sentiment categorization is essentially a classification problem, where features that contain opinions or sentiment information should be identified before the classification.

Title of the work: *Polarity shift detection,elimination and ensemble: A three-stage model for document-level sentiment analysis*
Authors: *RuiXiaa, FengXub, JianfeiYua, Yong Qia and Erik Cambriac*

Publication details: *Information processing and management volume 52,Issue 1,January 2016,page 36-45,Emotion and Sentiment in Social and Expressive Media*
Description: The polarity shift problem is a majour factor that affects classification of machine- learning- based sentiment analysis systems. In this paper a three-stage cascade model has been proposed to address the polarity shift problem in the context of document-level sentiment classification. At first each document is split into a set of sub sentences and a hybrid module is build that employees rules and statistical method to detect explicit and implicit polarity-shifts, respectively, secondly a polarity-shift elimination method is proposed to remove polarity-shift in negations. Finally,base classifier on training subsets are trained and divided by different types of polarity shifts and weighted combination of the component classifier for sentiment classification is used. The results on a range of experiments illustrate that the approach significantly out performs several alternative methods for polarity-shift detection and elimination.

Title of the work: *Sentiment analysis using subjectivity summarization based on minimum cut*
Authors: *Pang B and Lee L*
Publication details: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL'04.. Association for Computational Linguistics, Stroudsburg, PA, USA,2004, Pages 1-8*
Description: A novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document. The sentences in the document as either subjective or objective, discarding the latter; and then apply a standard machine-learning classifier to the resulting extract. This can prevent the polarity classifier from considering irrelevant or even potentially misleading text. Subjectivity extracts can be provided to users as a summary of the sentiment-oriented content of the document. The results show that the subjectivity extracts create accurately represent the sentiment information of the originating documents in a much more compact form: depending on choice of downstream polarity classifier, we can achieve highly statistically significant improvement (from 82.8% to 86.4%) or maintain the same level of performance for the polarity classification task while retaining

only 60% of the reviews'words. Also, we explore extraction methods based on a minimum cut formulation, which provides an efficient, intuitive, and effective means for integrating inter-sentence-level contextual information with traditional bag-of words features.

Title of the work: *Sentiment analysis approach to adapt a shallow parsing based sentiment lexicon*
Authors: *Jayraj M. Desai and Swapnil R. Andhariya*
Publication details: *Innovation in Information, Embedded and Communication, Systems(ICIIECS), 2015 International conference on, march 2015*
Description: With the rapid growing of IT development and e-commerce web sites, increasing trends in people to posting online reviews. Sentiment lexicons has offend used to analyzing the large volume of online review data available and gain useful knowledge from it. Most of the sentiment lexicon are aspect base, uses dependence parsing for extracting the word which are not be able to classify the sentimental word so accurately. Try to propose a method which combines sentiment lexicon and shallow parsing. Which determine aspect and domain base sentiment analysis and then assign polarity to a lexicon.Main merits of proposed methods is that it highly accurate and automatically generating structured to avoiding the cost of manually labelling data. The shallow parsing used to analyses sentence and get the constituents words. It will not considering the internal structure of constituent word, nor specifying their value in sentence. Then using polarity of words positive or negative evolution of the product.

Title of the work: *Predictive Analytics Integrated with and Social Media*
Authors: *Madhusudhan V and Shilpa N R*
Publication details: *International Research Journal of Engineering and Technology(IRJET),Volume 3 Issue:4, April-2016, Pages 1-5*
Description: Social media is often used as a very huge source of data that is freely available for mining and analysis. This is often used for sentiment analysis and predictive analysis. Sentiment analysis using social media provides the sentiments of

the people based on their posts on social media. It is also used to predict the future situations by predictive analysis. It provide the introduction to Social media analytics, predictive analysis and sentiment analysis and provide a survey on various applications of predictive analytics integrated with social media such as the usage of predictive analytics with social media, prediction of the next president to be elected, stock predictions and even predicting the success rate of movies by analysing tweets and YouTube comments. Predictive analytics with social media is also applied in the field of crime analysis, finding the strength of ties between two people over social media and recommending music and events based on a person's preferences.

Title of the work: *Predictive Modeling and Sentiment Analysis: Data Mining Approach*
Authors: *Mr. Vijay D. Chougule and Mrs. Anis N. Mulla*
Publication details: *International Research Journal of Engineering and Technology(IRJET),Volume 3 Issue:8, Nov-2015,Pages 1-5*
Description: Data mining technology have widely been applied in various businesses and manufacturing companies. Sharing data has become a trend among business partnerships, as it is supposed to be a mutually beneficial way of increasing productivity.Use sentiment analysis and prediction modeling to determine future scope of product. For sentiment analysis we take as an example online review of peoples towards the product they bought and services they received. Analysis of different online reviews on large scale will help to produce useful actionable knowledge. Conducting extensive experiments on large data set confirms the effectiveness of the proposed approach.

Title of the work: *Product Sales Prediction Based on Sentiment Analysis Using Twitter Data*
Authors: *Dipak Gaikar and Bijith Marakarkandy*
Publication details: *International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 6 (3), 2015, Pages 1-10*
Description: Online social media websites represent how fundamental information is

created, transferred and consumed. Social media, user generated content in the form of comments, blog posts and tweets establish a connection between the producers and the consumers of information. Today's world is connected to each other via social network like Twitter millions of people connected to each other through that network. Tracking the pulse of the social media contain, enables companies to gain feedback and insight in how to improve and market products better. It continues to offer new opportunities for organizations to directly interact with their customers or audience, the aimed at monitoring the online reputation of an organization, brand or person, social media and search engine result. This research paper uses a survey approach for movie sales prediction. This paper analyses, impact of the positive, negative, strongly positive and strongly negative online reviews of movies on the audience. It should be noted that the user feedback is given prior to watching the movie only on the basis of the online reviews. The result of this research will help the film industry to effectively address and meet the expectations of customers and stakeholder. This paper also investigates techniques for twitter data extraction using an API key.

1.3 Project Problem Statement

In the current scenario, we notice that customers purchase the product from the Amazon websites and give their review about the product after they use it for some period of time. The customer who wishes to buy the product from Amazon websites they go through the reviews about that product but reading all the reviews may take a lot of time and mixed reviews from other customers may confuse the customer. Reading only a few reviews may not give clear view about the product. Some reviews may be good and some may be bad. Every customers will have their own point of view about the product which they bought. Hence, an accurate review may not be found. Also the customers may want to know the future success of the product after using it. Customers would want to know till which extent the product may succeed in the future so that the product will not deteriorate in the future. What we require is the technique which crawls all the reviews of a particular product and give a rating to that problem. Further we should predict the future trend of the product.

1.4 Objective of Project

The main objectives of our project are:

- To generate most appropriate product rating based on reviews.
- People can easily decide whether the product posted is good or bad.
- To generate charts from the review given by the customers for the particular product.
- Predict future trends of the product.

1.5 Organization of the Report

The first chapter of this report deals with the overview of the review based sentiment analysis and related work in that area. It gives us in-depth review of our project which include problem statement and the objectives that needs to be achieved.

The second chapter mentions about the high level design of the project. It includes architectural design along with its module descriptions and the type of software development methodologies that will be implemented in this project. Functional requirements and non-functional requirements are also mentioned in this chapter.

Chapter 2

High-level Design

This chapter aims to present the software development methodology used in order to create and maintain the applications. This chapter also includes the architecture involved in designing the system and the corresponding modules. It also contains the functional requirements that are needed to be addressed in the project for efficient working of the system. Finally, there are the design constraints and assumptions that are considered in the project.

2.1 Software development methodology

A software development methodology is a framework that is used to structure, plan, and control the process of developing an information system. It is the process of computer programming, documenting, testing and bug fixing involved in creating applications and frameworks resulting in a software product.

The software development methodology followed here is the Spiral model. Spiral model is a combination of iterative development process model and sequential linear development model (Waterfall model) with very high emphasis on risk analysis. Based on the unique risk patterns of a given project, the spiral model guides a team to adopt elements of one or more process models such as incremental, waterfall or evolutionary prototyping. The risk in the project is that during analysis, the training dataset might contain group of words which are generally considered as negative but will depict positive results and vice versa. In this model, the software is developed in

a series of evolutionary releases. It couples the iterative nature of prototyping with the controlled and systematic aspects of the waterfall model.

Spiral model is best suited for this project as the project is going to be released in an evolutionary manner. First release generates the real time dataset which is obtained by implementing the review scrapping algorithm. The second release pre-processes the dataset and generates a list of refined words. The third release runs a sentiment identification algorithm on the list of refined words and the result will be presented as positive, negative or neutral. The review identifier, sentiment value, polarity along with the timestamp will be stored in a database in this release. The fourth release generates rating for the product. The last release predicts the rating for the nth week by using the rating of the previous weeks. Finally, the success or failure of the product for the nth week is predicted.

In this way, the complexity of the product is increased with every release. Figure 2.1 shows a diagram of spiral model. Spiral Life Cycle Model is one of the most flexible SDLC (Software Development Life-Cycle) models in place. Development phases can be determined by the project developers, according to the complexity of the project. Project monitoring is very easy and effective. This makes the model more transparent. Changes can be introduced later in the life cycle without any complexity.

The spiral model has four phases. A software project repeatedly passes through these phases in iterations called Spirals.

2.1.1 Activities undertaken during requirements gathering and analysis

This phase consists of two major activities: requirement gathering and requirement analysis. In the first step, all the reviews of the products from the amazon websites are scrapped (collected). This forms the requirement gathering phase. Then the unstructuredness in the reviews is resolved to eliminate all the inconsistencies, incompleteness and ambiguities existing in the product reviews, this forms the requirement analysis phase.

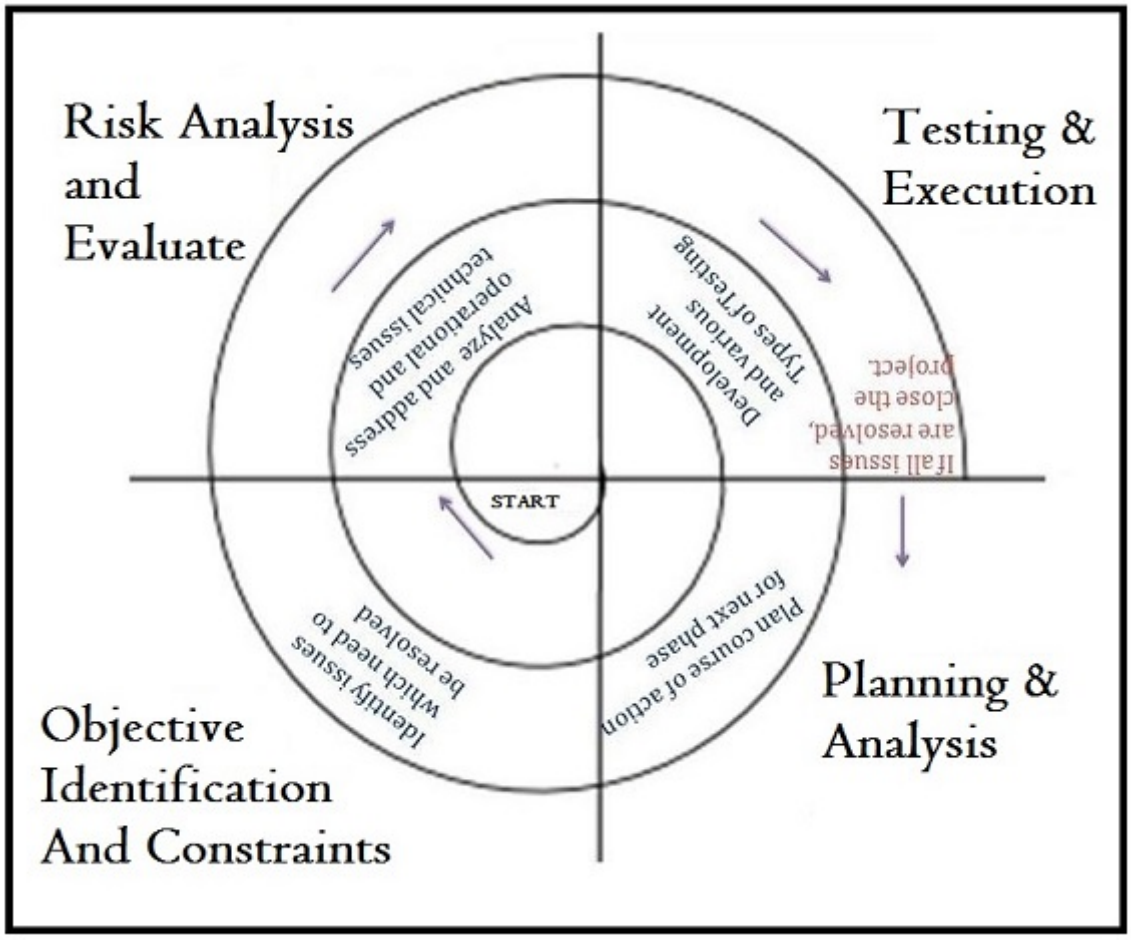


Figure 2.1: Diagram of Spiral Model

2.1.2 Activities undertaken during design

Design phase starts with the conceptual design on the baseline spiral and involves architectural design, logical design of modules and final design in the subsequent spirals.

2.1.3 Activities undertaken during construction or build

Construction phase refers to production of the actual software product at every spiral. In the baseline spiral when the product is just thought of and design is being developed POC (Proof of Concept) is developed in this phase to get customers feedback.

Then in the subsequent spirals with higher clarity on requirements and design details a working model of the software called build is produced with a version number. These builds are sent to customer for feedback.

2.1.4 Activities undertaken during Evaluation and Risk Analysis

Risk Analysis includes identification estimating, and monitoring technical feasibility and management risks, such as scheduling slippage and cost over run. After testing the build, at the end of first iteration, the customer evaluates the software and provide feedback. Based on the customer evaluation, software development process enters into the linear iteration and subsequently follows the linear approach to implement the feedback suggested by the customer. The process of iteration along the spiral continues throughout the life of the software.

2.2 Architecture

2.2.1 System Architecture

The top view of the system architecture is shown in Figure 2.2. The real time datasets are extracted from the Amazon website through review log and saved in a file. The pre-processor breaks the reviews from the file into tokens and part-of-speech is attached to each token. The base words are extracted from each token using stemming. Finally, the tokens are filtered to retain the verbs, adverbs and adjectives. The Sentiment Identifier determines the polarity of the review and classifies it as positive, negative or neutral. The sentiment database contains the details of each review such as review identifier, sentiment value, polarity and timestamp. The rating generator calculates the rating of the product on a weekly basis using the attributes from the sentiment database. The Rating Predictor predicts the rating for the nth week using time-series algorithm. Finally, the results are displayed through pi-charts and bar graphs.

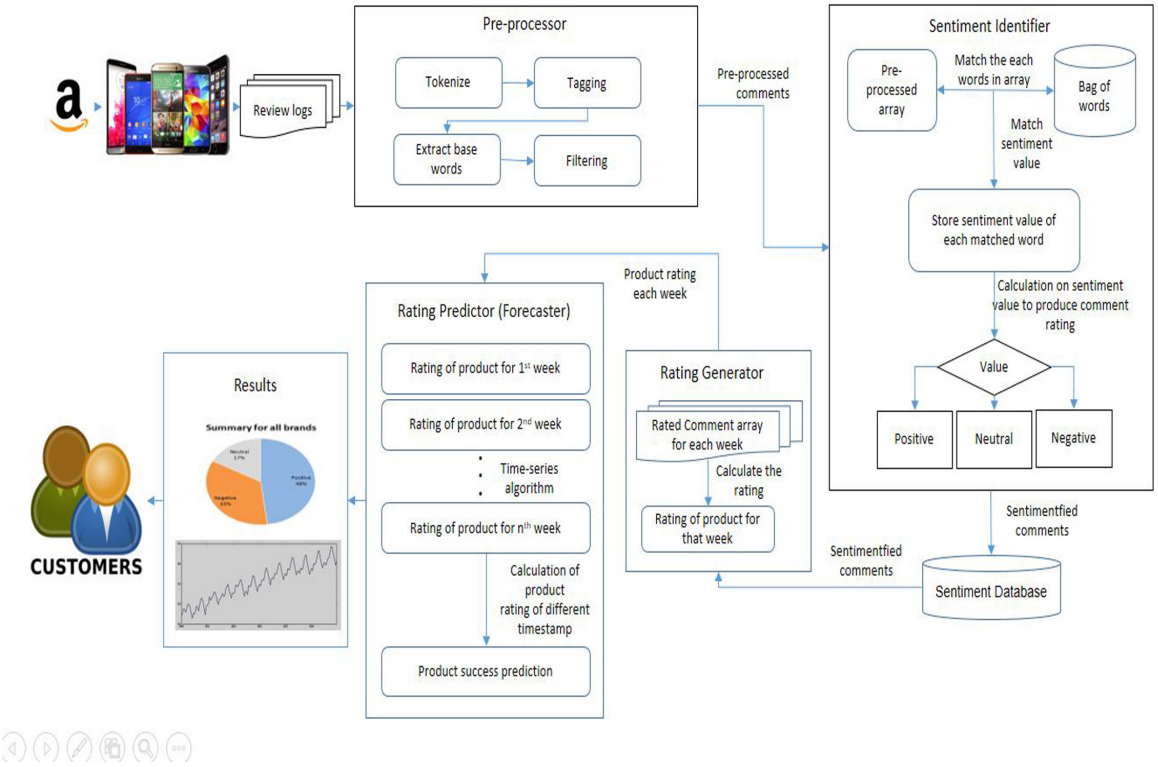


Figure 2.2: System Architecture

2.2.2 Data Flow Diagram

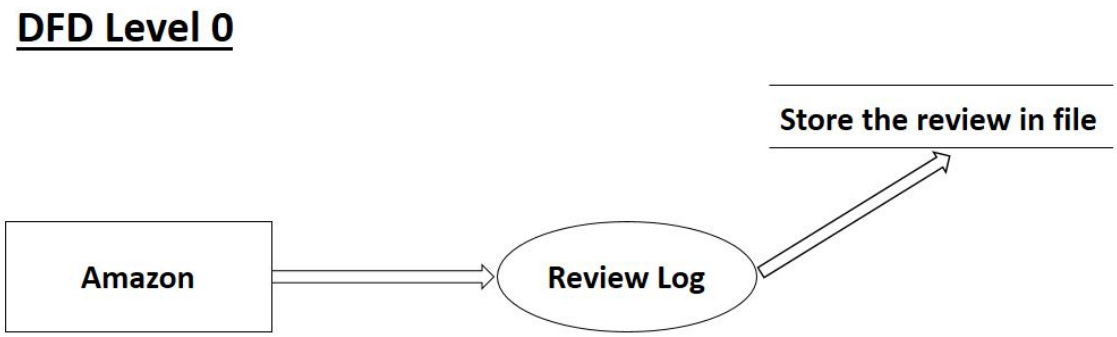


Figure 2.3: Level 0 DFD

At this level we scrap the reviews of the particular product from the amazon websites and store them in the review log files which is a json file which contain title of the product, url of the site, date on which the review was posted and review of the product.

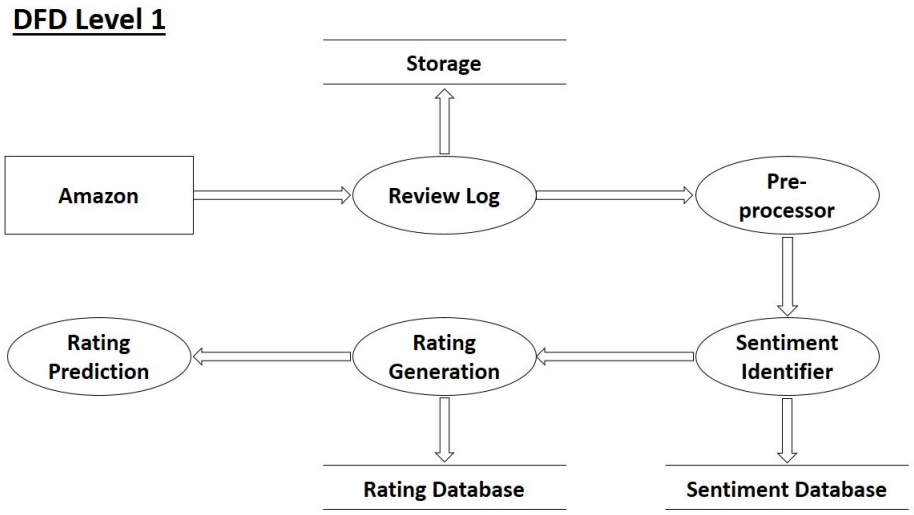


Figure 2.4: Level 1 DFD

At this level we we process the review which is stored in the json file which is done by pre-processor module. The reviews are stored in an array and then matched with the bag of words, the sentiment values are compared and according to that the classification of reviews are done and stored in an sentiment database. The rating is generated by using the data from the sentiment database and the rating is stored in a rating database for further prediction of the success and failure of the product.

2.2.3 Usecase Diagram

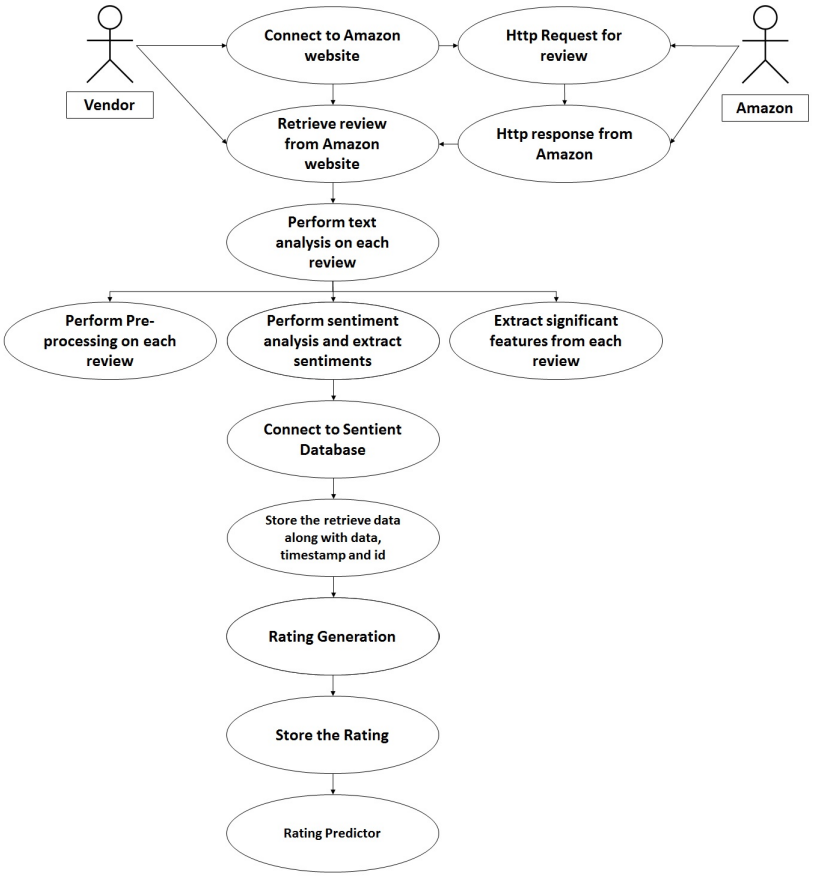


Figure 2.5: Usecase of the project

The above figure 2.5 represent the use case diagram of the project. First the vendor connect with the amazon website by entering the valid url and sends the request for retrieving the reviews of the products. Amazon sends the response and the review are scraped from the website to the review log file to perform the text analysis on each review. The text analysis performs the pre-processing task and sentiment identification task and extract some significant features from each reviews and store the result in a database for calculating the rating of the product and further predicting the success and failure of the product.

2.3 Modules Description

The five major modules from the system architecture described in Figure 2.2 has been further analysed here.

2.3.1 Review Log

Customers express their opinion about the product on the e-commerce site from where they have viewed the product. The opinions are expressed in different ways, with different vocabulary, content of writing and usage of short forms and slangs making it unstructured and disorganized. The reviews for a particular product are collected by the Review Log using review scrapping algorithm.

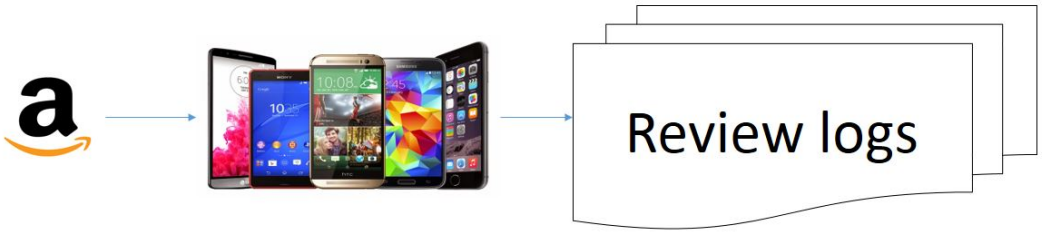


Figure 2.6: Process of Review Logging

2.3.2 Pre-processor

The reviews extracted by the Review Log are unstructured and contains words which have no scope in Sentiment Analysis. The reviews need to be refined before performing any sort of analytics. Text pre-processing is the filtering the extracted reviews before analysis. It includes identifying and eliminating non-textual content and content that is irrelevant to the area of study.

The pre-processor module first breaks down the unstructured review into tokens. Tokenization is the task of chopping the review up into pieces, called tokens, and at the same time throwing away certain characters, such as punctuation.

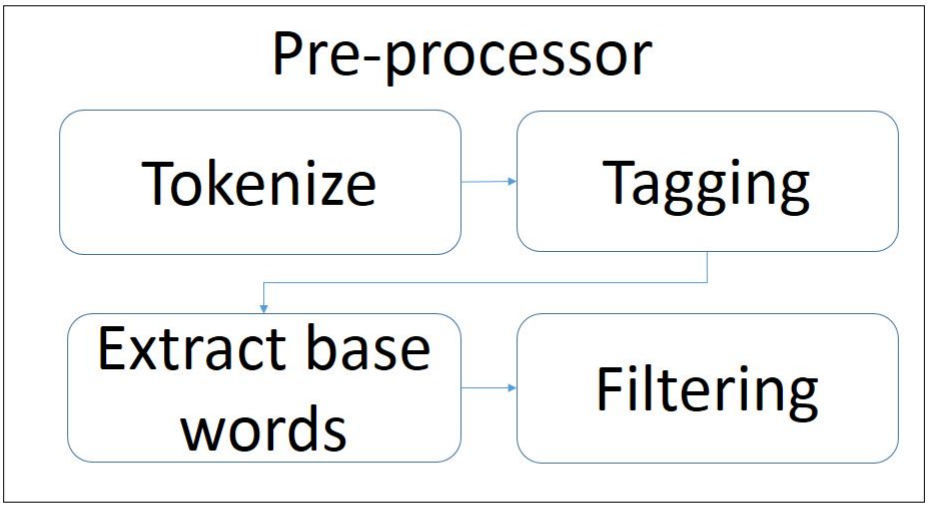


Figure 2.7: Component of Pre-processor

A Part-Of-Speech Tagger assigns parts of speech to each token, such as noun, pronoun, verb, adjective, adverb and preposition. The Part-Of-Speech required for sentiment analysis are mainly verbs, adverbs and adjectives.

The base words are extracted using a process called stemming. It involves clubbing of two or more subsequent words and removing the duplicates to improve the performance of methods. Stop words play a negative role in the task of sentiment classification. They do not carry any sentiment information and are of no use to us. Thus, the stop words like he, she, at and on needs to be discarded.

The Part-Of-Speech required for sentiment analysis are mainly verbs, adverbs and adjectives. Finally, the tokens are filtered to retain the verbs, adverbs and adjectives. This module yields a pre-processed list on which further analytics can be applied.

2.3.3 Sentiment Identifier

The Sentiment Identifier takes the pre-processed list as input and compares each element in the list with the Bag of Words. The Bag of Words contains many pre-defined words with a sentiment value attached with each word. If the element in the pre-processed list matches with the Bag of Words, the corresponding sentiment value are stored in sentiment array. The stored sentiment value is processed and rating for that particular review is generated.

Based on the rating, the review is classified as good, bad or neutral. The attributes such as review id, sentiment value, polarity and timestamp are stored in a sentiment database for each review.

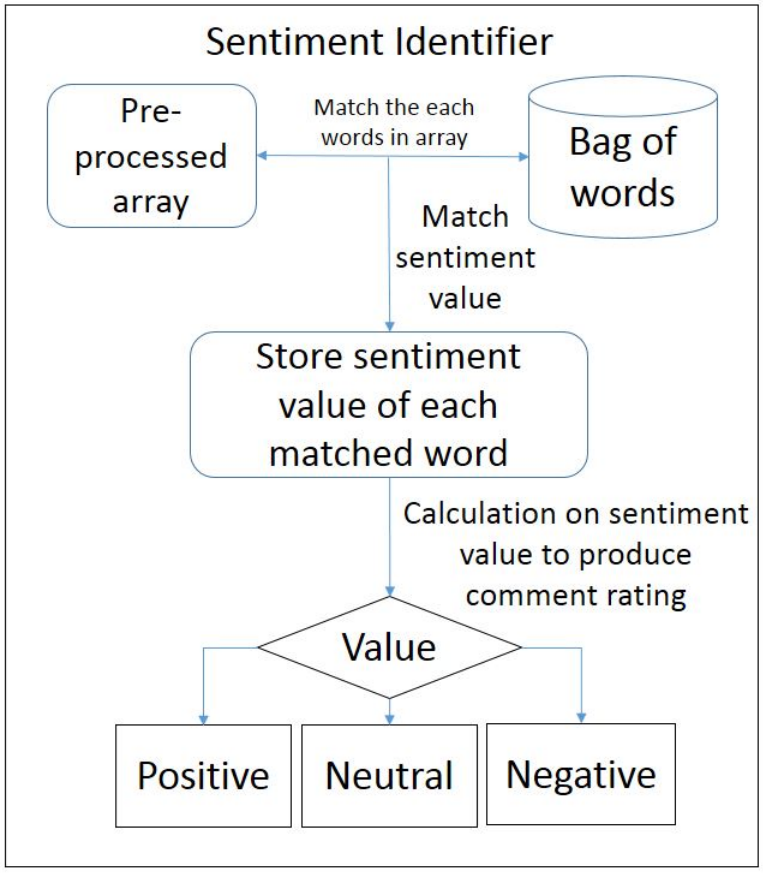


Figure 2.8: Sentiment Analysis

2.3.4 Rating Generator

The Rating Generator module makes use of the records from the sentiment database. Using the timestamp attribute, the product rating for each week is calculated and stored. The product rating for each week is generated and is recorded. Thus, the final product rating is obtained by taking the average of the ratings till the nth week.

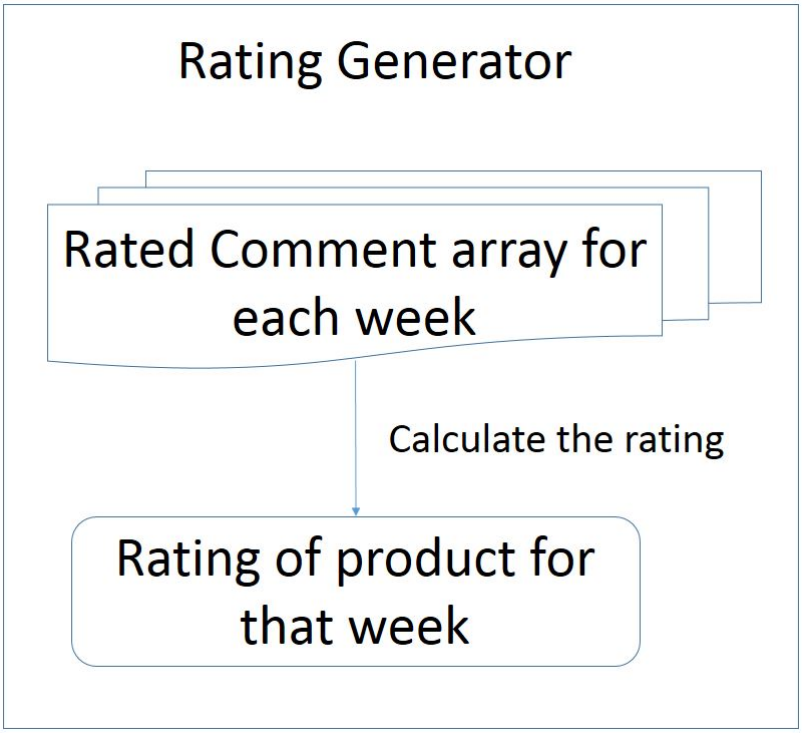


Figure 2.9: Workflow of Rating Generator

2.3.5 Rating Predictor

Time Series is a collection of data points collected at constant time intervals and are very frequently plotted via line charts. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, astronomy, communications engineering and largely in any domain of applied science and engineering which involves temporal measurements. These are analysed to determine the long term trend so as to forecast the future or perform some other form of analysis. Using this algorithm, the Rating Predictor module calculates the rating for the nth week using the records of the previous weeks. This results in the prediction of the success and failure of the product.

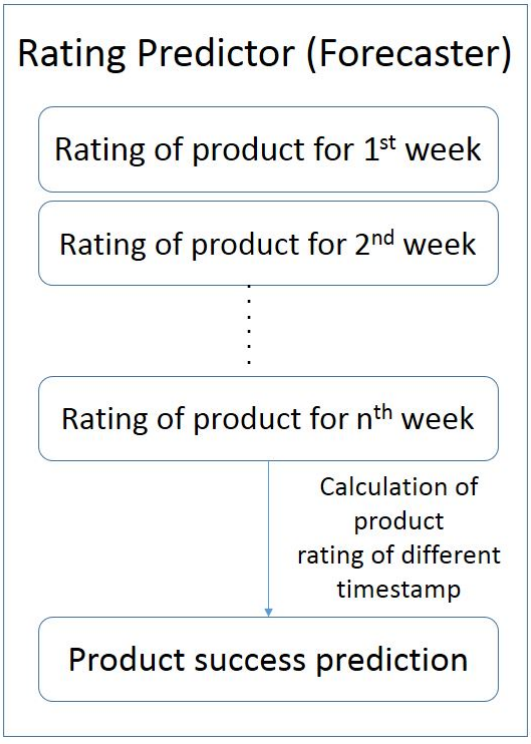


Figure 2.10: Flow of rating predictor

2.4 Functional Requirements

2.4.1 Retrieving Input

Name of the module: *Review Log*
Parameters: *Reviews*
Purpose: Collect reviews of the product onto the review log file. This function is the base of sentiment analysis. Customers express their opinion about the product on the e-commerce site from where they have viewed the product, this reviews act as a inputs for this analysis and this reviews are scrapped by using web scrapping tools.Here we retrieve the reviews from Amazon websites which are stored in the review log for the further pre-processing.

2.4.2 Featuring Evaluation/Processing

Name of the module: *Pre-processor*
Parameters: *Reviews*
Purpose: Since it is a machine learning approach, it divides the sentences into individual words which are called tokens and it is done by the process called tokenization. These tokens will be appended with part-of-speech and then base words are extracted using the process called steaming.

2.4.3 Frequency Distribution

Name of the module: *Sentiment Identifier*
Parameters: *Reviews, Bag of Words, Sentiment Value*
Purpose: Naive bayes classifier function are use to classify the sentiments into three classes i.e. positive, negative and neutral sentiments.

Positive Sentiments: These are the good words about the target in consideration, if the positive sentiments are increased, it is referred to be good. In case of product reviews, if the positive reviews about the product are more then that product are bought by many customers.

Negative Sentiments: These are the bad words about the target in consideration. In case of product reviews, if the negative reviews about the product are more, no one intend to buy it.

Neutral Sentiments: These are neither good nor bad words about the target. Hence it is neither preferred nor neglected.

2.4.4 Product Rating

Name of the module: *Rating Generator*
Parameters: *Reviews, Sentiment Value*
Purpose: The purpose of Product Rating is to generate the appropriate rating about the product. This module makes use of the records from the sentiment database. Using the timestamp attribute, the product rating for each week is calculated and stored. The product rating for each week is generated and is recorded. Thus, the final product rating is obtained by taking the average of the ratings till the nth week.

2.4.5 Product Rating Predictor

Name of the module: *Rating Predictor*
Parameters: *Reviews, Rating, Timestamp*
Purpose: The purpose of this module is to predict the future trend of the product. Rating Predictor module calculates the rating for the nth week using the records of the previous weeks. This results in the prediction of the success and failure of the product.

2.5 Constraints and Assumptions

Sentiment is inherently subjective from person to person, and can even be irrational. It's critical to mine a large and relevant sample of data when attempting to measure sentiment. No particular data point is necessarily relevant. It's the aggregate that matters. An individual's sentiment toward a product may be influenced by one or more indirect causes; someone might have a bad day and give a negative remark about something they otherwise had a pretty neutral opinion about. Also, since sentiment very likely changes over time according to a person's mood, world events, and so forth, it's usually important to look at data from the standpoint of time.

The biggest constraint of sentiment analysis is that it tries to analyse exact human sentiments through their reviews thus it will not fetch 100 percent accurate result. The analysis is always about how close we can get to the 100 percent mark. People use sarcasm and other forms of figures of speech which cannot be judges and are problematic for machines to detect. In the assumption part of the project, we have assumed that most of the reviews given by the customers are true and are not fake.

Chapter 3

Detailed Design

This chapter deals with the detailed design of the project. It encloses the dataset, algorithms and data structures used in the project. Initially this chapter comprises of the interface design that contains the topics like, how our project interact with the people, what are the software and hardware we used. Following this, next we have the data structures and algorithm that are used in the development of the project. These fields above are mentioned for each functions in the project. Finally there is a section that mentions the data source/databases used and also the formats that it encloses. So, in sum it encloses every aspect involved in the designing of the project.

3.1 Interface design

The purpose of user interface is to enable people to interact with the application. The interface should be simple and easy to use so that the people wonâĂŹt get confuse and frustrated. The environment runs on windows. The interface contain the button which scrap the reviews from the amazon websites and stores in the file for further processing and finally the charts are generated as a final result.

3.1.1 Product Entry page

End user specifies the product about which he wishes to see the rating. The interface is simple and user friendly it contains two fields in which user entry the product name along with the asin number which is the unique number for each product of the amazon. After pressing the submit button the reviews start scraping and stored in the json file which is further used for pre-processing and sentiment identification of the review, which will further generate the graph as an output.

Figure 3.1: Product Entry Page

3.1.2 Machine Learning Approach

The software which is being created will always work on a dataset which in our case is the text files. We are using machine learning approach for predictive analysis whose generic architecture is shown in figure 3.1. Within the machine learning approach, a series of feature vectors are chosen and a collection of tagged corpus of text. In a machine learning approach, the selection of features is crucial to the success rate of the classification.

Most commonly, a variety of unigrams (single words from a document) or n-grams (two or more words from a document in sequential order) are chosen as feature vectors. Other proposed features includes the number of positive words, number of negative words, and the length of a document. Support Vector Machines and the Naive Bayes algorithm are the most commonly employed classification techniques. The reported classification accuracy ranges between 63 percent and 84 percent, but these results are dependent upon the features selected.

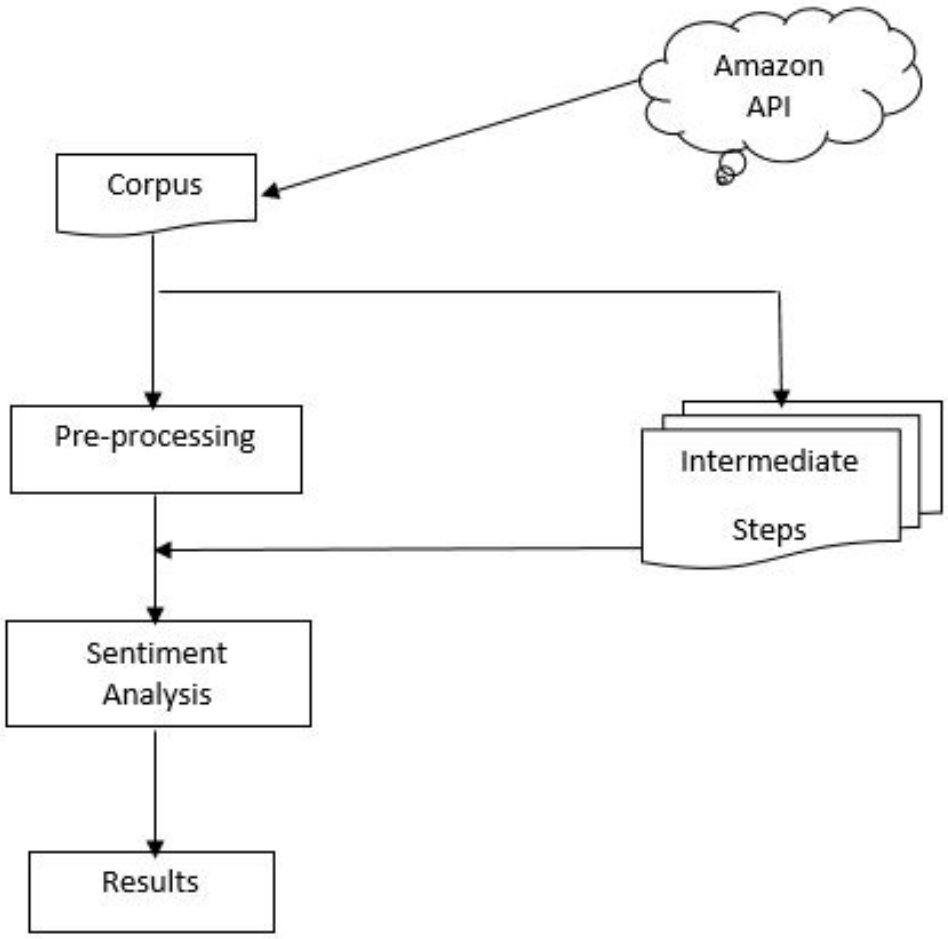


Figure 3.2: Generic Architecture of Machine Learning approach

3.1.3 Design and Analysis of proposed Approach

The superiority of both lexical approach (for its speed) and machine learning approach (for its accuracy) are not unknown to the world. A lexical approach is fast because of the predefined features (e.g. dictionary) it employs for extracting sentiments. Having a dictionary to refer at runtime reduces the time consumption almost exponentially. To improve performance of lexical approaches the feature set has to be increased drastically, i.e. a very large dictionary of variety of words with their frequencies has to be provided at runtime. This increases the overhead of the system and hence the performance suffers.

Thus, there is a constant trade-off between Performance vs Time. On the other hand, machine learning approach employ a recursively learning and tuning of their features, given large input datasets, improves its performance way beyond any lexical approach can achieve. However, due to this runtime performance tuning and learning the system undergoes drastic fall in time constraints. Our goal is to propose an approach that is a combination of both lexical and machine learning, hence exploit the best features of both in one. Our hybrid naÃŕve bayes follows the ritual four steps namely: Data collection, Preprocessing, Sentiment classification and rating generation.

3.1.4 Data Collection: Amazon E-commerce Website

For processing and classification of review we are collecting the reviews from amazon websites by using the tool called web scraper which is also called url scraper or web crawler. This collected data or review from amazon is stored in the review log file which is the json file, this data stored in the file are used for further processing and analysis.

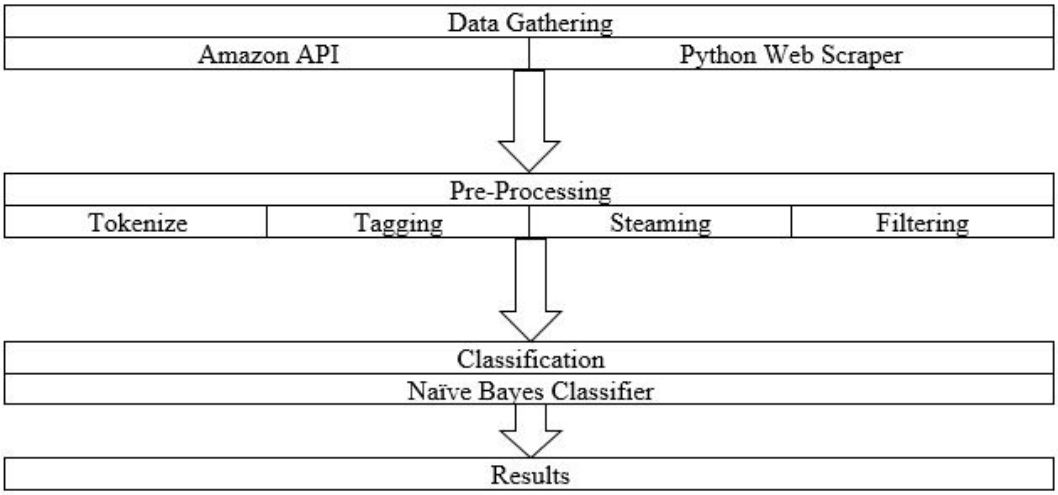


Figure 3.3: Processed Steps followed by Proposed Architecture

3.1.5 Preprocessing

The reviews gathered from amazon websites are a mixture of urls, and other non-sentimental data like hashtags, annotation. To obtain tokens, we first have to tokenize the text input then tagging parts of speech to each tokens are performed and finally extracting the base words are done in pre-processing of the reviews.

3.1.6 Sentiment Analysis: The classifier

NaÃŕve Bayes was our choice based on the interface from literature review carried out. NaÃŕve bayes is Bayesian probability distribution model based algorithm. In general all Bayesian models are derivatives of the well known Bayes Rule, which suggests that the probability of a hypothesis given a certain evidence, i.e. the posterior probability of a hypothesis, can be obtained in terms of the prior probability of the evidence, the prior probability of the hypothesis and the conditional probability of the evidence given the hypothesis.

3.2 Software used

To test the proposed approach, we created a setup with the following system requirements as shown in above table. The major software packages used are as follows:

Component	Type
Operating System	Windows XP /7 /8, Linux (Ubuntu)
Processor	i3 /i5 /i7 (32 /64 bits)
Min. Memory (RAM)	âĽě4GB
Min. Storage	20GB
Bandwidth	Uninterrupted High-speed Internet (1Mbps connection)
Software and Third-party tools	NLTK 2.0

Table 3.1: General System Requirements for our approach

3.2.1 Python 2.7 / 3

Python is our implementation language. Python is a general-purpose, inter-preted high-level programming language whose design philosophy code readability. Its syntax is clear and expressive. Python has a large and comprehensive standard library and more than 25 thousand extension modules. We use python for developing the backend of the test application crawler. This and the modules implemented are discussed later.

3.2.2 NLTK (Natural Language Toolkit)

It provides language processing modules and validation tools for natural language processing. The Natural Language Processing Toolkit (NLTK) is an open source language processing modules of human language in python. Created in 2001 as a part of computational linguistics course in the Department of computer and Information Science at the University of Pennsylvania. NLTK provides inbuilt support for easy to use interface over 50 lexicon corpora. NLTK was designed with four goals in mind.

1. **Simplicity** : Provide and intuitive framework along with substantial building blocks, giving users a pratical knowledge of NLP without getting bogged down in the tedious house-keeping usually associated with processing annoted language data.

2. **Consistency** : Provide a uniform framework with consistent interface and data structures, and easily guessable method names.

3. **Extensibility** : Provide a structure into which new software modules can easily by accommodated, including alternative implementations and competing approaches on the same task.

4. **Modularity** : Provide components that can be used independently without needing to understand the rest of the toolkit.

3.3 Data Structures and Algorithms

The detailed discussion on every module of the architecture along with their algorithm is given in this section.

3.3.1 Pre-processor

Purpose:The purpose of pre-processor is that reviews extracted by the Review Log are unstructured and contains words which have no scope in Sentiment Analysis. The reviews need to be refined before performing any sort of analytics. Text pre-processing is the filtering the extracted reviews before analysis. It includes identifying and eliminating non-textual content and content that is irrelevant to the area of study.

The pre-processor module first breaks down the unstructured review into tokens. Tokenization is the task of chopping the review up into pieces, called tokens, and at the same time throwing away certain characters, such as punctuation. A Part-Of-Speech Tagger assigns parts of speech to each token, such as noun, pronoun, verb, adjective, adverb and preposition. The Part-Of-Speech required for sentiment analysis are mainly verbs, adverbs and adjectives. The base words are extracted using a process called stemming. It involves clubbing of two or more subsequent words and removing the duplicates to improve the performance of methods. Stop words play a negative role in the task of sentiment classification. They do not carry any sentiment information and are of no use to us. Thus, the stop words like he, she, at and on needs to be discarded. The Part-Of-Speech required for sentiment analysis are mainly verbs, adverbs and adjectives. Finally, the tokens are filtered to retain the verbs, adverbs and adjectives. This module yields a pre-processed list on which further analytics can be applied.
Data Structures used : Arraylist

Algorithm:

Step1: Start

Step2: For each review fetched from the review log,
word_tokenize the review Wi

tag part-of-speech to each token {(w1, POS), (w2, POS), (w3, POS)…}

extract base words using stemming process

filter the words based on part-of-speech

Step 3: Store the result in a review array, R[]

Step 4: End

3.3.2 Sentiment Identifer

Purpose: The purpose of Sentiment Identifier is that it takes the pre-processed list as input and compares each element in the list with the Bag of Words. The Bag of Words contains many pre-defined words with a sentiment value attached with each word. If the element in the pre-processed list matches with the Bag of Words, the corresponding sentiment value are stored in sentiment array. The stored sentiment value is processed and rating for that particular review is generated. Based on the rating, the review is classified as good, bad or neutral. The attributes such as review id, sentiment value, polarity and timestamp are stored in a sentiment database for each review.
Data Structures used:Arraylist

Algorithm:

Step1: Start

Step2: For each review fetched from the review log,
word_tokenize the review Wi

tag part-of-speech to each token {(w1, POS), (w2, POS), (w3, POS)…}

extract base words using stemming process

filter the words based on part-of-speech

Step 3: Store the result in a review array, R[]

Step 4: End

3.3.3 Rating Generator

Purpose: The purpose of Rating Generator module is that it makes use of the records from the sentiment database. Using the timestamp attribute, the product rating for each week is calculated and stored. The product rating for each week is generated and is recorded. Thus, the final product rating is obtained by taking the average of the ratings till the nth week.
Data Structures used:Arraylist

Algorithm:

Step 1: Start

Step 2: Accept review_ratings Ci from sentiment identification

Step 2: For each Ci

Sum = ∑review_ ratings {c1, c2, c3…}

Average {pi} =sum/total no. of reviews in that time period

Step 3: Rate the product Pi according to its average review rating and timestamp

Step 4: End

3.3.4 Rating Predictor

Purpose: The purpose of Rating Predictor is that it make use of Time Series Technique to predict the success or failure of the product. Time Series is a collection of data points collected at constant time intervals and are very frequently plotted via line charts. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, astronomy, communications engineering and largely in any domain of applied science and engineering which involves temporal measurements. These are analysed to determine the long term trend so as to forecast the future or perform some other form of analysis. Using this algorithm, the Rating Predictor module calculates the rating for the nth week using the records of the previous weeks. This results in the prediction of the success and failure of the product.
Data Structures used:Arraylist

Algorithm:

Step 1: Start

Step 2: Accept the product_rating Pi of each timestamp

Step 3: For each product_rating {p1, p2, p3…}

If pi-1=<pi then

Product will be a success

Else

Product will fail

Step 4: End

3.4 UML diagrams with discussions

The sequence diagram is represented below in figure 3.3 in which the Vendor connect with the amazon websites and the response is provided by the amazon websites. Interface provides the vendor with a product name entry option as well as the asin number of the particular product.The vendor press the submit button and the scraping of the reviews takes place and are stores in the json file. The review in the file are used for further pre-processing of each review and sentiment identification are carried out and the rating of the product are generated and provided to the vendor then further prediction are done for the success and failure of the product and the final chart are displayed to the vendor.

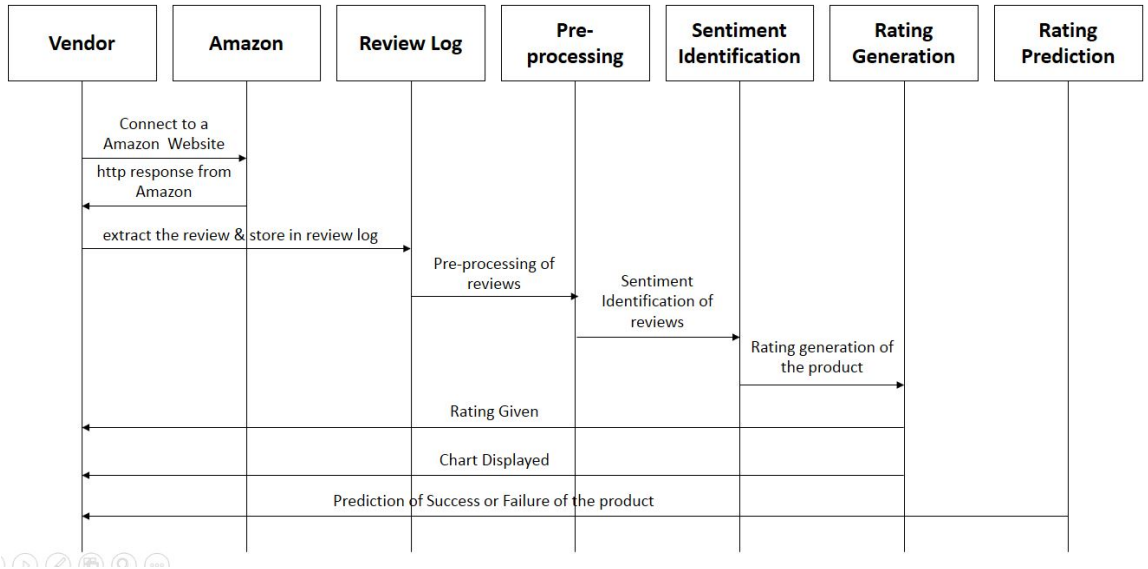


Figure 3.4: Sequence Diagram of the Project

3.5 Data Source

E-commerce companies like amazon, flipkart, snapdeal etc allows user to post real time reviews about the products. Due to these service, people use acronyms, make spelling mistake, use emotions and other characters that express special meanings. Following is a brief terminology associated with reviews.

- Emoticons: These are facial expressions: pictorially represented using punctuation and letters; they express the userâĂŹs mood
- Hashtags: Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their reviews.

3.6 Data Formats

In this project we use the dataset provided by the amazon Application Programming Interface(API). The dataset consists of title of the product, url from which it is extracted, review posted by the customers about the product and date on which this review was posted.Figure 4.5 shows a snapshot of a dataset usedd in the project. This dataset are used for processing and analysis of the product rating as well as the it predict the success and failure of the product based on the reviews.

Figure 3.5: Dataset format used

Chapter 4

Implementation

4.1 Tools and Technologies

In this project we use few tools and technologies like web scrapper, natural language tool kit, etc. which are used to extract the reviews and process it.

4.1.1 Web Scraper

Web scraping or web data extraction is the tool used for extracting data from websites. This software access the World Wide Web directly through a web browser.This process is an automated process which are implemented by using web crawler. In this the data are extracted and copied from the web for later analysis. Scraping a web page includes fetching it and extracting from it. Fetching process is done by browser, when you view the page whereas web crawling plays a major role in web scraping. The content of a page will be searched, parsed, reformatted and data are copied into a spreadsheet and so on. Latest forms of web scraping includes extarcting the data feeds from web servers. For example, JSON is commonly being used as a transport storage mechanism between the client and the web server. In this project we are using this tool to extarct the product reviews from amazon websites and storing it into the JSON file for further processing.

Figure 4.1: Diagram of web scraper tooll

4.1.2 Natural Language Toolkit

The Natural Language Toolkit (NLTK) is a suite of libraries and programs for symbolic and statistical natural language processing(NLP) for the Python programming language. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems. While NLTK comes with a number of corpora that have been pre-processed(often manually) to various degrees, conceptually each layer relies on the processing in the adjacent lower layer. Tokenization comes first; then words are tagged; then groups of words are parsed into grammatical units can be classified. Along the way, NLTK gives you the ability to generate statistics about occurrences of various elements, and draw graphs that representeither the processing itself, or statistical aggregates in results.

Steps to download and install NLTK

- Install Python: https://www.python.org/download/releases/2.7/
- Install Numpy(optional):http://sourceforge.net/projects/numpy/files/NumPy/1.8.1/num1.8.1-win32-superpack-python3.4.exe/download
- Install NLTK:https://pypi.python.org/pypi/nltk
- Test the installation

4.1.3 Bootstrap

Bootstrap is a free and open-source front-end web framework for designing websites and web applications. It contains HTML and CSS based design templates for forms, buttons, navigation, etc ,.Bootstrap 3 supports the latest versions of the Google Chrome, Firefox, Internet Explorer, Opera and Safari. The bootstrap supports responsive web design. This means the layout of web pages adjust dynamically, taking into account the characteristics of the device used(desktop, tablet, mobile phones).

4.1.4 Flask framework

Flask is a micro web framework written in Python which does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common function.However flask supports extensions that can add application features as if they were implemented in Flask itself.

4.2 Experimental Setup

4.2.1 System Requirements

- Hardware Requirements
  - Processor Pentium-dual core or equivalent or more
  - 1GB RAM or more

- Software Requirements
  - Operating System: Ubuntu version 12.02 or above
  - Python 2.7 or above version
  - Python Modules

Few of the resources and libraries of the python platform that has been used for processes such as tokenization, POS Tagging, SVM Classification are stated as follows:

- BeautifulSoup Python, the library provides an interface for crawling the web-page. We use this for crawling the review of the product.
- Review NLP, it is a specific review tokenizer and tagger.It provide a fast and robust Python-based btokenizer and part-of-speech tagger for Review.
- Scikit Python Library, for Naive Bayes Classifier.
- LibSVM is a integrated software for support vector classification. It supports multiclass classification.

4.3 Coding Standards followed

General coding standards mentioned in python open source organization has been used in the project. Some of them are listed below:

4.3.1 Packages and Import statement

Python has only one type of modules object, and all modules are of this type. To help organize modules and provide a naming hierarchy, Python has a concept of packages.The regular packages used are:

- The init file can contain the same python code that any other module can contain, and Python will add some additional attributes to the module when it is imported.
- A namespace packages is a composite of various portions, where each portion contributes a subpackages to the parent packages.

Imports are always put at the top of the file, just after any module comments and docstrings, and before module globals and constants.
Imports should be grouped in the following order:

- standard library imports
- related third party imports
- local application library specific imports
- You should put a blank line between each group of imports.

The import statements are listed below:

- The importlib module provides a rich API for interacting with the import system.
- The import nltk import the nltk module which are used in text processing, tokenization, steaming, tagging, parsing, etc,.

4.3.2 Indentation

We use 4 spaces per indentation level. We never use tabs. We use 8 spaces indents for line wraps, including function calls and assignments. For flowing long blocks of text with fewer structural restrictions the line length should be limited to 72 characters. Add a comment, which will provide some distinction in editors.

4.4 Code Integration details

The project has been divided into 5 modules and the net outcomes that is seen at the end is the result of the integation of the following modules.
This modules are integrated with a following python code given below. It has an init.py file which makes the integration possible and the python files which are to be integrated should be in the same folder. A python file can use the functions or methods of other python file by importing, it use import keyword and then the imported python file can be used in that python file.

```
#import amazonscrapt
from analyzer import Sentiment
import rough
import ratinggen
import jsut
import sqlite
import createtable
import predict
import tocsv

myObj = Sentiment()


sqlite.main()
createtable.main()
myObj.main()
rough.main()
ratinggen.main()
jsut.main()
predict.main()
tocsv.main()
```

Figure 4.2: Python Code for Integrating the modules

4.4.1 Review Log

This is the first module which is integrated. It is a json file which contains the scraped reviews from the amazon websites which are used for further processing.

4.4.2 Pre-processor

This is the second module which ia integrated. It process the review from the review log file and it removes the unwanted characters and make the review clean by the process called tokenization, steaming and filtering.

4.4.3 Sentiment Identifier

This is the third module which is integrated. It is used to classify the reviews into three groups positive, negative and neutral based on the sentiment values.

4.4.4 Rating Generator

This is the fourth module which is integrated.The Rating Generator module generate the rating of the particular product.

4.4.5 Rating Predictor

This is the final module which is integrated. Rating predictor uses Time Series model to predict the success and failure of the product in the coming future.

4.5 Implementation work flow

A particular approach is being followed in making the project. Analysis has been done in several steps and the output are displayed. Diagram 4.3 shows the work flow of the project which consist of following:

- Web scraper: It scraps the review of the products from the amazon websites.
- Review log: The scraped reviews are stores in the json file for further processing.
- Pre-processing: The review in the file contains noisy characters which are processed and cleaned by using different process like tokenization, steaming and filtering.
- Sentiment Identification: This module is used to classify the reviews into positive, negative or neutral based on sentiment values.
- Results: The final result is displayed as a charts which contains the rating generated of the product and prediction of success and failure of the product.

4.6 Execution Results and Discussions

The result has been depicted in the form of a graphs that represent the rating of the product of a week.
In the figure 4.4 we can see that in first week the rating is around 5.8 and it has slightly decreased in the third week and again it has been increased to 7.5 and then later it remained saturated.

4.7 Non-functional requirements results

4.7.1 Reliability

The system should meet all the functional requirements without any unexpected behaviour. The result should not be incorrect or outdated and free from errors.

4.7.2 Availability

The system will be available at all times to the customer through an interactive web application. The functionality of the system will depend on the external services such as internet access. If those services are unavailable, the user should be alerted.

4.7.3 Privacy

The system should never collect and disclose any personal information of the customers.

4.7.4 Maintenance

The software should be written clearly and concisely. The code should be well documented. Particular care will be taken to design the software modularity to ensure easy maintenance.

4.7.5 Portability

The system will be designed to run on any operating system using a thin client such as web browser. The system will be forward compatible for all currently released operating system.

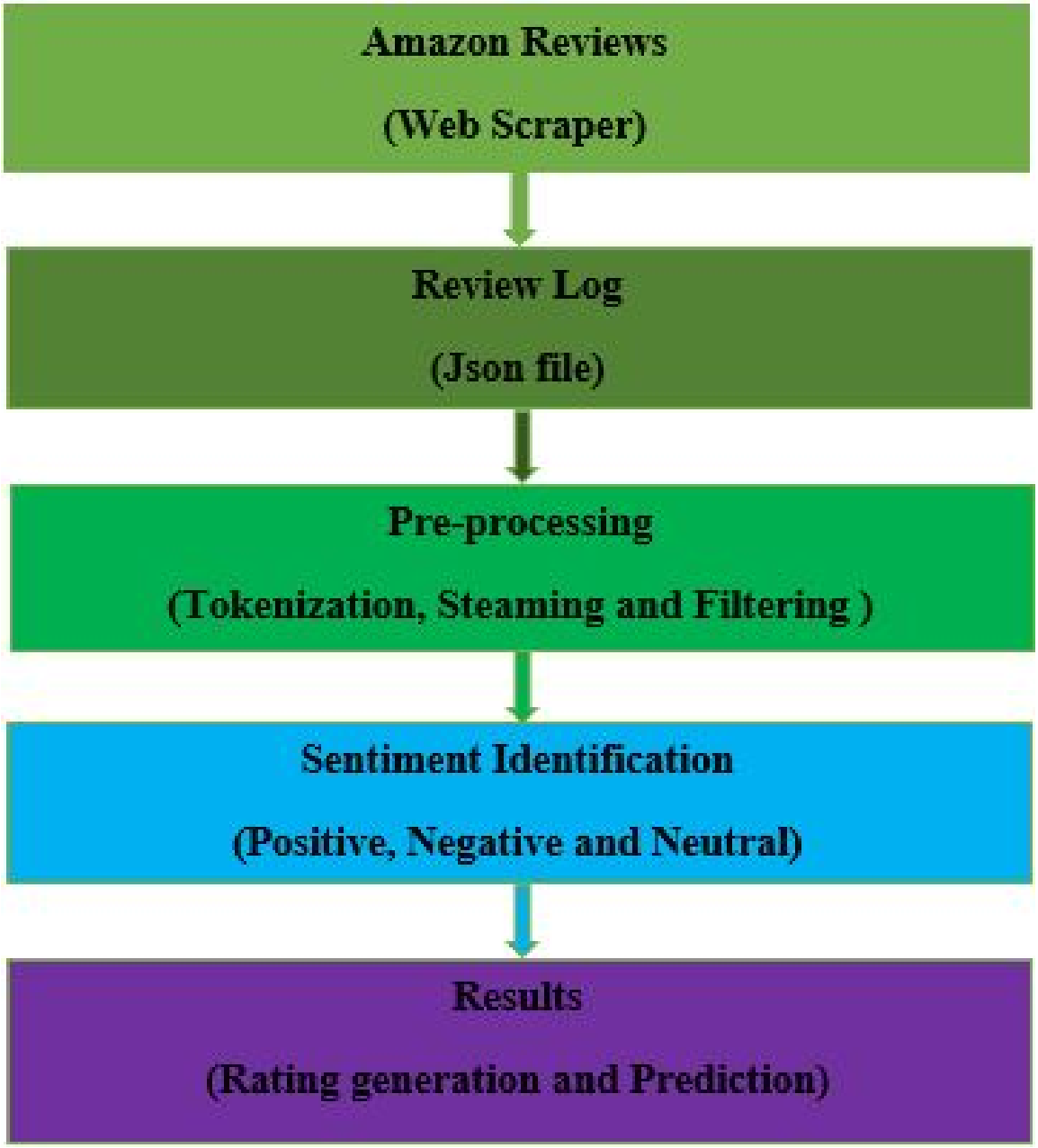


Figure 4.3: Approach used to implement the project

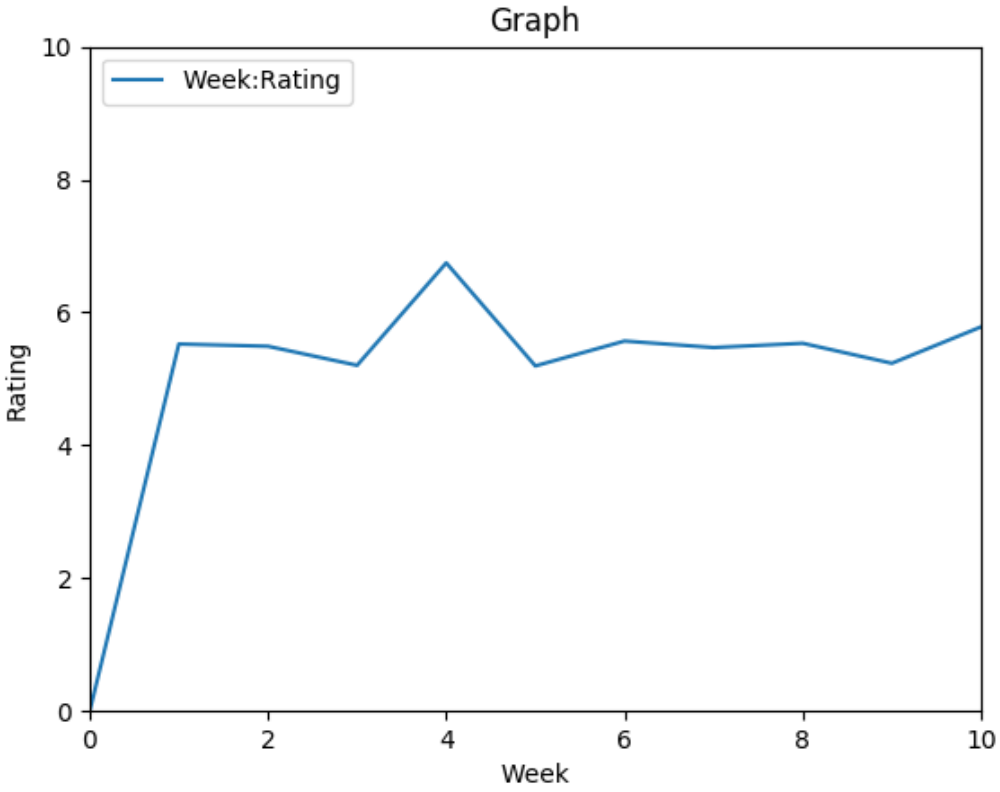


Figure 4.4: Rating Generated Graph

Chapter 5

Testing

The project requires analysis of around 2480 reviews on IPhone 5s scrapped from Amazon website. Since the huge dataset has to be processed in six steps, the complete execution is time consuming. Our testing is integrated in the execution steps itself, so there is no need for explicit testing and building the test cases. Under ideal circumstances, the result should be 100 percent accurate. Since, every human sentiment cannot be guessed accurately, ideal condition cannot be achieved. Different testing modules and their accuracy obtained is mentioned below.

5.1 Test workflow

The workflow that we followed for testing purpose is the incremental model in which each review is tested for its rating.

5.1.1 Unit Testing

Unit testing is a software development process in which the smallest testable parts of an application, called units, are individually and independently certified for proper operation. In our project, every word in the review has a role to play in finding the overall rating of the sentence.

5.1.2 Integration Testing

Integration testing is a phase in software testing in which individual software modules are combined and tested as a group. It is carried out after unit testing and before validation testing.

5.1.3 Validation Testing

Validation testing is the phase in software testing where we test that the software meets the required specification and it fulfills its intended purpose. As in the case of our project, we are trying to generate a rating for each review. The rating is then calculated on a weekly basis and a line graph is generated to analyze the trend of the product. Our final result should be a readable output which will validate our application.

5.1.4 Output Testing

After performing the validation testing, the next steps proceeds to output testing of the proposed system. Output testing plays a vital role since no system is useful if it does not generate the required result that meets the specified objectives.

5.2 Test case details

5.2.1 Test case id:01

Unit to test:The dataset format that the framework can handle.
Assumptions: Scrapped reviews are informative and every review is useful for analysis.
Test data: Real-time dataset on IPhone 5s from Amazon website.
Steps to be executed: Scrapping all the reviews and storing it in a file.
Expected result: Structured dataset in json format.
Actual result:Dataset in json format including the title and review.
Pass/Fail:Pass
Comments: It is a slow process as scrapping each review takes time.

5.2.2 Test case id:02

Unit to test: The list of words generated by the pre-processor.
Assumptions:Each review contains words that are useful for sentiment analysis.
Test data: The dataset in json format.
Steps to be executed:Each review is tokenized, tagged with the parts of speech, base word is extracted and filtered.
Expected result:The preprocessed list contains only the words that are useful for sentiment and stored in a json file.
Actual result:The preprocessed list contains only the verbs, adverbs and adjectives that are necessary for sentiment analysis. The obtained lists are stored in a json file.
Pass/Fail:Pass
Comments:: The review which do not contain any useful words result in an empty list.

5.2.3 Test case id:03

Unit to test: The classifiaction of review.
Assumptions:Each review have its sentiment values which are used to classify based on that values.
Test data: The review stored in the array for identification.
Steps to be executed:Each review is taken from the array and compared with the bag of words and the sentiment values is compared and the review were classified.
Expected result:The classification of review in positive, negative and neutral.
Actual result:The review were classified into positive, negative and neutral.
Pass/Fail:Pass
Comments:: The review classified are the result of the sentiment analysis.

5.2.4 Test case id:04

Unit to test: The generation of the review.
Assumptions:Each review has a rating and it is informative.
Test data: one week of reviews.
Steps to be executed:Each review is stored in sentiment database which contain date and time when the review were posted and review id.
Expected result:The generation of rating of the product.
Actual result:The rating of the product were displayed.
Pass/Fail:Pass
Comments:: The rating generated helps the customer to know the quality of the product of viewing the rating of the product directly instead of viewing the numerous review posted by the user.

Chapter 6

Conclusions and Future Scope

6.1 Conclusion

Sentiment analysis is an emerging field and we have made a small attempt to work on this field to extract knowledge from huge volume of data entered by the customers in the E-commerce websites. In this project, the reviews are collected from the Amazon E-commerce website and stored in a file. Various sentiment analysis operations are performed on the file to classify the reviews as positive or negative and generate the rating on a weekly basis. The final rating is then calculated and the success or failure of the product is predicted. This project is targeted for the customers who buy products online as well as the for the vendors who wish to see the future trend of their product.

6.2 Future Scope

There are several future works that can be proposed with some of the being improving the performance while others can be built on top of the work done here. Here are some of the works we believe can be performed.

- The methods can be improved to achieve better performance ratio.
- Use other supervised machine learning classifier other than NaÃŕve Bayes classifier.
- To extend the system to support analysis of products from other E-commerce websites other than Amazon.
- To extend the project in terms of storage for storing huge number of reviews in cloud.