A

MINI PROJECT REPORT

ON

Amazon Product Sentiment Analysis

Submitted in partial fulfillment of the

requirements of the degree of

Bachelor of Engineering

In

Computer Engineering

By

CO34 Asavari Shejwal
CO23 Sahil Sheikh
CO37 Sailee Shirke

Under the guidance of

Prof. Pragati

Chandankhede

Department of Computer Engineering

K.C. College of Engineering and Management Studies

and Research, Thane (E)

Mumbai University

2020-2021

ACKNOWLEDGEMENT

No project is ever complete without the guidance of those experts who have already traded this

past before and hence become master of it and as a result, our leader. So, we would like to take this opportunity to take all those individuals who have helped us in visualizing this project.

We express our deep gratitude to our mini project guide Prof. Pragati Chandankhede for providing timely assistance to our query and guidance that she gave owing to her experience in this field for the past many years. She has indeed been a lighthouse for us in this journey.

We extend our sincerity appreciation to all our Professors from K.C. College of Engineering & Management Studies & Research for their valuable suggestions during the designing of the project. Their contributions have been valuable in so many ways that we find it difficult to acknowledge them individually.

We also express our gratitude to our HOD Prof. Mandar Ganjapurkar for extending his help directly and indirectly through various channels in our project work. Thank you.

CERTIFICATE

This is to certify that the report on mini project entitled,

Amazon Products Sentiment Analysis

Project Members

CO34 Asavari Shejwal
CO33 Sahil Sheikh
CO37 Sailee Shirke

As a partial fulfillment for Bachelor of Computer Engineering
Degree course of Mumbai University for Academic year 2020-2021

INTERNAL GUIDE

Prof. Pragati
Chandankhede

INTERNAL EXAMINER EXTERNAL EXAMINER

HOD PRINCIPAL

(Prof. Mandar Ganjapurkar) (Dr. Vilas Nitnaware)

ABSTRACT

The world we see nowadays is becoming more digitized. In this digitalized world e-commerce is taking the ascendancy by making products available within the reach of customers where the customer doesn't have to go out of their house. Nowadays people are relying on online products so the importance of a review is going higher. For selecting a product, a customer needs to go through thousands of reviews to understand a product. But in this prosperous day of machine learning, going through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from it. We used supervised learning methods on a large scale amazon dataset to polarize it and get satisfactory accuracy. We did an extensive literature review of the problem and various

methodologies to arrive at its solution. We then decided to work on Amazon Products Sentiment Analysis. The objective of this paper is to categorize the positive and negative feedback of the customers over different products and build a supervised learning model to polarize large amounts of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with a large amount of positive reviews provides a powerful comment on the legitimacy of the item. Conversely, books, or any other online item, without reviews puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing. People value the consent and experience of others and the review on a material is the only way to understand others' impression on the product.

List of Figures

5	Result	13

TABLE OF CONTENTS

Sr.No. Topic

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Rait Pravleen Singh

Rajpurohit ChandanSingh

Harish Unnikrishnan

Date:

1. Introduction

Machine Learning is the process of analysis and study of the data statistics and analytics which can provide raw sources of data to identify the common usage patterns of the users and collect this information to deliver their services to the appropriate users by making effective business decisions.

There are two main approaches to sentiment analysis:

1. Machine Learning Technique:

   Machine learning is interpreted as an important bureau of Artificial intelligence that takes action with the use of practicing a code that allows the system to comprehend. This technique makes use of morphological and expressive features. It examines Sentiment analysis as an issue of periodic text categorization in which we have several documentations for priming and categorization. The prototype is instructed to forecast the class mark for the latest example. Some widely used classifiers include the decision tree, the neural tree network, Naïve Bayes, logistic regression, and Support Vector Machine. We carry out these classifiers using supervised and unsupervised learning.

2. Lexicon Based Technique:

The Lexicon Based approach turns on detecting a point of view. Lexicon incorporates a multitude of words of opinions that are known and precompiled. To analyze the text, this viewpoint lexicon can be used. The lexicon-based approach mostly includes three methods, including the dictionary based method, the corpus-based method, and the manual approach to opinion. This is preferably used with the rest of the two methodologies in combination.

2. Report on present investigation

2.1. Requirement Analysis

2.1.1 Scope

The scope of this document is to give a detailed description of the sentiment analysis project. It will illustrate the purpose and complete deceleration for the development of the system. It will also explain system constraints, interface and interactions. This document is primarily intended for anyone who wants to get an overview of how machine learning and sentiment analysis works, its outcomes and possible usage in future.

2.1.2 Software Requirements

i. Windows or Linux operating system

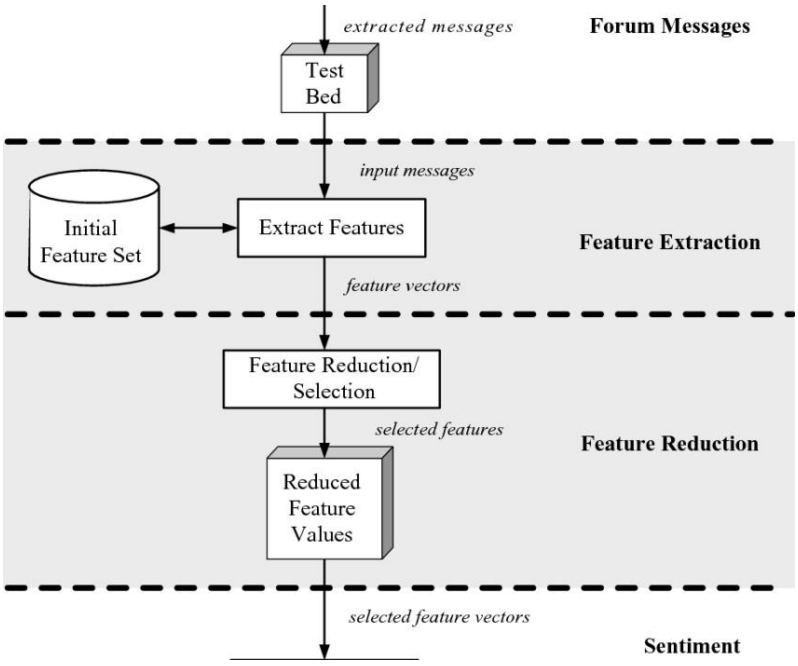
ii. Python: Python 3.x with “nltk” library installed

iii. Web Browser

2.2. Problem Statement:

Sentiment classification aims to determine the overall intention of a written text which can be of admiration or criticism type. This can be achieved by using machine learning algorithms such as Na¨ıve Bayes and Support Vector Machine. So, the problem that is going to be investigated in the project is as follows: Which machine learning approach performs better in terms of accuracy on the Amazon products reviews.

2.3. Project Design



2.4. Methodology

1. DATA COLLECTION: We used data named Amazon product data which was provided by researchers from [14]. The dataset was unlabeled and to use it in a supervised learning model we had to label the data. We used three JSON files where the structure of the data is as follows: "reviewerID": ID of the reviewer "asin": ID of the product "reviewerName": name of the reviewer.

	asin	imUrl	description	categories	title	price	salesRank	related	brand
0	0132793040	http://ecx.images-amazon.com/images/I/31JIPhp%...	The Kelby Training DVD Mastering Blend Modes i...	[[Electronics, Computers & Accessories, Cables...	Kelby Training DVD: Mastering Blend Modes in A...	NaN	NaN	NaN	NaN
1	0321732944	http://ecx.images-amazon.com/images/I/31uogm6Y...	NaN	[[Electronics, Computers & Accessories, Cables...	Kelby Training DVD: Adobe Photoshop CS5 Crash ...	NaN	NaN	NaN	NaN
2	0439886341	http://ecx.images-amazon.com/images/I/51k0qa8f...	Digital Organizer and Messenger	[[Electronics, Computers & Accessories, PDAs, ...	Digital Organizer and Messenger	8.15	{'Electronics': 144944}	{'also_viewed': ['0545016266', 'B009ECM8QY', '...	NaN

reviewerID : ID of the reviewer, e.g. A2SUAM1J3GNN3B - type: object
asin : ID of the product , e.g. 0000013714 – type: object
reviewerName : name of the reviewer – type: object
helpful : helpfulness of the review, e.g. 2/3 – type: object
reviewText : text of the review – type: object
overall : Rating (1,2,3,4,5)– type: float64
summary : summary of the review – type: object
unixReviewTime : time of the review (unix time) – type: int64
reviewTime : time of the review (raw) – type: object

2. PREPROCESSING RAW DATA

In this study, two experiments have been conducted. In the first experiment the whole data set was used. Since the number of reviews were quite enough to get a reasonable result from the classifiers the reviews with three stars were omitted to avoid any complication while training the algorithms. However, in the second experiment due to the small number of data the reviews with three stars were also considered as negative. The same code was then used to label the data. The reviews that were considered as positive got a score of ”1” and the remaining ones got a ”0” score.

Data preparation is divided into two subsections:

i) **Tokenization**: Tokenization is designated as the technique by which a succession of words is separated into individual words, such as names, keywords, expressions, and tokens. Tokens could be phrases, even full sentences of single or individual words. Certain characters are discarded, like

punctuation marks, vowels and many more in the tokenization process. For the distinct course of action such as parsing and text mining, the tokens act as input values.

ii) **Lower case:** Word to lowercase (TREE -> tree) conversion. All of the words are converted into a lowercase letter to bring uniformity and a direction in one way.

iii) **Cleaning Stop Words:** Stop words are terms in an expression that are not required or needed in text analytics in any field. So, to enhance the efficiency of the study, such terms are mostly disregarded. Depending on the country and their culture and language etc., there are various kinds of stop words in different formats. There are some stop words in the English format, so it has to be deleted.

iv) **Stemming:** It is the method of normalizing the word into its root form. Assigning sentiment scores: sentiment scores are assigned for different ratings. A score of 0 to reviews having ratings 1 and 2, score 1 to reviews having ratings 3, and score 2 to reviews having ratings 4 and 5 respectively.

Score 0: negative sentiment Score 1: neutral sentiment Score 2: positive sentiment

3. Feature Extraction

a. Bag of Words: Bag of word is a process of extracting features by representing simplified text or data, used in natural language processing and information retrieval. In this model, a text or a document is represented as the bag (multiple set) of its words. So, simply a bag of words in sentiment analysis is creating a list of useful words. We have used a bag of words approach to extract our feature sets. After the preprocessed dataset we used pos tagging to separate different parts of speech and from that we select nouns and adjectives and use those to create a bag of words. Then we run it through supervised learning and find our results and also the top used words from the review dataset.

b. TF-IDF:TF-IDF: It is an information retrieval technique which weighs a term's frequency (TF) and also inverse document frequency (IDF). Each word or term has its own TF and IDF score. The TF and IDF product scores of a term is referred to the TF*IDF weight of that term. Simply we can state that the higher the TF*IDF score (weight) the rarer the term and vice versa. TF of a word is the frequency of a word.

c. Chi Square: Chi square(X^2): It is a calculation that is used to determine the smaller the difference between the observed data and the expected data . In this approach we have preprocessed our dataset then we have divided data into training and testing sets. We used a pipeline method to apply TF-IDF, Chi square and other classifiers onto our dataset and got the results.

Algorithm for proposed approach :

Input: Labeled Data=labeled data obtained after active learning process.

Output: Accuracy of classifiers; Precision,Recall,F-1 Measure for positive and deceptive values.

//product review polarity accuracy

1. Load labeled data positive & negative
2. Preprocess labeled data
3. for every X= {X1…Xn} in labeled data
4. Extract Feature(Xi)
5. Cross validate into training & testing set
6. Classifier.train()
7. Accuracy= classifier.accuracy()
8. majority_voting(accuracy) using vote classifier
9. show result(accuracy,precision,recall,f1 measure)
10. end extract feature(text) return n-gram feature majority_voting(accuracy) return highest accuracy

2.5. Implementation plan

2.5.1 Semester VIII

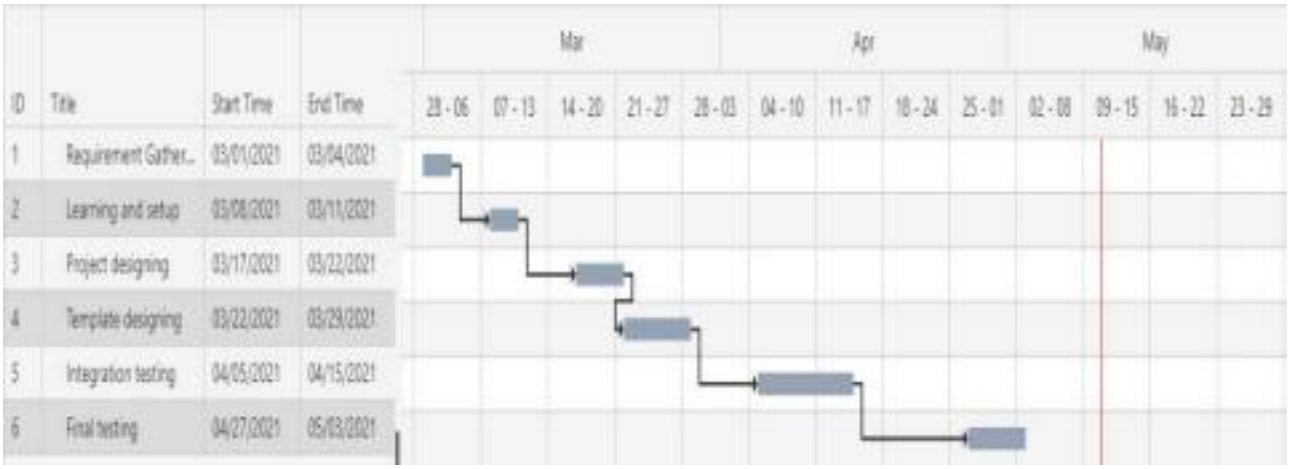


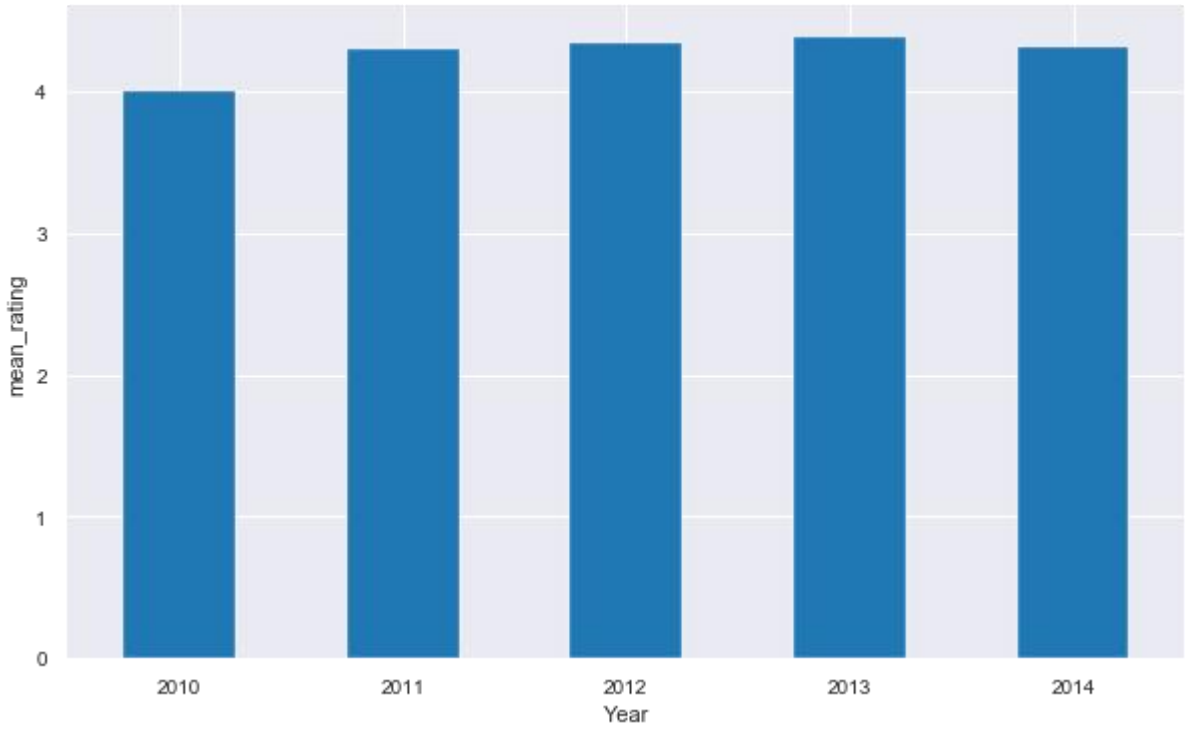
Figure 2.5.1. Implementation Plan

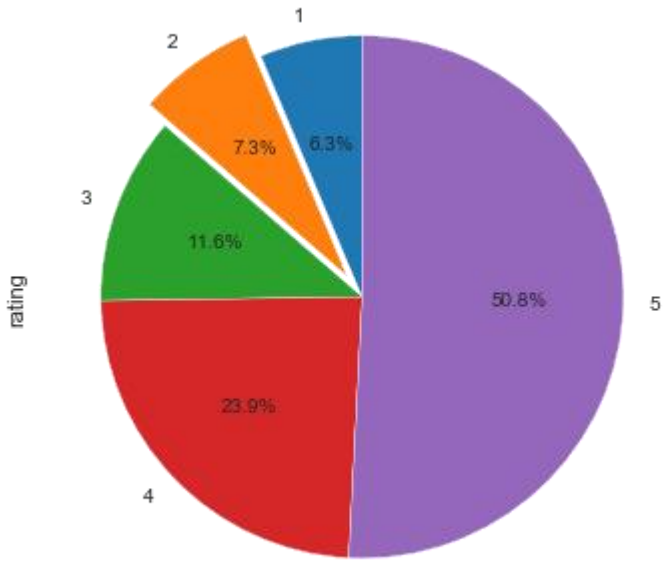
3. Results and Discussion:

In assessing the classification efficiency of a model, evaluating metrics plays an important role. The most widely used technique for this purpose is precision measurement. The efficiency of a sequencer on a given experimenting dataset is the proportion of those datasets that are properly grouped by the sequencer, and the accuracy measure of the text mining method is often not sufficient to provide proper decision or result, so some others metrics should be taken to assess the

output of the classifier. Three other significant indicators widely used are memory, accuracy, and Fmeasurement.

There were several machine learning algorithms used in our experiment like Naïve Bayesian, Support vector Machine Classifier (SVC), Stochastic Gradient Descent (SGD), Linear Regression (LR), Random Forest and Decision Tree. We have conducted cross validation methods and 10 fold gave the best accuracy. We conduct the best classifiers on 3 categories of product reviews and see the results according to the evaluation.





4. Conclusion:

In this research we proposed a supervised learning model to polarize a large amount of product review dataset which was unlabeled. We proposed our model which is a supervised learning method and used a mix of 2 kinds of feature extractor approach. We described the basic theory behind the model, approaches we used in our research and the performance measure for the conducted experiment over quite a large data. We tried different simulations using cross validation, training-testing ratio, and different feature extraction processes for comparing varying amounts of data to achieve promising results. In most of the cases 10 fold provided a better accuracy while Support Vector Machine (SVM) provided best classifying results. It is hard to gather a huge amount of gold standard dataset for this purpose as e-commerce sites have their limitations on giving data publicly. Also scraping data can be a problem as we can't scrape enough data to consider it as real-life public reviews over different products.A radical change from virtual platforms to digitalized platforms can be seen in a new age. The dependence of clients and consumers on online feedback has increased especially. Digital opinions have enhanced a forum for raising belief and shaping the trends of customer purchasing. By performing an opinion analysis of Amazon product checks and categorizing the opinions into optimistic, neutral, and negative feelings, our project aims to accomplish this. Four classification models were used to identify reviews after combining the data with some neutral and negative opinions.

5. References

- https://sci-hub.hkvisa.net/10.1109/ICCCI50826.2021.9402414
- https://ieeexplore.ieee.org/document/9402414
- https://towardsdatascience.com/sentiment-analysis-and-product-recommendation-on-amazons-electronics-dataset-reviews-part-1-6b340de660c2