Sentiment Analysis On Large Scale Amazon Product Review

A. Sunitha M.E, Assistant Professor of Computer Science and Engineering,
Vignesh.K, Student of Computer Science and Engineering,
Vinoth Kumar.B, Student of Computer Science and Engineering,
St. Joseph College of Engineering, Sriperumudur, Chennai.

Abstract :

The world we see now a days is becoming more digitalized. 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. As now a day's 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 prospering 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 method on a large scale amazon dataset to polarize it and get satisfactory accuracy**.** The intense competition to attract and maintain customers online is compelling businesses to implement novel strategies to enhance the customer experiences. It is becoming necessary for companies to examine customer reviews on online platforms such as Amazon to understand better how customers rate their products and services. The purpose of this study is to investigate how companies can conduct sentiment analysis based on Amazon reviews to gain more insights into customer experiences. The dataset selected for this capstone consists of customer reviews and ratings from consumer reviews of Amazon products. Amazon product reviews enable a business to gain insights on customer experiences regarding specific products and services.

Introduction:

As the commercial site of the world is almost fully undergone in online platform people is trading products through different ecommerce website. And for that reason reviewing products before buying is also a common scenario. Also now a day,customers are more inclined towards the reviews to buy a product. So analyzing the data from those customer reviews to make the data more dynamic is an essential field nowadays. In this age of increasing machine learning based algorithms reading thousands of reviews to understand a product is rather time consuming where we can polarize a review on particular category to understand its popularity among the buyers all over the world. The objective of this paper is to categorize the positive and negative feedbacks of the customers over different products and build a supervised learning model to polarize large amount of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations

Objectives:

- The aim of the study is to classify customer reviews into positive or negative sentiment.
- To measure the intensity of the sentiments generated from the customer reviews.
- To analyze the association between customer reviews concerning different amazon products.

Literature Survey:

1.Sales Prediction using Online Sentiment with Regression Model

Abstract

- The main aim of this paper is to predict the sales of a vehicle using sentiment analysis from various places on the internet. The online presence of a vehicle , as well as its brand, plays a key role in the sales of the vehicle. However, many other parameters are required and will be discussed in this paper. Sales prediction in today's market is not only beneficial for the manufacturer but also for the various other companies that manufacture parts or accessories for vehicles. It can also be considered as a boon for retailers, showroom owners, and service mechanics. In this application, we have made use of linear regression for sentiment analysis and polynomial regression for sales prediction.

2.Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics

Abstract:

- With the rapid growth of the Internet, the ability of users to create and publish content has created active electronic communities that provide a wealth of product information. However, the high volume of reviews that are typically published for a single product makes harder for individuals as well as manufacturers to locate the best reviews and understand the true underlying quality of a product. In this paper, we reexamine the impact of reviews on economic outcomes like product sales and see how different factors affect social outcomes such as their perceived usefulness.

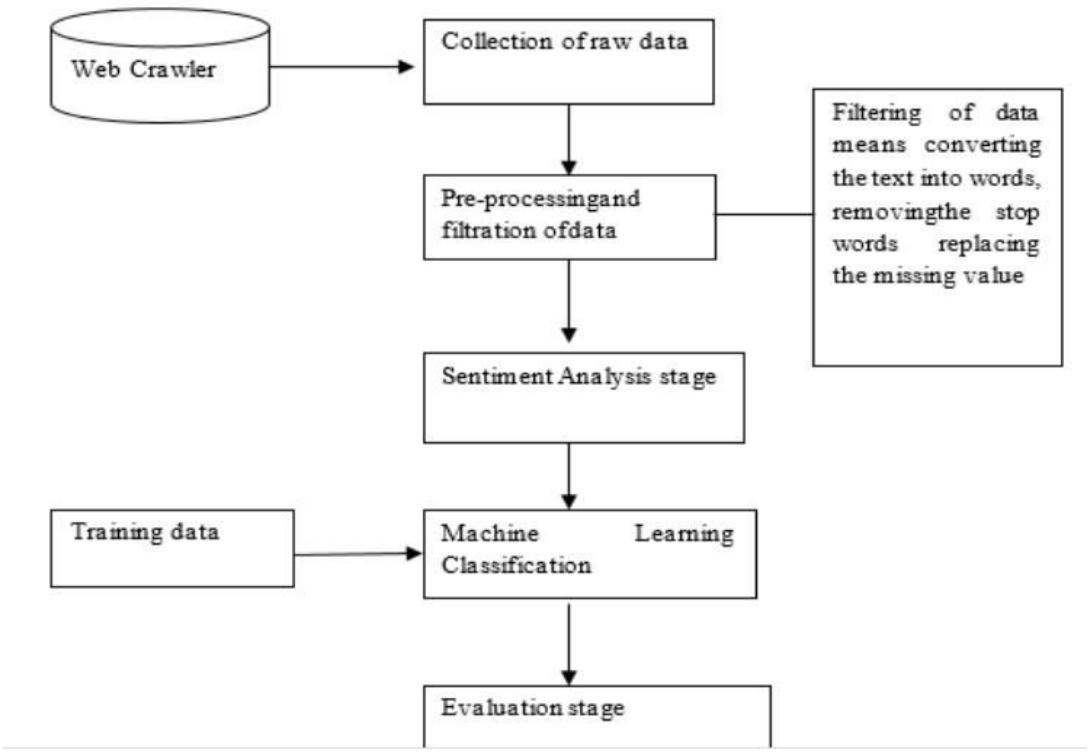
Existing system:

- So far, much of the research papers related to product reviews,sentiment analysis or opinion mining has been done recently. They collected Amazon dataset at first and then performed preprocessing for stop words and special characters' removal. They applied phrase level, single word and multiword feature selection or extraction technique. Naive Bayes is used as the classifier.

Proposed system:

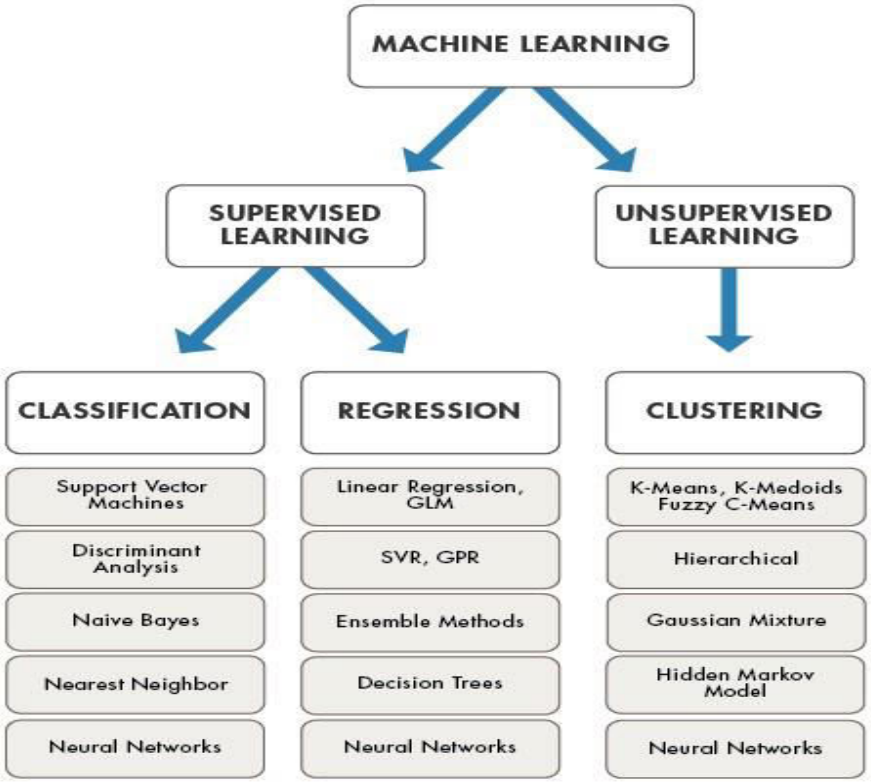
- Sentiment analysis methodology will be used in this research. The method is also referred to as opinion mining. The process relies on machine learning (ML) algorithms and natural language processing (NLP) to determine the emotions behind online reviews. The research will focus on analyzing the sentiments on Amazon product reviews

Architecture Diagram:



Modules:

- Data collection
- Data cleaning
- Data Pre-processing
- Data preparation
- Feature Extraction
- second stage filtering
- Model training
- Performance Evaluation
- Prediction



Classification Diagram

Data collection:

- We acquired our dataset of 3 different JSON formats and labeled our dataset. As we have a large amount or reviews manually labeling was quite impossible for us. Therefor we preprocessed our data and used Active learner to label the datasets.

Data preparation:

- After completing the pre-processing, social media posts are transformed into vectors to generate the feature vectors. The vectors are used in the learning phase for machine learning algorithms.

Data cleaning:

- In this module data cleaning is done to prepare the data for analysis by removing or modifying the data that may be incorrect, incomplete, duplicated or improperly formatted.

Data Pre-Processing:

- **Tokenization:** It is the process of separating a sequence of strings into individuals such as words, keywords, phrases,symbols and other elements known as tokens. Tokens can be

individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens work as the input for different process like parsing and text mining.

Feature Extraction:

- 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

Algorithm:

Random forests

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.
- "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset
- " Instead of relying on one decision tree, the random forest takes the prediction"

Natural language processing(NLP)

- NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence. It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks such as translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation
- Now, modern NLP consists of various applications, like speech recognition, machine translation, and machine text reading.
- **Intentions of Using Emojis:**
- Expressing sentiment: same as facial expressions during face-to-face communication, emoji usage in order to express sentiments or emotions, such as anger, happiness, fear, and so on.
- Strengthening expression: using emojis to strengthen their expression,for instance, adding an angry face emoji by the end of a text in order to express a more negative sentiment.

Future enhancement:

- Finally, as we conclude that our algorithm would work upon the user reviews about the product that is 'good' or 'bad' accordingly. The work which could be undertaken in the

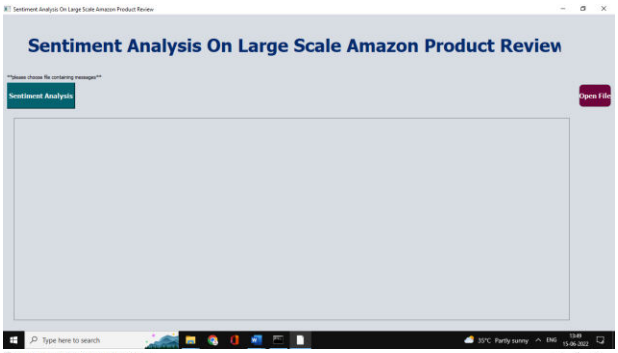
future is that, we can predict why the product unconsidered ‘bad’ or ‘good’ based on the negative/ positive reviews. We could design an algorithm based on products which could link these mixed reviews with a keyword of the specific product attributes.

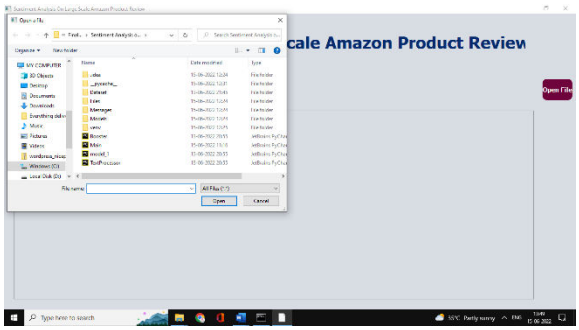
SOURCE CODE:

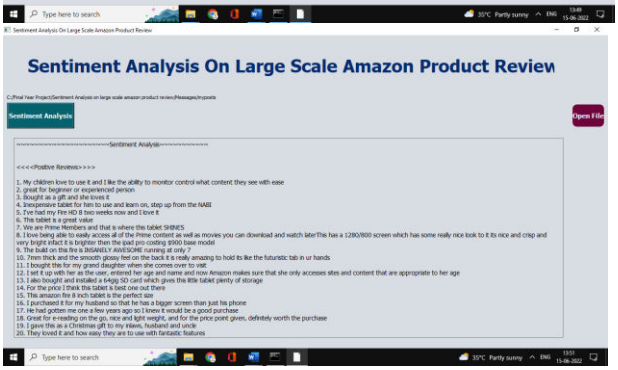
```
import re
class DetectorBooster:
    def _init_(self,filename):
        self.file = open(filename,"r")
        self.key_words = []

    def detect(self):
        crime_sentences = ""
        self.extractor()
        tot = 0
        # sentences =  self.file.read().split(".")
        sentences = re.split('\.|\?|!', self.file.read())
        # print(sentences)
        for s in sentences:
            s = s.strip()
            if len(s) > 4: # meaningful sentence
                ar_s = s.split(" ")
                for word in  ar_s:
                    if word.lower() in self.key_words:
                        crime_sentences += s + "."
                        break
                tot+=1
        my_list = crime_sentences.split(".")[:-1]
        crime_rate = len(my_list) /  tot
        return my_list, crime_rate
    def extractor(self):
        att = open("Files/Positive_Review.txt", "r")
        for line in att:
            self.key_words.append(line.lower().strip())
        dg = open("Files/Negative_Review.txt", "r")
        for line in dg:
            self.key_words.append(line.lower().strip())
class AnalyzorBooster:
    def _init_(self,filename):
        self.Positive_key_words = []
        self.Negative_key_words = []
        self.file = open(filename,"r")
        self.key_words = []
   def analyse(self):
        Negative_sentences = ""
        Positive_sentences = ""
        self.extractor()
        tot = 0
        # sentences =  self.file.read().split(".")
        sentences = re.split('\.|\?|!|\n', self.file.read())
```

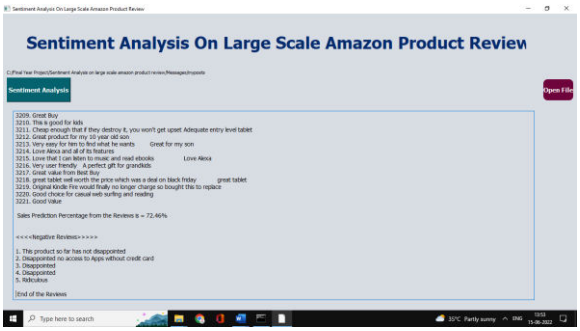
```
    for s in sentences:
        s = s.strip()
        if len(s) > 4: # meaningful sentence
            ar_s = s.split(" ")
            for word in  ar_s:
                if word.lower() in self.Negative_key_words:
                    Negative_sentences += s + "."
                    break
                elif word.lower() in self.Positive_key_words:
                    Positive_sentences += s + "."
                    break
                tot+=1
        my_list = Positive_sentences.split(".")[:-1]
        my_list_2 = Negative_sentences.split(".")[:-1]
        Positive_rate = len(my_list) /  tot
        Negative_rate = len(my_list_2) /  tot
        return my_list,my_list_2, Positive_rate, Negative_rate
    def extractor(self):
        att = open("Files/Positive_Review.txt", "r")
        for line in att:
            self.Positive_key_words.append(line.lower().strip())
        dg = open("Files/Negative_Review.txt", "r")
        for line in dg:
            self.Negative_key_words.append(line.lower().strip())
```

Snapshots:









CONCLUSION:

- Posting online reviews is a contemporary approach for the public to share their sentiments and reviews, and unfussiness intelligence, it is more useful. Posting online surveys is a modern approach for the public to share their reviews and in business intelligence, it's more beneficial.
- Using random forest, we summarize the sentiment data from online reviews and feedback, using sentiment information and past product sales results, we used the random forest model to predict sales performance of products and we will achieve the greater predictive accuracy of sales with the help of this model. Finally, we would conclude that the product is 'good' or 'bad'.

Biography:

Mrs.A.Sunitha M.E., is an Assistant Professor in the Department of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, Tamil Nadu. She has completed his M.E, CSE under Anna University Affiliation College in the year 2014. she has done his B.E, CSE under Anna University Affiliation College in the year 2012. Mrs.A.Sunitha has 07 years of teaching experience and 5 publications in International Journals and Conferences. Her area of interests includes Network Security, Computer Networks, Data Science and Machine Learning. She is an active member of CSI and IEANG. She has organized various International Conferences, workshops and Seminars in the area of Computer Networks, Cloud Computing & Machine Learning respectively.

Mr. K. Vignesh B.E., Student in the stream of Computer Science and Engineering atSt.Joseph College of Engineering, Sriperumbudur,Chennai, Tamilnadu .I had started a Ecommerce Start-up in my 3rd year College Days , Now the Start-up in the R&D Phase, Now I'm the Web Developer of the Start-up Company as Drawlead is owned by me.

Mr. B. Vinoth Kumar B.E, Student in Stream of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai, Tamilnadu. I had Founded a Ecommerce Start-up as Mart2020 in my 3rd year college Days, Now the Start-up in the R&D Phase. Now We Founded the Drawlead as a Web development and Digital Marketing Company to Provide a Growth Strategy for the Small Scale Business and Start-up Companies.