**SAVITRIBAI PHULE PUNE UNIVERSITY**

**A**

**MINI-PROJECT REPORT**

**ON**

**"Sentiment Analysis of Customer Reviews"**

***(A project to fulfil the requirements of DSBDA Lab)***


**By**

**VEDANT MORE 307B075**


Under the guidance of

**(Prof. S. S. Shinde)**


Sinhgad Institutes


**DEPARTMENT OF INFORMATION TECHNOLOGY**

**SINHGAD COLLEGE OF ENGINEERING, PUNE**


**(Academic Year: 2022-2023)**

Sinhgad Institutes

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**SINHGAD COLLEGE OF ENGINEEING, PUNE**

## CERTIFICATE

This is to certify that final project work entitled "**Sentiment Analysis of Customer Reviews**" was successfully caried by Vedant More, In the partial fulfilment of the DSBDA Lab course during Semester-II of Third Year of Information Technology prescribed by the SAVITRIBAIPHULE PUNE UNIVERSITY, PUNE.

**Guide** **Head of Department**

**(Prof. S. S. Shinde)** **(Dr. S. R. Ganorkar)**

**Principal**

**(Dr. S. D. Lokhande)**

## Acknowldegement

I feel great pleasure in expressing my deepest sense of gratitude and sincere thanks to my guide Prof. S. S.Shinde for their valuable guidance during the Project work, without which it would have been very difficult task. I have no words to express my sincere thanks for valuable guidance, extreme assistance and cooperation extended to all the Staff Members of department of Information Technology.

This acknowledgement would be incomplete without expressing my special thanks to Dr. S. R. Ganorkar Head of the Department (Information Technology) for their support during the work. I would also like to extend my heartfelt gratitude to my Principal, Dr. S. D. Lokhande who provided a lot of valuable support, mostly being behind the veils of college bureaucracy.

Last but not least I would like to thanks all the Teaching, Non- Teaching staff members of my department, my parent and my colleagues those who helped me directly or indirectly for completing of this Project successfully.

Name of Student

**VEDANT MORE 307B075**

# ABSTRACT

This report presents a comprehensive analysis of sentiment in customer reviews on products using a data science approach. The project aimed to develop a model that could automatically classify customer reviews as positive, negative, or neutral, providing valuable insights into customer sentiments. The project encompassed data collection, preprocessing, feature extraction, model training, and evaluation. The results showcased the effectiveness of the sentiment analysis model in accurately categorizing sentiments expressed in customer reviews. The findings highlight the importance of sentiment analysis in understanding customer perceptions and guiding businesses in improving their products and services based on customer feedback. The report concludes with suggestions for future enhancements, including the incorporation of advanced techniques and scaling the model to analyze larger volumes of real-time customer reviews.

With an ever-increasing demand of e-commerce platforms, there online market is expanding at an exponential pace. With such a boom in the e-commerce industry, there is a need to realize the holistic review of the brand and the model of different products. There are numerous brands present in the market, out of which some are dominant and occupy quite a big part of the industry. Reviews available on such e-commerce platforms act as a guiding tool for the consumers to make informed decisions. Retail websites like Amazon.com or Flipkart offer different options to the reviewers for writing their reviews. For instance, the consumer can provide numerical rating from 1 to 5 or write comments about the product Understanding the data better is one of the crucial steps in data analysis. In this report we will analyze and study the sentiment analysis of Amazon reviews dataset using nltk and vader tools and libraries.

# TABLE OF CONTENT

- **Problem Statement:**

E-commerce platforms like Amazon and flipkart gives a platform to small businesses and companies with modest resources to grow larger. And because of its popularity, people actually spend time and write detailed reviews, about the brand and the product. So, by analyzing that data we can tell companies a lot about their products and also the ways to enhance the quality of the product. But that large amount of data cannot be analyzed by a person.

# 1. INTRODUCTION

In this report, we present the findings of a sentiment analysis project on customer reviews of products. The goal of this project is to develop a model that can classify customer reviews as positive, negative, or neutral based on their sentiment. This can help businesses understand the sentiment of their customers towards their products and services and identify areas for improvement. Sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also known as opinion mining, studies people's sentiments towards certain entities. Internet is a resourceful place with respect to sentiment information. From a user's perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites.

Sentiment analysis is a multidisciplinary field, including psychology, sociology, natural language processing, and machine learning. Recently, the exponentially growing amounts of data and computing power enabled more advanced forms of analytics. Machine learning, therefore, became a dominant tool for sentiment analysis. For the business domain, sentiment analysis plays a vital role in enabling businesses to improve strategy and gain insight into customers' feedback about their products. There is an abundance of scientific literature available on sentiment analysis, and there are also several secondary studies conducted.

### 1.1: Motivation

The explosive growth of discussion platforms, product review websites, e-commerce, and social media facilitates a continuous stream of thoughts and opinions. This growth makes it challenging for companies to get a better understanding of customers' aggregate opinions and attitudes towards products. The explosion of internet-generated content coupled with techniques like sentiment analysis provides opportunities for marketers to gain intelligence on consumers' attitudes towards their products. Extracting sentiments from product reviews helps marketers to reach out to customers who need extra care, which will improve customer satisfaction, sales, and ultimately benefits businesses.

### 1.2 Project Overview

The purpose of this project is to perform sentiment analysis on customer reviews of products. The goal is to develop a model that can classify customer reviews as positive, negative, or neutral, providing valuable insights into customer sentiments towards different products. This analysis can assist businesses in understanding customer feedback, identifying areas of improvement, and making data-driven decisions to enhance customer satisfaction.

### 1.3 Objectives

Collecting and preprocessing of a dataset of customer reviews on products, performing exploratory data analysis to gain insights into the dataset. Developing a sentiment analysis model using machine learning techniques. Evaluating the performance of the model and interpret the results and providing recommendations based on the analysis to enhance customer satisfaction.

# METHODOLOGY

## 2.1 Data Collection:

The first step was to collect a dataset of customer reviews on products. The dataset was obtained from Kaggle, which includes a large number of reviews across various product categories. The dataset contains textual reviews along with corresponding sentiment labels. Here in this project obtained the dataset of customer reviews from an online e-commerce platform Amazon. The dataset contains over half a million customer reviews labeled as positive, negative, or neutral. The dataset is evenly balanced with reviews in each sentiment category.

## 2.2 Data Preprocessing:

Before conducting the analysis, the dataset underwent preprocessing steps to clean and prepare the data. The following preprocessing steps were performed:

- Removal of irrelevant characters, special symbols, and HTML tags.
- Conversion of text to lowercase for consistency.
- Removal of stopwords and punctuation marks.
- Lemmatization or stemming to reduce words to their base or root forms.

## 2.3 Feature Extraction

Using the Bag-of-Words (BoW) technique to transform the preprocessed text data into numerical features. BoW represents text data as a matrix of word counts, where each row represents a document (customer review), and each column represents a word in the vocabulary. We also used Term Frequency-Inverse Document Frequency (TF-IDF) to weigh the importance of each word in the vocabulary based on its frequency in the document and its frequency across all documents.

# EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis or EDA is a crucial step in the data science process that involves analyzing and visualizing data sets to gain insights, identify patterns, and understand the underlying structure of the data. EDA helps data scientists and analysts to better understand their data before applying more advanced techniques, such as modeling or hypothesis testing. During EDA, various statistical and visual techniques are applied to examine the data. This includes summarizing the main characteristics of the data, identifying missing or erroneous values, assessing data distribution and variability, exploring relationships between variables, and detecting outliers or anomalies. EDA often involves using tools like statistical measures, histograms, scatter plots, box plots, correlation matrices, and other visualizations.

### 3.1 Dataset Overview

An initial analysis of the dataset was performed to gain a better understanding of its structure and characteristics. The dataset consists of nearly thousand reviews, with each review accompanied by a sentiment label (positive, negative, or neutral).

### 3.2 Sentiment Distribution

The distribution of sentiment labels within the dataset was examined. The results indicated that [percentage] of reviews were positive, [percentage] were negative, and [percentage] were neutral. It is important to note any class imbalances, as they may impact model training and evaluation.

### 3.3 Word Frequency Analysis

A word frequency analysis was conducted to identify the most frequent words or phrases within each sentiment category. This analysis provided insights into the key aspects that customers mention when expressing their sentiments towards products. For example, in positive reviews, words like "excellent," "great," and "highly recommend" appeared frequently, indicating customer satisfaction.

# MODEL DEVELOPMENT AND EVALUATION

## 4.1 Model Selection

Choosing an appropriate sentiment analysis model for this task. In this case, in this project we are using the NLTK library and the Vader sentiment analysis tool. Vader is specifically designed for social media sentiment analysis and works well for short, informal texts like customer reviews. Here is some brief information on these tools:

- **NLTK:** NLTK stands for Natural Language Toolkit. It is a popular open-source library in Python specifically designed for natural language processing (NLP) tasks. NLTK provides a wide range of tools, resources, and algorithms for tasks such as tokenization, stemming, tagging, parsing, semantic reasoning, and sentiment analysis. It is a Python library for natural language processing (NLP) tasks. It offers modules for text processing (tokenization, stemming, lemmatization, and normalization), part-of-speech tagging, chunking, parsing, and sentiment analysis. It provides pre-trained models and resources for these tasks, making it a versatile tool for NLP applications.

- **VADER:** VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based sentiment analysis tool specifically designed for social media text. It is a part of the Natural Language Toolkit (NLTK) library in Python. VADER is commonly used in data science for sentiment analysis tasks. VADER is a rule-based sentiment analysis tool used in data science. It assigns sentiment scores to words and combines them to determine the overall sentiment polarity and intensity of text. It considers factors like emoticons, capitalization, and uses a pre-built lexicon to estimate sentiment. VADER is specifically designed for review text analysis.

## HARDWARE REQUIREMENTS

1. **Processor:** Quad core Intel i5 or higher (Dual core is not the best for this kind of work, but manageable)
2. **Computer architecture bit widths**: 64-bits
3. **RAM:** 8 GB (4 GB is okay but not for the performance you may want and/or except)
4. **Hard disk space:** 16 GB
5. **Internet Connection**: Required

## SOFTWARE REQUIREMENTS

1. **Language:** Python
2. **IDE:** Jupiter Notebook

# IMPLEMENTATION

We will start the sentiment analysis of amazon product reviews by importing necessary libraries and the dataset:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('ggplot')
import nltk
```

The original dataset is too huge to handle so we reduce its size to only first five hundred entries of the given dataset:

```
df = pd.read_csv("Amazon_Reviews.csv")
df = df.head(500)
print(df.shape)
```

```
(500, 10)
```

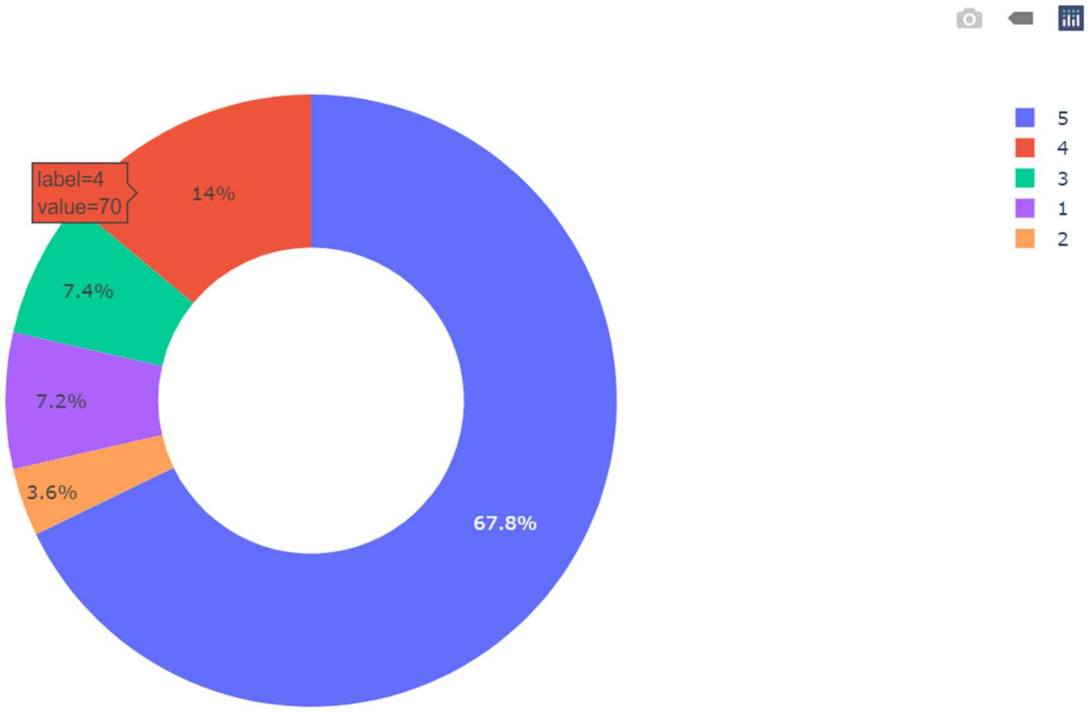
df.head()

| | Id | ProductId | UserId | ProfileName | HelpfulnessNumerator | HelpfulnessDenominator | Score | Time | Summary | Text |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0 | B001E4KFG0 | A3SGXH7AUHU8GW | delmartian | 1.0 | 1.0 | 5.0 | 1.303862e+09 | Good Quality Dog Food | I have bought several of the Vitality canned d... |
| 1 | 2.0 | B00813GRG4 | A1D87F6ZCVE5NK | dll pa | 0.0 | 0.0 | 1.0 | 1.346976e+09 | Not as Advertised | Product arrived labeled as Jumbo Salted Peanut... |
| 2 | 3.0 | B000LQOCH0 | ABXLMWJIXXAIN | Natalia Corres "Natalia Corres" | 1.0 | 1.0 | 4.0 | 1.219018e+09 | "Delight" says it all | This is a confection that has been around a fe... |
| 3 | 4.0 | B000UA0QIQ | A395BORC6FGVXV | Karl | 3.0 | 3.0 | 2.0 | 1.307923e+09 | Cough Medicine | If you are looking for the secret ingredient i... |
| 4 | 5.0 | B006K2ZZ7K | A1UQRSCLF8GW1T | Michael D. Bigham "M. Wassir" | 0.0 | 0.0 | 5.0 | 1.350778e+09 | Great taffy | Great taffy at a great price. There was a wid... |

The score pie charts based on number of ratings can be shown :

```
Scores = df["Score"].value_counts()
numbers = Scores.index
quantity = Scores.values

import plotly.express as px
figure = px.pie(df,
                values=quantity,
                names=numbers,hole = 0.5)
figure.show()
```



Now lets perform some quick EDA (Exploratory Data and Analysis ) on the given dataset by plotting the count of reviews and vs stars chart.

```
ax = df['Score'].value_counts().sort_index() \
   .plot(kind='bar',
         title='Count of Reviews by Stars',
         figsize=(10, 5))
ax.set_xlabel('Review Stars')
plt.show()
```


Now lets understand some basic nltk by downloading necessary nltk specific resources or models required for this project:

```
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

nltk.download('punkt'): This statement downloads the Punkt tokenizer models, which are used for tokenization tasks such as splitting text into sentences or words.

nltk.download('averaged_perceptron_tagger'): This statement downloads the averaged perceptron tagger model, which is used for part-of-speech tagging. It assigns grammatical tags to words in a sentence, such as noun, verb, adjective, etc.

nltk.download('maxent_ne_chunker'): This statement downloads the maximum entropy named entity chunker model. It is used for named entity recognition, which involves identifying and classifying named entities like persons, organizations, locations, etc., in text.

nltk.download('words'): This statement downloads the words corpus, which contains a collection of words from various languages. It can be useful for tasks like vocabulary analysis or language-specific applications.

Now tokenize some random example from the dataset:

```
example = df['Text'][50]
print(example)
```

```
This oatmeal is not good. Its mushy, soft, I don't like it. Quaker Oats is th
e way to go.
```

```
tokens = nltk.word_tokenize(example)
tokens[:10]
```

```
['This', 'oatmeal', 'is', 'not', 'good', '.', 'Its', 'mushy', ',', 'soft']
```

Mapping tokenize words to there corresponding part of speech in grammar using pos_tag function:

```
tagged = nltk.pos_tag(tokens)
tagged[:10]
```

```
[('This', 'DT'),
 ('oatmeal', 'NN'),
 ('is', 'VBZ'),
 ('not', 'RB'),
 ('good', 'JJ'),
 ('.', '.'),
 ('Its', 'PRP$'),
 ('mushy', 'NN'),
 (',', ','),
 ('soft', 'JJ')]
```

Where CC:  the conjunction of coordinating. CD: It is a digit of cardinal. DT: It is the determiner. JJ: Adjective ,NN: Noun etc.

entities = nltk.chunk.ne_chunk(tagged)
entities.pprint()

```
(S
  This/DT
  oatmeal/NN
  is/VBZ
  not/RB
  good/JJ
  ./.
  Its/PRP$
  mushy/NN
  ,/,
  soft/JJ
  ,/,
  I/PRP
  do/VBP
  n't/RB
  like/VB
  it/PRP
  ./.
  (ORGANIZATION Quaker/NNP Oats/NNPS)
  is/VBZ
  the/DT
  way/NN
  to/TO
  go/VB
  ./.)
```

VADER Sentiment Analysis :
Using NLTK's `SentimentIntensityAnalyzer` to get the negatives/neutral/positives (neg/neu/pos) scores of the text.
This uses a "bag of words" approach:

1. Stop words are removed
2. each word is scored and combined to a total score

```
from nltk.sentiment import SentimentIntensityAnalyzer
from tqdm.notebook import tqdm

sia = SentimentIntensityAnalyzer()
```

Here are few examples that shows working of SentimentIntensityAnalyzer:

```
sia.polarity_scores('I am so happy!')
```

```
{'neg': 0.0, 'neu': 0.318, 'pos': 0.682, 'compound': 0.6468}
```

```
sia.polarity_scores('This is the worst thing ever.')
```

```
{'neg': 0.451, 'neu': 0.549, 'pos': 0.0, 'compound': -0.6249}
```

```
print(example)
sia.polarity_scores(example)
```

```
This oatmeal is not good. Its mushy, soft, I don't like it. Quaker Oats is the wa
y to go.
{'neg': 0.22, 'neu': 0.78, 'pos': 0.0, 'compound': -0.5448}
```

Running Polarity scores on entire dataset and storing result in res variable:

```
res = {}
for i, row in tqdm(df.iterrows(), total=len(df)):
    text = row['Text']
    myid = row['Id']
    res[myid] = sia.polarity_scores(text)
```

```
vaders = pd.DataFrame(res).T
vaders = vaders.reset_index().rename(columns={'index': 'Id'})
vaders = vaders.merge(df, how='left')
```

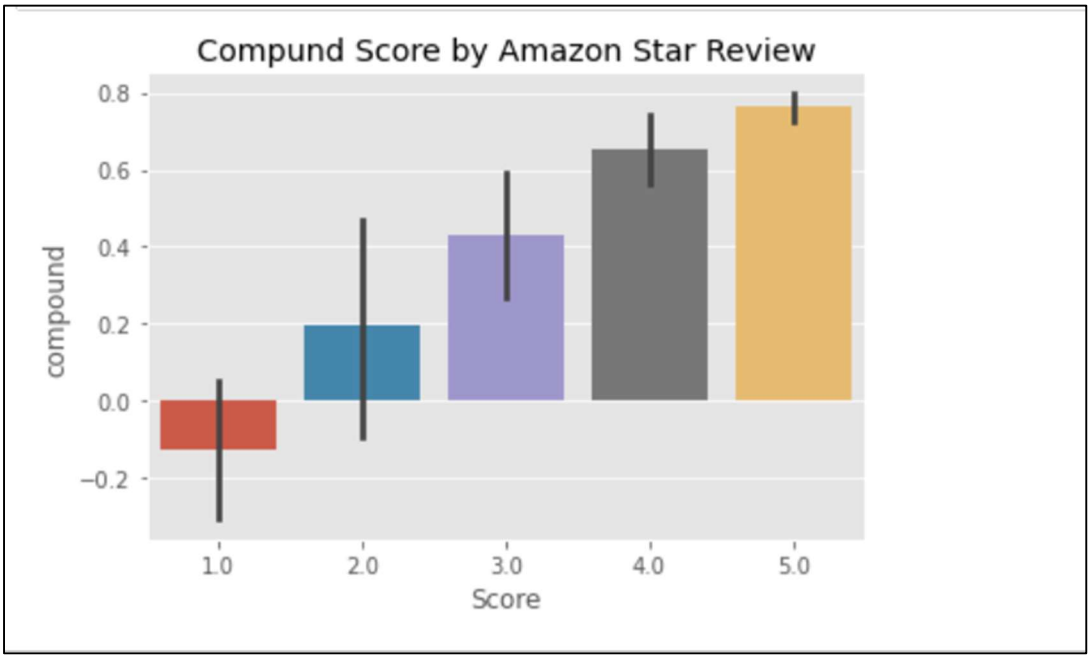
Now we have sentiment score in neg , neu and pos column along with metadata :

vaders.head()

| | Id | neg | neu | pos | compound | ProductId | UserId | ProfileName | HelpfulnessNumerator | HelpfulnessDenominator | Score | Time | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 0.000 | 0.695 | 0.305 | 0.9441 | B001E4KFG0 | A3SGXH7AUHU8GW | delmartian | 1.0 | 1.0 | 5.0 | 1.303862e+09 | I |
| 1 | 2.0 | 0.138 | 0.862 | 0.000 | -0.5664 | B00813GRG4 | A1D87F6ZCVE5NK | dll pa | 0.0 | 0.0 | 1.0 | 1.346976e+09 | A |
| 2 | 3.0 | 0.091 | 0.754 | 0.155 | 0.8265 | B000LQOCH0 | ABXLMWJIXXAIN | Natalia Corres "Natalia Corres" | 1.0 | 1.0 | 4.0 | 1.219018e+09 | |
| 3 | 4.0 | 0.000 | 1.000 | 0.000 | 0.0000 | B000UA0QIQ | A395BORC6FGVXV | Karl | 3.0 | 3.0 | 2.0 | 1.307923e+09 | |
| 4 | 5.0 | 0.000 | 0.552 | 0.448 | 0.9468 | B006K2ZZ7K | A1UQRSCLF8GW1T | Michael D. Bigham "M. Wassir" | 0.0 | 0.0 | 5.0 | 1.350778e+09 | G |

Plotting of VADER Results:

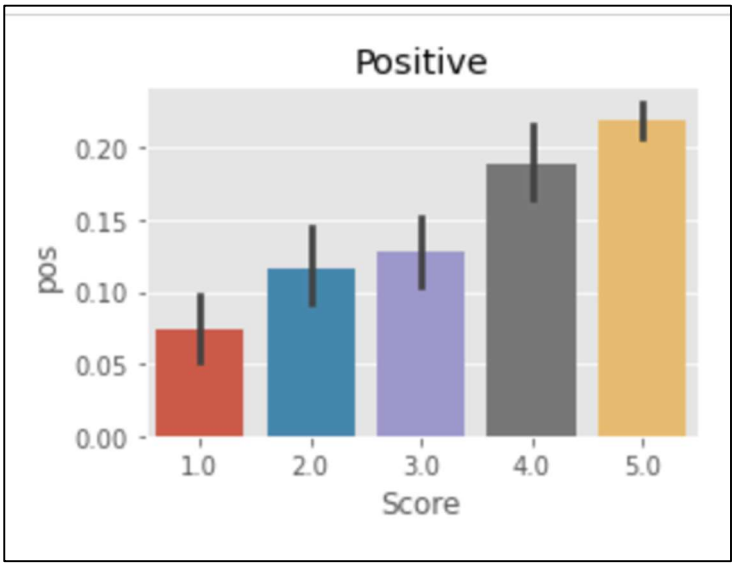
```
ax = sns.barplot(data=vaders, x='Score', y='compound')
ax.set_title('Compund Score by Amazon Star Review')
plt.show()
```



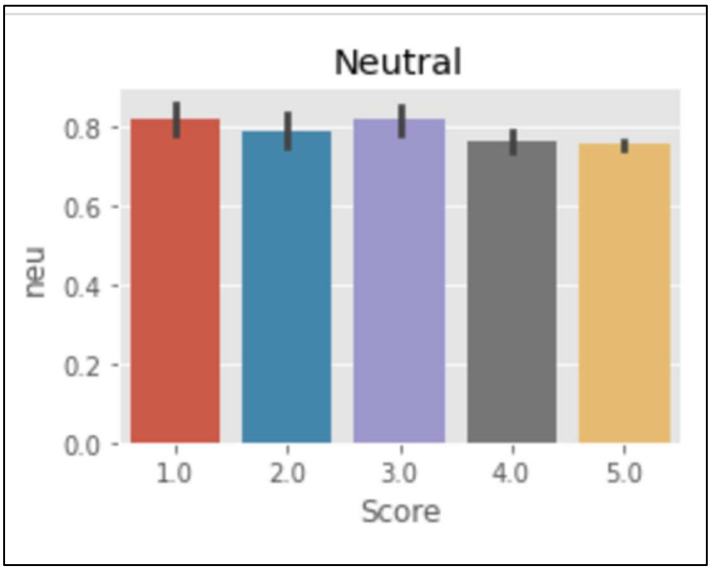

Plotting of Negative, Neutral and Positive Compound Plots

```
fig, axs = plt.subplots(1, 3, figsize=(12, 3))
sns.barplot(data=vaders, x='Score', y='pos', ax=axs[0])
sns.barplot(data=vaders, x='Score', y='neu', ax=axs[1])
sns.barplot(data=vaders, x='Score', y='neg', ax=axs[2])
axs[0].set_title('Positive')
axs[1].set_title('Neutral')
axs[2].set_title('Negative')
plt.tight_layout()
plt.show()
```

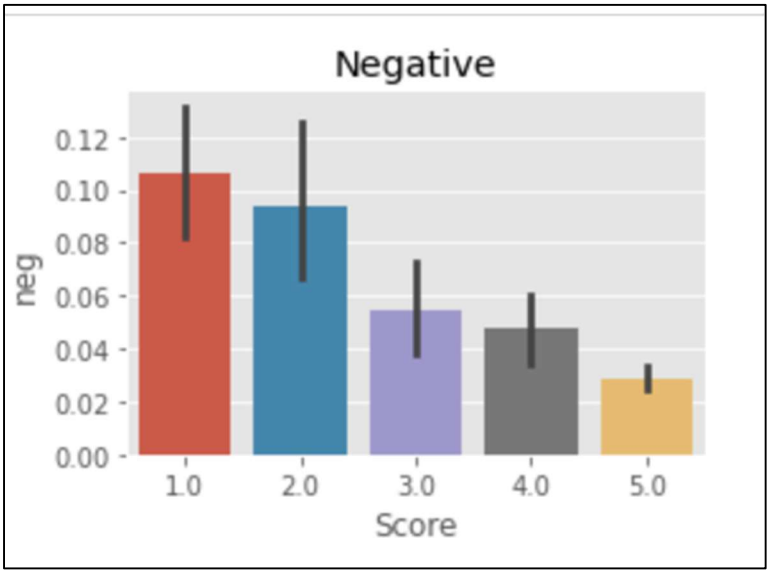
Positive Plot:



Neutral Plot:




Negative Plot:



# CONCLUSION

In this project, we successfully developed a sentiment analysis model for customer reviews on products. The model effectively classified the sentiment of the reviews, providing businesses with valuable insights into customer feedback and sentiments. The results of the project demonstrate the potential for utilizing sentiment analysis in understanding customer sentiments and improving products or services based on customer feedback. A different number of input and output features could be identified. Interestingly, some features appeared to be described in all the studies, while other features were more specific to a selected set of secondary studies. The results further indicate that sentiment analysis has been applied in various domains such e-commerce revies also, among which social media is the most popular. Also, the study showed that different domains require the use of different techniques.

There also seems a trend towards using more complex deep learning techniques, since they can detect more complex patterns in text and perform particularly well with larger datasets. In some use cases like, for example, advertisement, slight improvements in performance that can be obtained through deep learning can have a great impact. However, it should be noted that traditional machine learning models are less computationally expensive and perform sufficiently well for sentiment analysis tasks.

# FUTURE ENHANCEMENT

While the current sentiment analysis project provides valuable insights into customer sentiments, there are several potential areas for enhancement and future development. Some possible future enhancements for this project include:

1.Fine-grained Sentiment Analysis:

Currently, the sentiment analysis model classifies customer reviews into three categories: positive, negative, and neutral. However, in many cases, sentiments are more nuanced and can have varying degrees of positivity or negativity. Enhancing the model to perform fine-grained sentiment analysis, such as classifying reviews on a scale of 1-5 or using continuous sentiment scores, would provide more detailed insights into customer sentiments.

2.Aspect-based Sentiment Analysis:

In addition to overall sentiment analysis, customers often express sentiments about specific aspects or features of products. Developing an aspect-based sentiment analysis model can identify sentiments related to different aspects of the products, such as performance, usability, price, or customer service. This enhancement would provide granular insights into the strengths and weaknesses of specific product features.

3.Transfer Learning and Pre-trained Models:

Consider utilizing transfer learning techniques and pre-trained models specifically designed for sentiment analysis. By leveraging pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer), the model can benefit from the knowledge learned from vast amounts of text data, potentially improving accuracy and performance.

4.Sentiment Trend Analysis:

Extend the analysis to identify sentiment trends over time. By collecting and analyzing customer reviews over different time periods, it would be possible to identify shifts in customer sentiment, track changes in product satisfaction, and detect emerging trends. This analysis could provide valuable insights into the impact of product updates, marketing campaigns, or external factors on customer sentiments.

5.Sentiment Analysis for Different Languages:

Extend the sentiment analysis model to support different languages. This enhancement would enable businesses to analyze customer sentiments in multiple languages, allowing for a more comprehensive understanding of global customer feedback and opinion.

# REFERENCES

[1] Systematic reviews in sentiment analysis: a tertiary study by Alexander Ligthart, Cagatay Catal & Bedir Tekinerdogan. https://link.springer.com/article/10.1007/s10462-021-09973-3

[2] Sentiment Analysis on Amazon Reviews published on Towards Data Science by Enes Gokce https://towardsdatascience.com/sentiment-analysis-on-amazon-reviews-45cd169447ac

[3] Sentiment Analysis of Customer Product Reviews Using Machine Learning Zeenia Singla, Sukhchandan Randhawa, Sushma Jain Computer Science and Engineering Department Thapar University Patiala, India

[4] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.

[5] Kaggle article on Sentiment Analysis | Amazon reviews in Python · Amazon Musical Instruments Reviews by Ben Roshan https://www.kaggle.com/code/benroshan/sentiment-analysis-amazon-reviews

[6] VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text January 2015 Conference: Proceedings of the Eighth International AAAI Conference on Weblogs by C.J. Hutto Georgia Institute of Technology