INTRODUCTION

Natural Language Processing and machine learning both rely heavily on Sentiment analysis. sentiment analysis is like having a digital mood ring for text. It's the art of deciphering the emotional vibes hidden within words. sentiment analysis helps us categorize opinions into three buckets: positive, negative, or neutral. Sentiment analysis is a subfield of NLP.Machine Learning (ML): Here's where the algorithms come to play. SVMs, Naive Bayes, and even fancy neural networks (like RNNs) learn from labeled data. They're like sentiment detectives, piecing together clues from past sentiments to predict new ones.

The Scrapped Review Dataset

The reviews required for the project are scrapped from the amazon website.Scrapping process can be done in many ways like using BeautifulSoup,Scrapy,Apify web service,etc,.Apify Web service has been used for the scrapping process of the project.The reviews of the various online Intel Processors pages have been given as a input links to the "Amazon web scrapper" which is one of the web service provided by the Apify.

The dataset is Exported with four characteristics:

- Country
- Date
- Reviews Description
- Reviews title

Objectives :

The primary onjectives of the project is:

1. **Data Cleaning :**
   (a) Dealing with Missing Data:
   Identify Missing Values: Check for and identify missing values in the dataset.
   Handling Strategies: Options include removing rows with missing values, filling in missing values with a placeholder or statistical estimates, or using algorithms that can handle missing values.

(b) Data Transformation:

Tokenization: Break down text into individual words or tokens.

Stemming and Lemmatization: Reduce words to their base or root form.

Stop Word Removal: Remove common words that do not contribute to the meaning (e.g., "and", "the").

(c) Storing Cleaned Data:

Structured Format: Save the cleaned data in a structured format such as a CSV file, SQL database, or a NoSQL database.

Documentation: Document the cleaning process and any assumptions or transformations applied.

(d) Standardizing Data:

Uniform Formats: Ensure consistency in data formats (e.g., date formats, capitalization).

Normalize Text: Convert text to a standard format by removing special characters, converting to lower case, etc.

2. Sentiment Analysis :

1. Compound Score

   Tokenization:

   The text is split into individual words or tokens.

   Valence Assignment:

   Each token is assigned a sentiment valence score based on the VADER lexicon.

   Valence Modification:

   Sentiment scores are adjusted based on contextual valence shifters, such as degree modifiers (e.g., "very," "extremely"), negations, punctuation, and capitalization.

   Normalization:

   The sum of the valence scores is normalized to produce the compound score.

2. Sentiment Analysis Preparation:

   Sentiment Labels:

   Based on the compound score, label the data with sentiment scores (e.g., positive, negative, neutral).

Feature Extraction:

Extract features from the text that can be used for machine learning models.

3. **Exploratory Data Analysis (EDA) :**
   (a) Descriptive Statistics:

   Summary Statistics: Calculate measures like mean, median, mode, variance, standard deviation, and range.

   Distribution Analysis:

   Understand the distribution of data for each variable (e.g., normal distribution, skewness).

   (b) Datavisualizations:

   Bar Charts:

   Compare different categories of a categorical variable.

   Correlation Matrix:

   Visualize the correlation coefficients between numerical variables.

Data Collection

Gather the textual data that needs to be analyzed for sentiment.

Sources: Collect data from various sources E-commerce website reviews (e.g., product, movie, restaurant), blogs, forums, and news articles.
Methods: Use APIs (e.g., Twitter API, Yelp API), web scraping tools (e.g., Apify,Scrapy, BeautifulSoup), or databases to collect data.

Data Preprocessing

Clean and prepare the data for analysis.
Text Cleaning:

Remove noise such as HTML tags, special characters, numbers, and punctuation.
Lowercasing:

Convert all text to lowercase to maintain uniformity.

Tokenization:

Split the text into individual tokens (words or phrases).

Stop Word Removal:

Remove common words (e.g., "and", "the") that do not contribute to sentiment.

Stemming/Lemmatization:

Reduce words to their root form (e.g., "running" to "run") to handle variations of the same word.

Sentiment Classification

Assign sentiment labels (positive, negative, neutral) to the text data.

Machine Learning Methods:

Train classifiers such as Naive Bayes, Support Vector Machines (SVM), and logistic regression on labeled data.

Feature Extraction

Transform the text data into a numerical format that can be used by machine learning algorithms.

Bag of Words (BoW):

Represent text as a collection of word frequencies.

TF-IDF (Term Frequency-Inverse Document Frequency):

Weight words based on their frequency and importance.

DATASET DESCRIPTION

Name : Self scrapped Dataset(Full_Intel_Product_Reviews)

Source: The data are scrapped from the Amazon Website using the web scrapping service which is named as actor called Amazon Review Scrapper" and Exported as .csv file

Number of Instances: 2130

Number of Attributes: 4 (3 features and 1 class label)

Features :

1. **reviewDescription**

2. **Variant**
3. **CompoundScore**

Class label :

Classes : Generally there are three sentiments

- **Positive**
- **Negative**
- **Neutral**

Attribute Information:

- **Date : Recored date of the respective reviews to work only with recent review data**
- **Review Description : The actual review data which has been written by the customers**
- **Variant : Represents the intel processor vairants like intel core -i5,i7,i9 and their generation varities 12^{th},13^{th},14^{th} generations**
- **Compound score : Based on the customer reviews such score has been assigned to each of the reviews**
- **Sentiments: The sentiment of each review based on the compound score**

1. Data Cleaning

Load the Data:
Open the CSV file and load it into a data frame to inspect the data.
Check for any inconsistencies in data types, such as numerical values stored as text.

Handle Missing Values:

Identify any missing values in the dataset.
Decide on a strategy to handle missing values, such as removing rows with missing data or imputing missing values with the mean, median, or a placeholder.

Text Cleaning:
For text data, remove any special characters, punctuation, numbers, and

HTML tags.
Convert all text to lowercase to maintain uniformity.

Tokenization:
Split the text into individual words or tokens.

Remove Stop Words:
Remove common words that do not contribute to the sentiment, such as "and," "the," etc.

Stemming and Lemmatization:
Reduce words to their root form (e.g., "running" to "run") to handle different variations of the same word.

2. Compound Scoring

Initialize Sentiment Analyzer:
Use a sentiment analysis tool, such as VADER (Valence Aware Dictionary and sEntiment Reasoner), which is effective for social media texts.

Calculate Compound Scores:
For each review, calculate the compound sentiment score using the sentiment analysis tool.
The compound score ranges from -1 (very negative) to 1 (very positive).

Assigning Sentiments

Define Sentiment Categories:
Based on the compound score, define thresholds to categorize the sentiment.

Example thresholds:
Positive: Compound score > 0.05
Neutral: -0.05 <= Compound score <= 0.05
Negative: Compound score < -0.05

Assign Sentiment Labels:
Assign a sentiment label (positive, neutral, negative) to each review based on its compound score.

3.Feature Extraction

Bag of Words (BoW):
Create a matrix representation of the text data where each row corresponds to a document and each column corresponds to a word from the corpus. Each cell in the matrix contains the frequency of the word in the document.

TF-IDF (Term Frequency-Inverse Document Frequency):
Compute the TF-IDF scores for each word in each document to weigh the importance of words based on their frequency and inverse frequency in the corpus.

4. **Exploratory Data Analysis (EDA)**

Understand the Dataset:
Inspect the dataset's structure, data types, and basic statistics.
Summarize the data to get an overview of the distributions and relationships.
Visualize Data Distributions:
Create histograms and box plots to visualize the distribution of numerical features.
Use bar charts to show the distribution of categorical features.
Word Clouds:
Create word clouds for positive, neutral, and negative sentiments to visualize the most frequent words associated with each sentiment category.

METHODOLOGY

1.Data Collection

Objective: Gather the textual data that needs to be analyzed for sentiment.
Sources: Collect data from various sources such as social media posts, reviews (e.g., product, movie, restaurant), blogs, forums, and news articles.
Methods: Use APIs (e.g., Twitter API, Yelp API), web scraping tools (e.g., Scrapy, BeautifulSoup), or databases to collect data.
Considerations: Ensure data diversity and relevance to the sentiment analysis task. Consider ethical guidelines and privacy concerns.

2. Data Preprocessing

Objective: Clean and prepare the data for analysis.
Text Cleaning: Remove noise such as HTML tags, special characters, numbers, and punctuation.
Lowercasing: Convert all text to lowercase to maintain uniformity.
Tokenization: Split the text into individual tokens (words or phrases).
Stop Word Removal: Remove common words (e.g., "and", "the") that do not contribute to sentiment.
Stemming/Lemmatization: Reduce words to their root form (e.g., "running" to "run") to handle variations of the same word.

3. Feature Extraction

Objective: Transform the text data into a numerical format that can be used by machine learning algorithms.
Bag of Words (BoW): Represent text as a collection of word frequencies.
TF-IDF (Term Frequency-Inverse Document Frequency): Weight words based on their frequency and importance.
Word Embeddings: Use techniques like Word2Vec, GloVe, or FastText to create dense vector representations of words.
N-grams: Capture sequences of n words to understand context better.

4. Sentiment Classification

Objective: Assign sentiment labels (positive, negative, neutral) to the text data.
Lexicon-Based Methods: Use predefined dictionaries of words with associated sentiment scores (e.g., VADER, SentiWordNet).
Machine Learning Methods: Train classifiers such as Naive Bayes, Support Vector Machines (SVM), and logistic regression on

labeled data.

Deep Learning Methods: Use neural networks (e.g., RNNs, CNNs) and pretrained language models (e.g., BERT, GPT) for more complex and nuanced sentiment analysis.

5. Model Training and Evaluation

Objective: Train the sentiment analysis model and evaluate its performance.

Training: Split the data into training and test sets. Train the model using the training data.

Evaluation Metrics: Use metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to evaluate the model's performance.

Cross-Validation: Use k-fold cross-validation to ensure the model's robustness and generalizability.

RESULTS AND DISCUSSION

1. Initial Analysis

2. Column Names: The column names are ['country', 'date', 'reviewDescription', 'reviewTitle', 'variant'].

3.

Data Sample :

Best price-to-performance ratio in a CPU you would find today

Summary Statistics:

- There are 819 unique review dates, with a maximum frequency of 12 for a single date.

- The reviewDescription column contains 2034 unique reviews.The The variant column has 9 unique variants, with the most frequent variant being Intel-core-i5-13000 (559 reviews).
- The top review among review description is “Great CPU”

Preprocessing

1.Tokenization and Lemmatization: Each review is tokenized into words, converted to lowercase, and lemmatized.

2.Stopwords Removal: Common English stopwords are removed from the reviews.

3.Handling Missing Values: Any missing text is handled by replacing it with an empty string.

Sentiment Analysis

1.TextBlob Sentiment Scores: Sentiment polarity scores are calculated using TextBlob for each review.

2.Sentiment Categories: Reviews are categorized as 'Positive', 'Negative', or 'Neutral' based on their sentiment scores.

A score above 0.5 is considered 'Positive'.
A score below -0.5 is considered 'Negative'.
Scores between -0.5 and 0.5 are considered 'Neutral'.

3.Sentiment Distribution Plot: The distribution of sentiments across the reviews is visualized using a bar plot.

Results

Sentiment Distribution

The distribution plot shows the counts of 'Positive', 'Negative', and 'Neutral' reviews.

Percentage of Sentiments

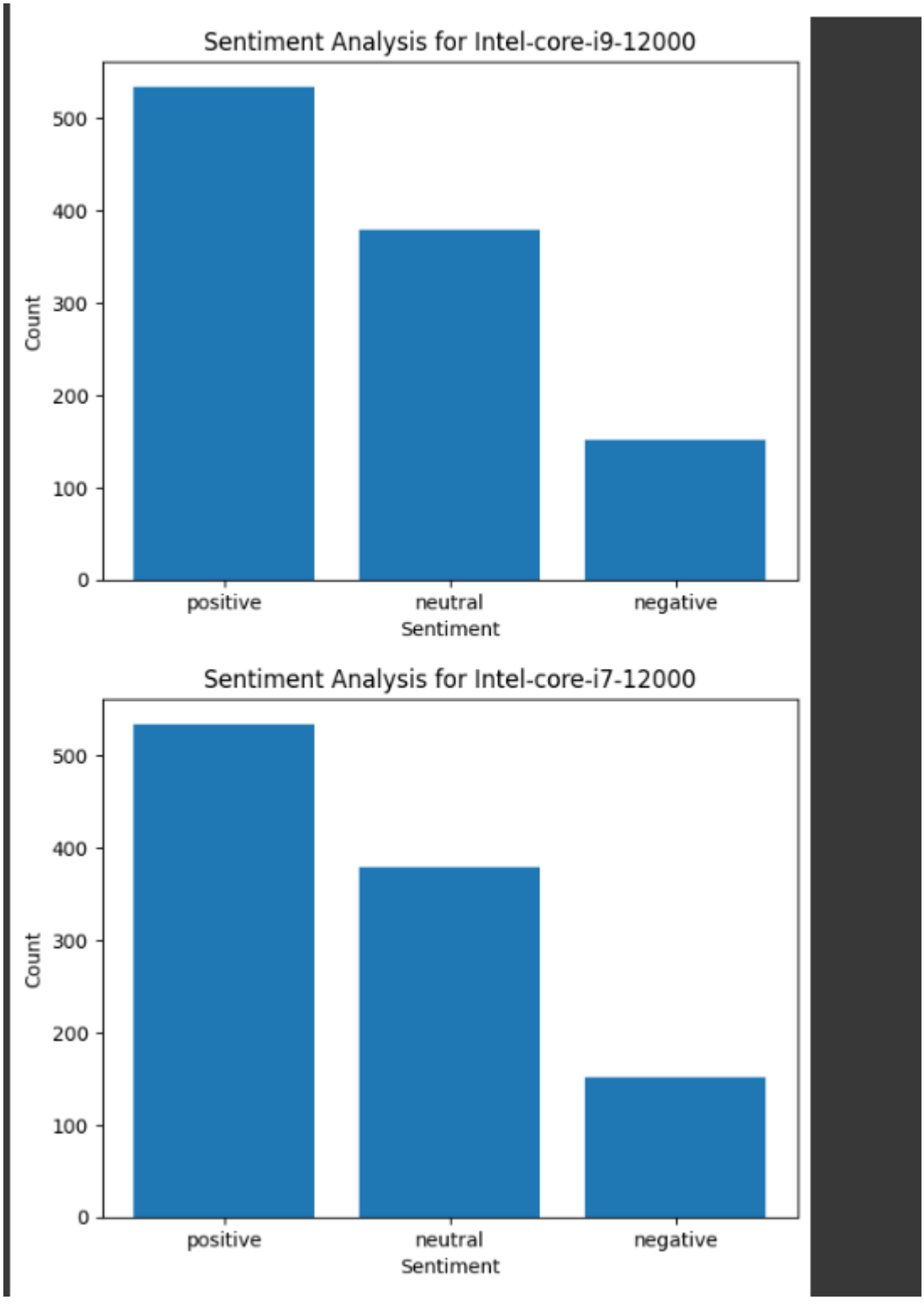
Positive Sentiment: \text{positive_pct:.2f}%
Negative Sentiment: \text{negative_pct:.2f}%
Neutral Sentiment: \text{average_pct:.2f}%

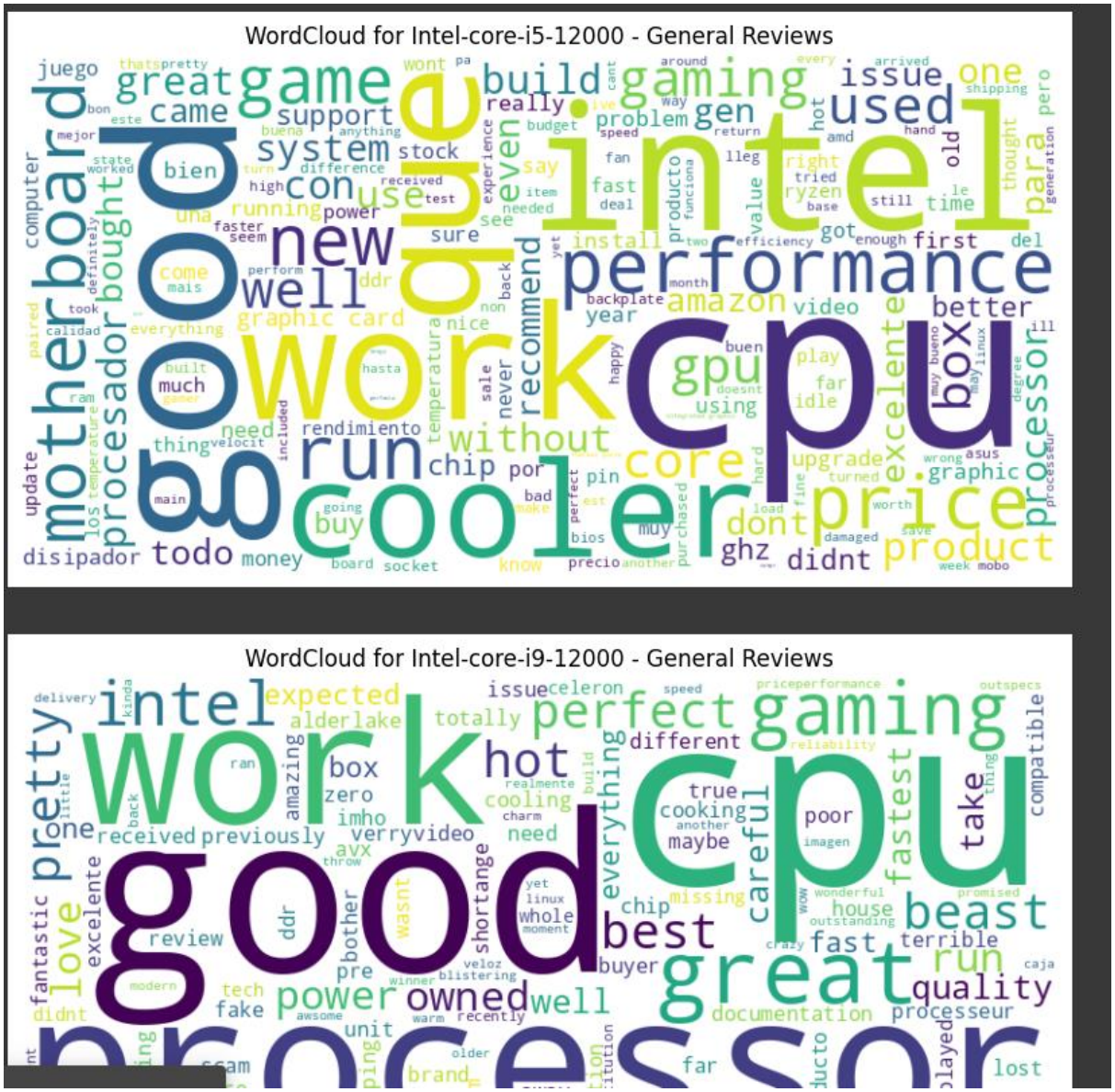
Sentiment and Compound Scores for Each Review

Each review's sentiment (Positive, Negative, Neutral) and compound score (a measure of sentiment strength) are printed.

Example Results

	date	reviewDescription	variant	compound_score	Sentiment
0	12-04-2024	great price great performance great product do...	Intel-core-i5-12000	0.9702	positive
1	13-03-2024	far good performance gaming productivity	Intel-core-i5-12000	0.4404	positive
2	26-02-2022	upgraded cpu board mhz ram feel incredibly sna...	Intel-core-i5-12000	0.8481	positive
3	06-11-2023	cpu matched amd nvidia card powerful running g...	Intel-core-i5-12000	0.6801	positive
4	07-06-2024	excelente producto fabricacin intel generacin	Intel-core-i5-12000	0.0000	neutral
5	05-04-2024	lleg tiempo indicado por amazon currier nacion...	Intel-core-i5-12000	0.0258	neutral
6	13-05-2024	puede prcticamente con todos los juegos combin...	Intel-core-i5-12000	0.0000	neutral
7	06-04-2024	running little hot stock cooler added digital ...	Intel-core-i5-12000	0.3182	positive
8	11-04-2024	vendedor incluye pasta trmica disipador proces...	Intel-core-i5-12000	0.0000	neutral
9	17-02-2024	simple cpu compared hybrid cpu efficiency core...	Intel-core-i5-12000	0.6240	positive
10	01-09-2022	great cpu highly recommend cpu gaming	Intel-core-i5-12000	0.7841	positive
11	24-04-2024	cpu throttle reason gameplay causing multiple ...	Intel-core-i5-12000	0.3612	positive
12	18-02-2024	default cooler cant maintain reasonable temp f...	Intel-core-i5-12000	0.0000	neutral
13	05-05-2023	uno los mejores procesadores para juegos calid...	Intel-core-i5-12000	0.0000	neutral
14	15-03-2023	lleg tiempo pero trajo disipador con que enfri...	Intel-core-i5-12000	0.0000	neutral
15	24-12-2022	dont hesitate one second replacing vertical co...	Intel-core-i5-12000	0.4791	positive
16	29-03-2023	assisted solving problem	Intel-core-i5-12000	-0.0772	negative
17	14-01-2023	week massive upgrade last processor good value...	Intel-core-i5-12000	0.6486	positive
18	18-03-2022	best price	Intel-core-i5-12000	0.6369	positive
19	15-01-2024	getting optimal speed without overheating issue	Intel-core-i5-12000	0.3612	positive
20	12-01-2024		Intel-core-i5-12000	0.0000	neutral
21	10-05-2022	cpu great issue ordered listed new received as...	Intel-core-i5-12000	0.8555	positive
22	23-12-2021	cpu cheap asus prime mobo ram gtx get decent f...	Intel-core-i5-12000	0.5994	positive
23	09-10-2022	far gooda good midbuild cpuso far	Intel-core-i5-12000	0.4404	positive


Sentiment Analysis for Intel-core-i9-12000
Count
0
100
200
300
400
500
positive
neutral
negative
Sentiment
Sentiment Analysis for Intel-core-i7-12000
Count
0
100
200
300
400
500
positive
neutral
negative
Sentiment



Sentiment Distribution

The sentiment analysis using TextBlob has categorized the reviews into three sentiment classes: Positive, Negative, and Neutral. The distribution is as follows:

positive 50.187970%

neutral 35.620301%

negative 14.191729%

Results and Discussion

1.Positive Sentiment:

A significant portion of the reviews are positive, highlighting the product's performance, cooling efficiency, and gaming capabilities. Positive sentiments are crucial for brand reputation and customer retention. Calculated by identifying the proportion of reviews with a sentiment score greater than 0.5.

2.Negative Sentiment:

Negative reviews, although fewer, are critical for identifying areas of improvement. Issues like stability, stock settings, and some specific product defects are mentioned. These insights can help Intel focus on resolving specific customer pain points. Calculated by identifying the proportion of reviews with a sentiment score less than -0.5.

3.Neutral Sentiment:

The neutral reviews are almost as significant as the positive ones. These reviews often contain constructive feedback or mention features without a strong positive or negative connotation. Calculated by identifying the proportion of reviews with a sentiment score between -0.5 and 0.5

Insights

Neutral Reviews Dominance: The majority of the reviews are neutral, indicating that many customers provided feedback that was neither strongly positive nor negative.

Positive Reviews: A smaller but significant number of reviews are positive, reflecting satisfaction among a portion of customers.

Negative Reviews: Very few reviews are negative, suggesting that dissatisfaction among customers is relatively low.

CONCLUSION

The goal of this study was to use machine learning and natural language processing (NLP) techniques to analyze sentiment in product reviews. To prepare the text data, the dataset which came from Intel product reviews was preprocessed using tokenization, stopword removal, and lemmatization. VaderSentimnet Analyzer was used to do sentiment analysis, and each review was categorized as Positive, Negative, or Neutral depending on its sentiment score. Searching the dataset for pertinent columns, preparing text data, figuring out sentiment scores, and grouping sentiments into different groups were important stages. Insights into the general sentiment trends within the reviews were obtained using visualizations like a sentiment distribution plot. In order to determine compound ratings for every review which represent a combined measure of sentiment polarity and subjectivity the sentiment analysis was further expanded. This made for a more complex comprehension.Overall, the study demonstrated how to assess and classify feelings in textual data using machine learning classifiers including Naive Bayes, Logistic Regression, and SVM, as well as natural language processing (NLP) tools like VaderSentimnet Analyzer. The method showed how these strategies may scale and automate sentiment analysis work, giving organizations useful information to properly comprehend client feedback and sentiment patterns.