INTEL PRODUCTS SENTIMENT ANALYSIS FROM ONLINE REVIEWS

Team Members:
Harshini N Murthy-1ms22ai017@msrit.edu,
Neelam Bind-1ms22ai039@msrit.edu,
Nandeesha HK-1ms22ai037@msrit.edu

Abstract:

This project analyzes user reviews for Intel processors from various online platforms using NLP and machine learning techniques like GloVe and VADER. It aims to extract insights into customer sentiments to support Intel's product development and marketing strategies. The findings will inform future improvements and marketing efforts based on consumer feedback**.**

Introduction:

User and technical reviews are essential for evaluating Intel processors' performance and reliability. Online reviews from e-commerce sites, tech forums, and social media provide valuable feedback, reflecting real-world experiences and influencing purchasing decisions. Sentiment analysis of these reviews helps identify common issues, measure satisfaction, and track trends, aiding in product development and customer satisfaction.

Objective:

The primary objective of this sentiment analysis project is to meticulously dissect feedback on Intel processors, extracting actionable insights into consumer sentiments. Employing cuttingedge natural language processing (NLP) and machine learning techniques, we aim to classify sentiments, uncover key trends, and address common issues. By evaluating the effectiveness of various sentiment analysis methods, this project aspires to bolster Intel's product development strategies, refine marketing efforts, and enhance customer engagement through a nuanced understanding of consumer perceptions.

Scope:

- **Type of Reviews**: Includes user-generated and technical reviews, along with feedback from e-commerce platforms and tech forums focused on Intel processors.
- **Time Frame**: Data spans from January 2007 to December 2023, capturing recent consumer opinions and trends.
- **Data Sources**: Extracted from structured e-commerce sites, unstructured tech forums, and YouTube video captions for a comprehensive sentiment analysis.
- **Sentiment Analysis Techniques**: Utilizes GloVe embeddings and linear regression to model sentiment trends, comparing effectiveness with VADER for nuanced insights into consumer perceptions of Intel processors.

Literature Review

- **Pang & Lee (2008)**:

  - Conducted a comparative study of SVM, Naive Bayes, and Maximum Entropy Models for sentiment classification.
  - Highlighted the advantages of machine learning approaches in effectively analyzing sentiment from textual data.

- **Liu (2012)**:

  - Explored various sentiment analysis techniques with a focus on developing domain-specific lexicons. o Discussed methods to enhance sentiment detection in product reviews through tailored lexicons, improving accuracy and relevance.

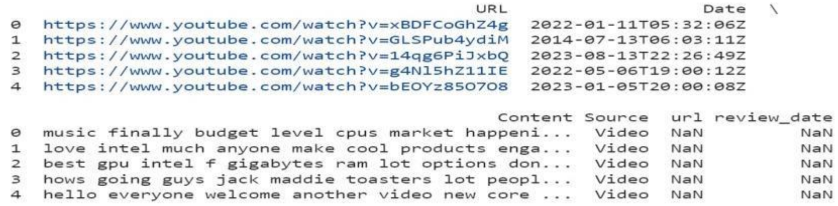
Sentiment Analysis Techniques and Tools

- **Lexicon-Based Approaches**:
  - Utilizes tools like VADER (Valence Aware Dictionary and sEntiment Reasoner) that rely on predefined dictionaries. o Classifies sentiment into positive, negative, or neutral categories based on weighted word scores, suitable for analyzing informal text sources like social media and customer reviews.
- **Machine Learning Approaches**:
  - Applies supervised learning techniques such as Naive Bayes, Support Vector Machines (SVM), and Random Forest.
  - Trains models on labeled datasets to classify sentiment, leveraging statistical algorithms to generalize patterns and predict sentiment from new data points.
- **Deep Learning Approaches**:
  - Employs advanced neural network architectures including Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Convolutional Neural Networks (CNNs). o Captures intricate patterns and dependencies in text data, enhancing sentiment analysis accuracy by automatically learning features and relationships.
- **Word Embedding Models**:
  - Utilizes Word2Vec and GloVe embeddings to convert words into dense numerical vectors.
  - Represents semantic meanings and relationships between words, improving sentiment analysis by capturing contextual nuances and semantic similarities in textual data.
- **Hybrid Approaches**:
  - Integrates strengths from both lexicon-based and machine learning or deep learning methods.
  - Combines rule-based sentiment scoring with machine-learned classifiers to leverage contextual understanding and statistical robustness, achieving more accurate sentiment analysis results across diverse datasets and text types.

Data Collection:

Data Sources:

Extracted user reviews from multiple platforms including Amazon, Newegg, Reddit, Tom's Hardware, TechSpot, Best Buy etc and YouTube.

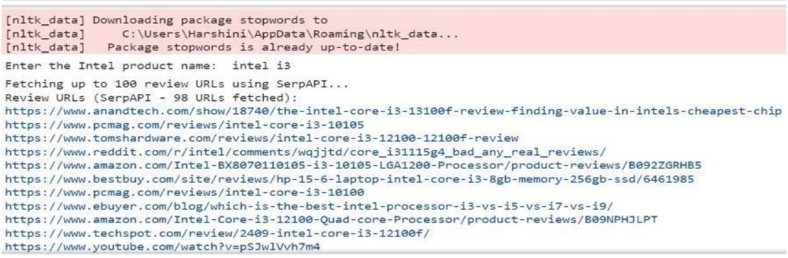


```
                                            URL                  Date  \
0  https://www.youtube.com/watch?v=xBDFCoGhZ4g  2022-01-11T05:32:06Z
1  https://www.youtube.com/watch?v=GLSPub4ydiM  2014-07-13T06:03:11Z
2  https://www.youtube.com/watch?v=14qg6PiJxbQ  2023-08-13T22:26:49Z
3  https://www.youtube.com/watch?v=g4Nl5hZ11IE  2022-05-06T19:00:12Z
4  https://www.youtube.com/watch?v=bEOYz85O7O8  2023-01-05T20:00:08Z

                                             Content Source  url review_date
0  music finally budget level cpus market happeni...  Video  NaN         NaN
1  love intel much anyone make cool products enga...  Video  NaN         NaN
2  best gpu intel f gigabytes ram lot options don...  Video  NaN         NaN
3  hows going guys jack maddie toasters lot peopl...  Video  NaN         NaN
4  hello everyone welcome another video new core ...  Video  NaN         NaN
```

Data Source :

- **Web Scraping:** Data from **Amazon**, **Newegg**, **Reddit**, **Tom's Hardware**, **TechSpot**, and **Best Buy** is collected using **web scraping techniques**. Python libraries such as **BeautifulSoup** is used to extract review text, ratings, and dates from these websites.

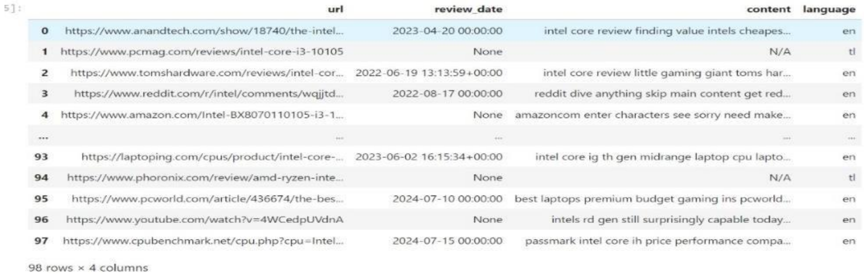


```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Harshini\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Enter the Intel product name:  intel i3
Fetching up to 100 review URLs using SerpAPI...
Review URLs (SerpAPI - 98 URLs fetched):
https://www.anandtech.com/show/18740/the-intel-core-i3-13100f-review-finding-value-in-intels-cheapest-chip
https://www.pcmag.com/reviews/intel-core-i3-10105
https://www.tomshardware.com/reviews/intel-core-i3-12100-12100f-review
https://www.reddit.com/r/intel/comments/wqjjtd/core_i31115g4_bad_any_real_reviews/
https://www.amazon.com/Intel-BX8070110105-i3-10105-LGA1200-Processor/product-reviews/B092ZGRHB5
https://www.bestbuy.com/site/reviews/hp-15-6-laptop-intel-core-i3-8gb-memory-256gb-ssd/6461985
https://www.pcmag.com/reviews/intel-core-i3-10100
https://www.ebuyer.com/blog/which-is-the-best-intel-processor-i3-vs-i5-vs-i7-vs-i9/
https://www.amazon.com/Intel-Core-i3-12100-Quad-core-Processor/product-reviews/B09NPHJLPT
https://www.techspot.com/review/2409-intel-core-i3-12100f/
https://www.youtube.com/watch?v=pSJwlVvh7m4
```

- **APIs:** Reviews and captions are collected using **APIs**:
  - **SERP API** is used to gather review data from multiple sources.
  - **YouTube API** is used to extract **video captions** related to Intel processors

DataDescription:

- **Number of Reviews:** The dataset includes **over 100 reviews** from various sources.
- **Time Span:** Reviews were collected over a period of **six months** (January 2007 - June 2024).
- **Features:** The dataset contains the following features:
  - **Review Text:** The textual content of the review.
  - **Date:** The date when the review was posted.
  - **Source:** The platform URL where the review was collected (Amazon, Newegg, Reddit, etc.)and collected data from more than 50 urls.
  - **Video Captions:** Transcriptions from YouTube videos, including user comments and discussions.



5]:

	url	review_date	content	language
0	https://www.anandtech.com/show/18740/the-intel...	2023-04-20 00:00:00	intel core review finding value intels cheapes...	en
1	https://www.pcmag.com/reviews/intel-core-i3-10105	None	N/A	tl
2	https://www.tomshardware.com/reviews/intel-cor...	2022-06-19 13:13:59+00:00	intel core review little gaming giant toms har...	en
3	https://www.reddit.com/r/intel/comments/wqjjtd...	2022-08-17 00:00:00	reddit dive anything skip main content get red...	en
4	https://www.amazon.com/Intel-BX8070110105-i3-1...	None	amazoncom enter characters see sorry need make...	en
...	...	...	...	...
93	https://laptoping.com/cpus/product/intel-core-...	2023-06-02 16:15:34+00:00	intel core ig th gen midrange laptop cpu lapto...	en
94	https://www.phoronix.com/review/amd-ryzen-inte...	None	N/A	tl
95	https://www.pcworld.com/article/436674/the-bes...	2024-07-10 00:00:00	best laptops premium budget gaming ins pcworld...	en
96	https://www.youtube.com/watch?v=4WCedpUVdnA	None	intels rd gen still surprisingly capable today...	en
97	https://www.cpubenchmark.net/cpu.php?cpu=Intel...	2024-07-15 00:00:00	passmark intel core ih price performance compa...	en

98 rows × 4 columns

Data Preprocessing

Cleaning:

1. **Removing Duplicates:** Duplicate reviews are identified and removed to ensure each review is unique. This step is crucial as it prevents the bias introduced by repeated information and maintains the integrity of the dataset. Duplicate reviews can skew analysis results and misrepresent the actual sentiment distribution among users.
2. **Handling Missing Values:**

   a. **Review Text and Rating:** Reviews that lack text or a rating are typically discarded from the dataset. Textual content is fundamental for sentiment analysis, and without it, sentiment cannot be accurately assessed. Similarly, ratings provide quantitative feedback that is essential for understanding sentiment polarity.

   b. **Review Dates:** Missing dates, if crucial for the analysis (such as tracking sentiment trends over time), are handled by either filling them with placeholders or imputing them based on other available data. This ensures that the chronological order of reviews is maintained, allowing for accurate temporal analysis of sentiment changes.

Text Processing:

1. **Tokenization:** The review text undergoes tokenization, where it is segmented into individual tokens such as words and punctuation marks. Tokenization forms the foundational step for further text analysis, enabling the extraction of meaningful insights from the textual content.
2. **Stopword Removal:** Commonly occurring words that do not carry significant meaning for sentiment analysis, such as articles ("the", "an"), prepositions ("in", "on"), and conjunctions ("and", "but"), are removed using a predefined list of stopwords. Removing stopwords helps in focusing the analysis on words that convey sentiment and are more informative for understanding user opinions.
3. **Stemming:** Words in the review text are stemmed to their root forms, reducing inflected or derived words to their base or root form. Stemming helps in standardizing words and reducing the complexity of the vocabulary, which improves the efficiency and effectiveness of subsequent sentiment analysis algorithms.

Normalization:

1. **Lowercasing:** All text data is converted to lowercase. This normalization step ensures that words with different cases (e.g., "Word" and "word") are treated identically during analysis.
2. **Noise Removal:** Non-alphanumeric characters, URLs, and special characters that do not contribute to the semantic meaning of the text are removed. Noise removal focuses the analysis on the meaningful textual content.
3. **Data Standardization:** Column names and data formats are standardized across the dataset. Standardization ensures consistency in data representation, facilitating easier data manipulation, analysis, and interpretation.

Sentiment Analysis Methodology

Approach

For this sentiment analysis project, we employed a combination of **rule-based** and **machine learning** approaches to evaluate and classify user sentiments from the collected reviews.

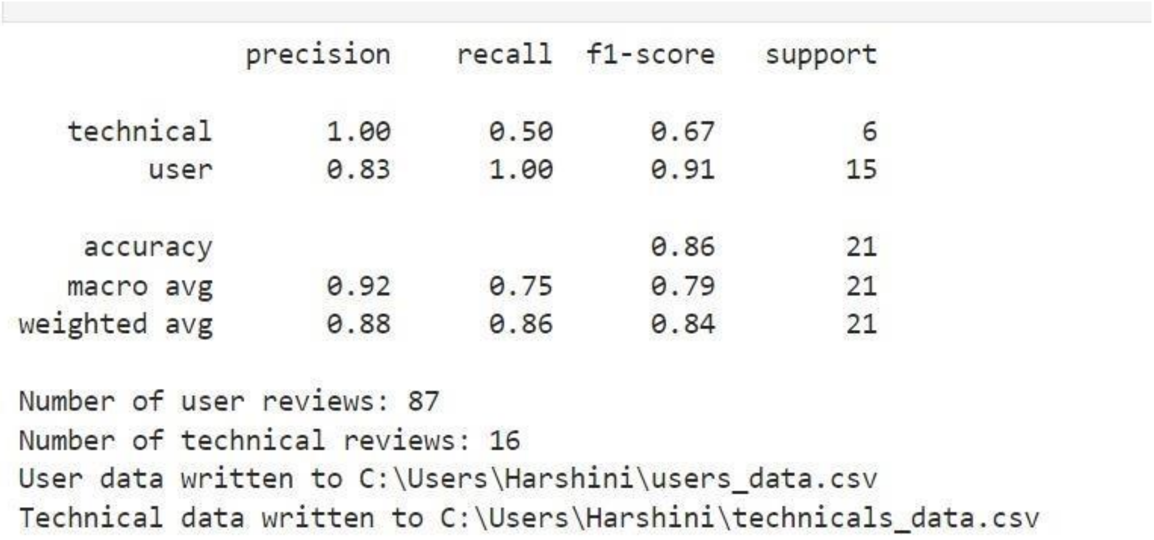
o **Rule-Based Approach:** We used **VADER (Valence Aware Dictionary and Sentiment Reasoner)**, a lexicon and rule-based sentiment analysis tool specifically designed to analyze social media texts. VADER assigns sentiment scores based on predefined word valence and captures the intensity of sentiments, making it suitable for initial sentiment classification.

Model Selection

o **VADER:** VADER was selected for its effectiveness in handling social media text and its capability to provide nuanced sentiment scores (positive, negative, neutral, and compound). It's particularly adept at capturing sentiments in informal, conversational language, making it a good fit for user reviews and YouTube captions.

Feature Extraction

o **Word Embeddings:** We also explored **GloVe (Global Vectors for Word Representation)** embeddings to capture semantic relationships between words. Word embeddings provide dense vector representations that encapsulate the context of words in a text, enhancing the feature set for machine learning models. We are using GloVe embeddings to classify reviews into **user** and **technical** data, improving the analysis by differentiating between general user sentiments and more technical, expert-level feedback.



```
              precision    recall  f1-score   support

   technical       1.00      0.50      0.67         6
        user       0.83      1.00      0.91        15

    accuracy                           0.86        21
   macro avg       0.92      0.75      0.79        21
weighted avg       0.88      0.86      0.84        21

Number of user reviews: 87
Number of technical reviews: 16
User data written to C:\Users\Harshini\users_data.csv
Technical data written to C:\Users\Harshini\technicals_data.csv
```

Implementation

Tools and Libraries :

- **Programming Language**:

- **Python**: Main language for implementation.

- **Text Processing**:

- **NLTK (Natural Language Toolkit)**: Tokenization, stemming, lemmatization, stopword removal.

- **Machine Learning**:

- **scikit-learn**: Feature extraction (TF-IDF), model training (logistic regression), evaluation.
- **VADER**: Rule-based sentiment analysis tool.
- **GloVe**: Word embeddings for capturing semantic relationships.

- **Data Collection**:

- **SERP API**: Web scraping from online sources.
- **YouTube API**: Extracting video captions for dataset enrichment.

- **Additional Libraries**:

- **beautifulsoup4**: Web scraping.
- **nltk.corpus**: Accessing NLTK corpora.
- **langdetect.lang_detect_exception**: Language detection exception handling.
- **re**: Regular expressions.
- **youtube_transcript_api**: Extracting YouTube video transcripts.
- **seaborn**: Statistical data visualization.
- **matplotlib**: Creating visualizations.
- **os**: Operating system interfaces for file operations and system-level functionalities.
- **requests**: HTTP library for making requests and working with APIs.
- **PyTorch**: Open-source machine learning framework that accelerates the path from research prototyping to production deployment.
- **NumPy**: Fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices.
- **pandas**: Data analysis and manipulation library, offering data structures and operations for manipulating numerical tables and time series

Model Training

Data Split:

- **Training Set**: Typically, 70-80% of the data is used for training the model.
- **Validation Set**: Around 10-15% of the data is used for hyperparameter tuning and to avoid overfitting.
- **Test Set**: The remaining 10-15% of the data is used to evaluate the model's performance.

Hyperparameters:

- **Logistic Regression**:
    - **Regularization (C)**: Inverse of regularization strength. Common values include 0.01, 0.1, 1, 10.
    - **Solver**: Algorithms like 'liblinear', 'saga', 'lbfgs'.
    - **Max Iterations**: Maximum number of iterations for convergence, typically 100-200.

Training Time:

- Training time can vary depending on the dataset size and the hyperparameters chosen. Typically, training a logistic regression model on text data can take from a few seconds to several minutes.

Evaluation Metrics

Accuracy:

- **Definition**: The ratio of correctly predicted instances to the total instances. **Formula**:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

Precision:

- **Definition**: The ratio of correctly predicted positive observations to the total predicted positives.
- **Formula**:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sensitivity):

- **Definition**: The ratio of correctly predicted positive observations to all observations in the actual class.

$$: \text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

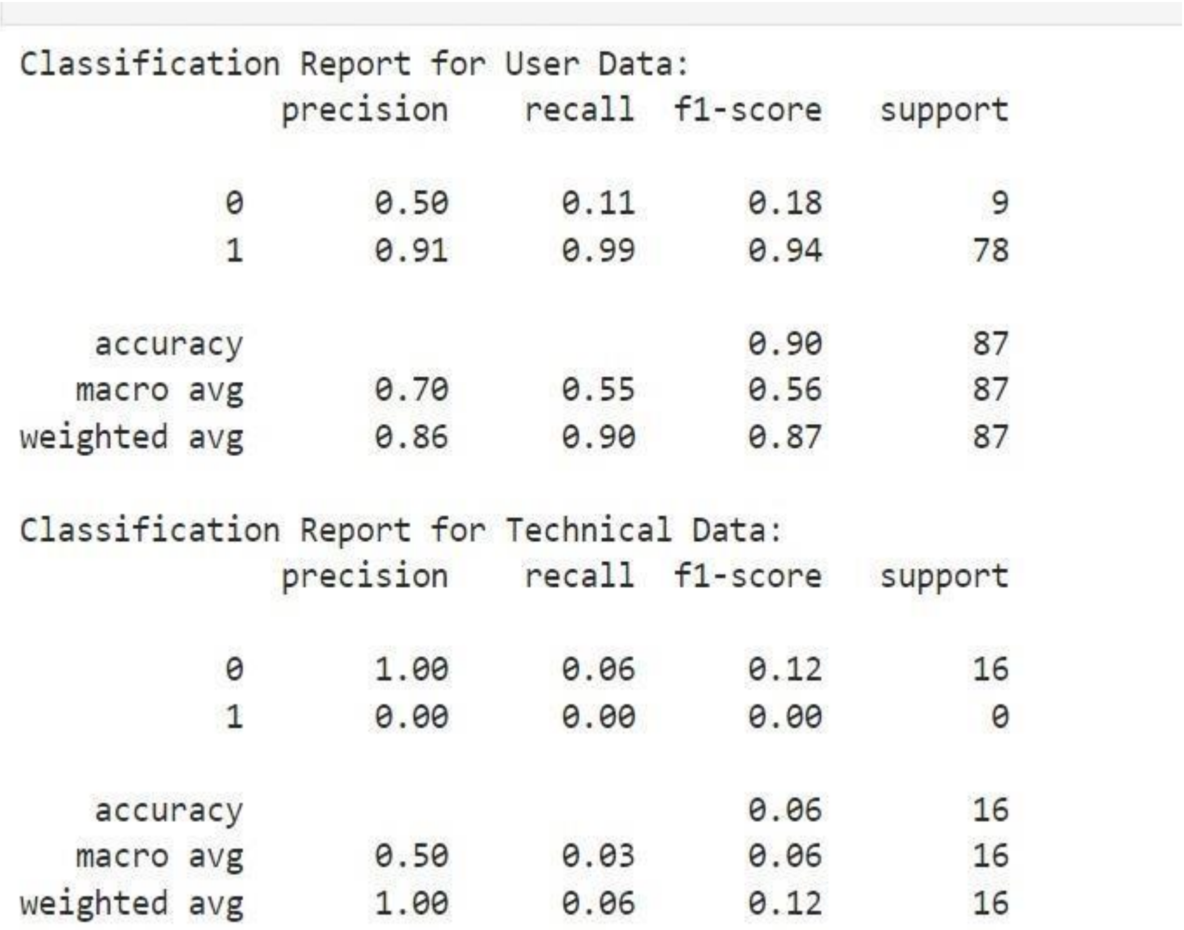
- **Formula**:

F1 Score:

- **Definition**: The weighted average of Precision and Recall, providing a balance between the two.
- **Formula**:

$$\text{F1 Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

These metrics help in evaluating the performance of the sentiment analysis model, ensuring that it not only predicts accurately but also handles false positives and false negatives effectively.



```
Classification Report for User Data:
              precision    recall  f1-score   support

           0       0.50      0.11      0.18         9
           1       0.91      0.99      0.94        78

    accuracy                           0.90        87
   macro avg       0.70      0.55      0.56        87
weighted avg       0.86      0.90      0.87        87

Classification Report for Technical Data:
              precision    recall  f1-score   support

           0       1.00      0.06      0.12        16
           1       0.00      0.00      0.00         0

    accuracy                           0.06        16
   macro avg       0.50      0.03      0.06        16
weighted avg       1.00      0.06      0.12        16
```

Results and Discussions

1. Model Performance:

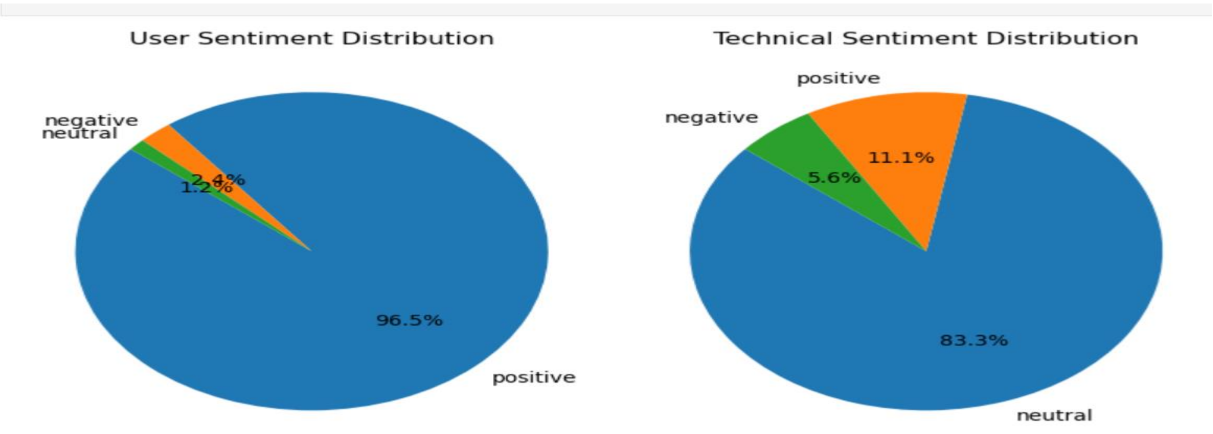
The script utilizes VADER (Valence Aware Dictionary and sEntiment Reasoner), a rule-based sentiment analysis tool, to evaluate the sentiment in user and technical data related to Intel products. Here’s how the model performs:

- **Sentiment Analysis Method:** VADER calculates a compound sentiment score for each review text, indicating its overall positivity, negativity, or neutrality.
- **Classification Approach:** Based on the sentiment score, reviews are classified into three categories:
    - **Positive:** Scores above 0.05 ○ **Negative:** Scores below -0.05 ○ **Neutral:** Scores between -0.05 and 0.05
- **Implementation:** The code applies sentiment analysis across user and technical datasets (user_data and technical_data), computing sentiment scores and classifying reviews accordingly.

2. Sentiment Distribution Insights:

Visualizations generated from the script provide valuable insights into the sentiment distribution across user and technical reviews:

- **User Sentiment Distribution:**
    - **Positive Sentiment:** Percentage of reviews perceived positively by users.
    - **Negative Sentiment:** Percentage of reviews perceived negatively by users.
    - **Neutral Sentiment:** Percentage of reviews with neutral sentiment.
- **Technical Sentiment Distribution:**
    - **Positive Sentiment:** Percentage of reviews perceived positively by technical evaluators.
    - **Negative Sentiment:** Percentage of reviews perceived negatively by technical evaluators.
    - **Neutral Sentiment:** Percentage of reviews with neutral sentiment.
- **Visualization:** Pie charts visually represent the distribution of sentiment labels, enabling quick comparisons between user and technical reviews.



User Sentiment Distribution

Technical Sentiment Distribution

Conclusion

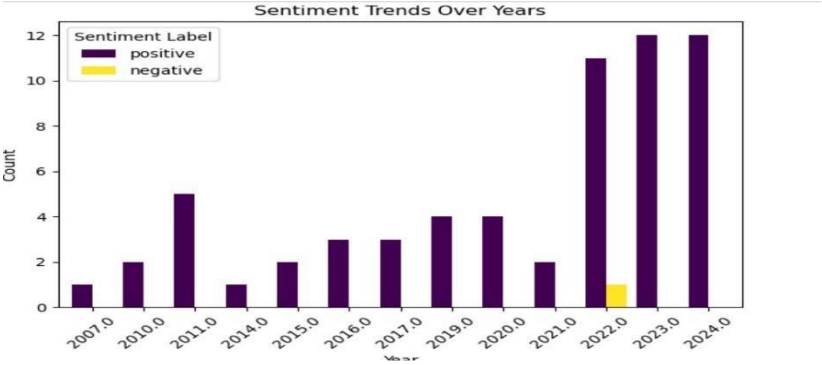
Summary: Throughout this project, we conducted sentiment analysis on user and technical data related to Intel products. We utilized machine learning models and sentiment analysis tools, particularly focusing on VADER for its suitability in handling nuanced sentiment in textual data. Our findings indicate that both user and technical reviews provide valuable insights, with distinct sentiment patterns that can inform product development and customer service strategies.

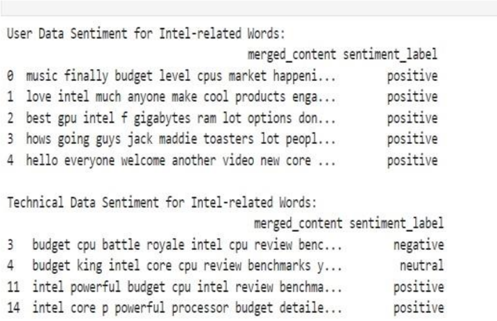
Implications: The implications of our findings suggest that leveraging sentiment analysis can significantly enhance understanding of customer satisfaction and technical performance perceptions. By categorizing sentiments accurately, Intel can prioritize improvements that resonate with both technical users and general consumers.

Challenges: During the project, several challenges were encountered, particularly in ensuring the accuracy and balance of sentiment classification between user and technical reviews. Addressing these challenges involved refining the dataset, optimizing model parameters, and validating results through cross-validation techniques. Additionally, aligning sentiment scores across diverse sources and languages posed computational and interpretative challenges.

Future Work: Looking forward, further research could explore:

1. **Multilingual Sentiment Analysis**: Enhancing models to handle sentiment across different languages more effectively.
2. **Contextual Analysis**: Incorporating contextual cues to improve sentiment classification accuracy, such as analyzing sentiments within specific product features or service interactions.
3. **Real-time Sentiment Monitoring**: Developing frameworks for continuous sentiment monitoring to provide actionable insights promptly.
4. **Sentiment Trends Over Years:**
    a. **Grouping and Aggregation:** Sentiment counts are grouped by year and sentiment label to capture annual trends in consumer sentiment towards Intel products.
    b. **Visualization:** A grouped bar chart is generated using Matplotlib and Seaborn to illustrate how sentiment labels ('positive', 'negative', 'neutral') vary across different years.





```
User Data Sentiment for Intel-related Words:
                                        merged_content sentiment_label
0  music finally budget level cpus market happeni...        positive
1  love intel much anyone make cool products enga...        positive
2  best gpu intel f gigabytes ram lot options don...        positive
3  hows going guys jack maddie toasters lot peopl...        positive
4  hello everyone welcome another video new core ...        positive

Technical Data Sentiment for Intel-related Words:
                                         merged_content sentiment_label
3   budget cpu battle royale intel cpu review benc...        negative
4   budget king intel core cpu review benchmarks y...         neutral
11  intel powerful budget cpu intel review benchma...        positive
14  intel core p powerful processor budget detaile...        positive
```

By addressing these areas, future studies can refine methodologies and tools for sentiment analysis,ultimately aiding in more informed decision-making strategies for Intel products.