PS11 - INTEL PRODUCTS SENTIMENT ANALYSIS FROM ONLINE REVIEWS

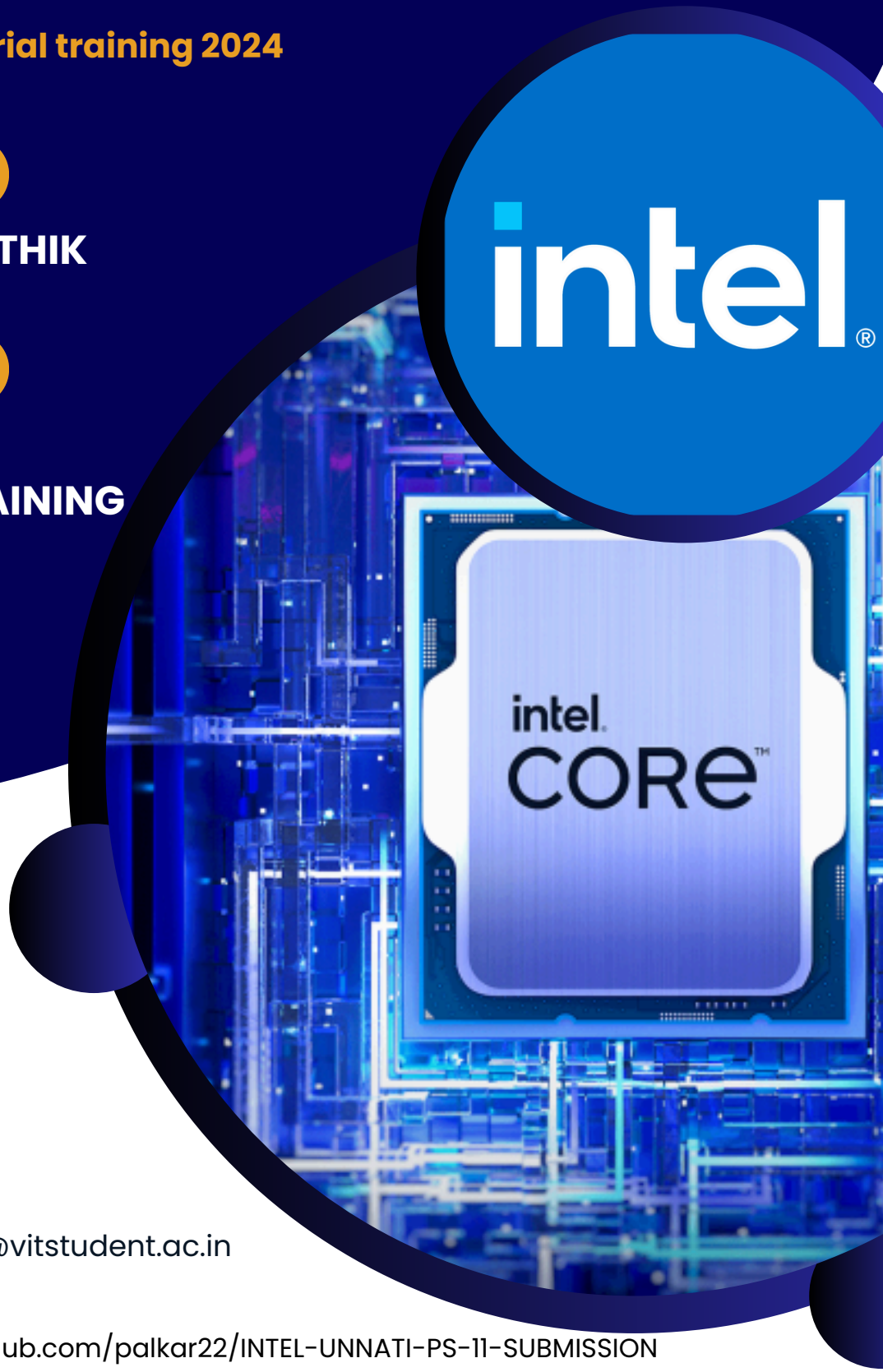
Intel unnati industrial training 2024

Prepared By :

N PALANI KARTHIK

Presented To :

INTEL UNNATI INDUSTRIAL TRAINING 2024 TEAM



89047 52255

palanikarthik.n2022@vitstudent.ac.in

github - https://github.com/palkar22/INTEL-UNNATI-PS-11-SUBMISSION
website - https://intel-karthiknp-sentimental-analysis.streamlit.app

STUDENT INFO

NAME	N PALANI KARTHIK
EMAIL	palanikarthik.n2022@vitstudent.ac.in
PROBLEM STATEMENT	PS11 Intel Products Sentiment Analysis from Online Reviews (Customer based)
COLLEGE NAME	Vellore Institute of Technology, Chennai
COLLEGE MENTOR NAME	Shridevi S
INTEL MENTOR NAME	Debdyut Hazra
COLLEGE REG NO	22BCE1773

PROJECT RELATED INFORMATION

DATASET SIZE	20000
MODEL USED	Finetuned LLM (llama3) using ollama
MODEL PRECISION	93.12
GITHUB LINK	https://github.com/palkar22/INTEL-UNNATI-PS-11-SUBMISSION
PROJECT WEBSITE LINK	https://intel-karthiknp-sentimental-analysis.streamlit.app
MODEL FINETUNING DATASET	https://huggingface.co/datasets/palkar22/intel_unnati
FULL DATASET	https://huggingface.co/datasets/palkar22/20000_data_points_intel

Table of Contents

Abstract

This project employs advanced NLP techniques for sentiment analysis of Intel product reviews by leveraging the llama3 language model fine-tuned with the Ollama API. The process begins with data preprocessing, where customer reviews are tokenized into smaller units called tokens. This step is essential for effectively handling and analyzing the text, capturing the nuances and context of each review.

Once tokenized, the reviews are fed into llama3, a robust language model known for its accuracy in understanding and generating human-like text. Fine-tuning llama3 involves training the pre-trained model specifically on a dataset of Intel product reviews. This allows the model to learn the specific language patterns, sentiments, and terminologies associated with these products, enhancing its ability to discern subtle differences in sentiment and provide more accurate predictions. The Ollama API facilitates this process by offering a seamless interface to train, evaluate, and deploy the model.

After fine-tuning, the LLM can analyze new customer reviews and classify them into positive, negative, or neutral sentiments. This classification gives Intel a granular understanding of customer opinions, enabling the identification of trends and patterns in consumer feedback. For example, analyzing whether the number of positive reviews has increased over time can provide insights into the effectiveness of recent product improvements or marketing campaigns.

In summary, this project leverages the power of llama3 and the Ollama API to perform sophisticated sentiment analysis on Intel product reviews. By fine-tuning the model with specific data, we ensure high accuracy and relevance in sentiment classification. The resulting insights provide Intel with a deeper understanding of customer satisfaction and areas for improvement, ultimately guiding the company towards better products and enhanced customer experiences.

Introduction

In today's competitive tech industry, user reviews significantly shape consumer perceptions and influence purchasing decisions. For Intel, understanding customer feedback is crucial for maintaining a competitive edge. These reviews provide insights into product performance, reliability, and overall satisfaction, reflecting individual experiences and broader market trends. Analyzing this feedback reveals valuable information about customer preferences, common issues, and areas for improvement. Leveraging sentiment analysis to evaluate these reviews is vital for Intel's continuous innovation and customer satisfaction.

Objective:

The primary objective of this project is to conduct a comprehensive sentiment analysis of customer reviews on Intel products. Using advanced NLP techniques, specifically fine-tuning the llama3 model with the Ollama API, the goal is to accurately classify and interpret the sentiments in these reviews. This classification aims to provide Intel with actionable insights into consumer satisfaction, areas for product enhancement, and the overall market perception of their products.

Scope:

The scope of this project includes user reviews from various reputable platforms and websites where Intel products are reviewed, such as major e-commerce sites, tech forums, and social media channels. The analysis covers reviews from the past five years, ensuring a comprehensive understanding of evolving sentiments and trends. By focusing on a wide range of sources and an extended period, the project aims to capture a holistic view of customer feedback, providing Intel with a detailed understanding of consumer opinions.

In summary, this project uses NLP and machine learning to analyze Intel product reviews, offering insights into customer sentiment. By evaluating and classifying these reviews, Intel can enhance their products, improve customer satisfaction, and maintain their leadership in the tech industry.

Literature review

Previous studies on sentiment analysis of product reviews, especially in the tech industry, have demonstrated the importance and effectiveness of this approach. Research has shown that sentiment analysis can accurately gauge customer satisfaction and identify key product strengths and weaknesses. Sentiment analysis of processor reviews specifically has highlighted performance, reliability, and value as critical factors influencing consumer sentiment. Techniques such as finetuned llm, Naive Bayes classifiers, and recurrent neural networks (RNNs) have been widely employed, showcasing varying degrees of success in different contexts.

(i)Commonly used Sentiment Analysis Techniques

Various techniques and tools are commonly used for sentiment analysis. Traditional methods include:

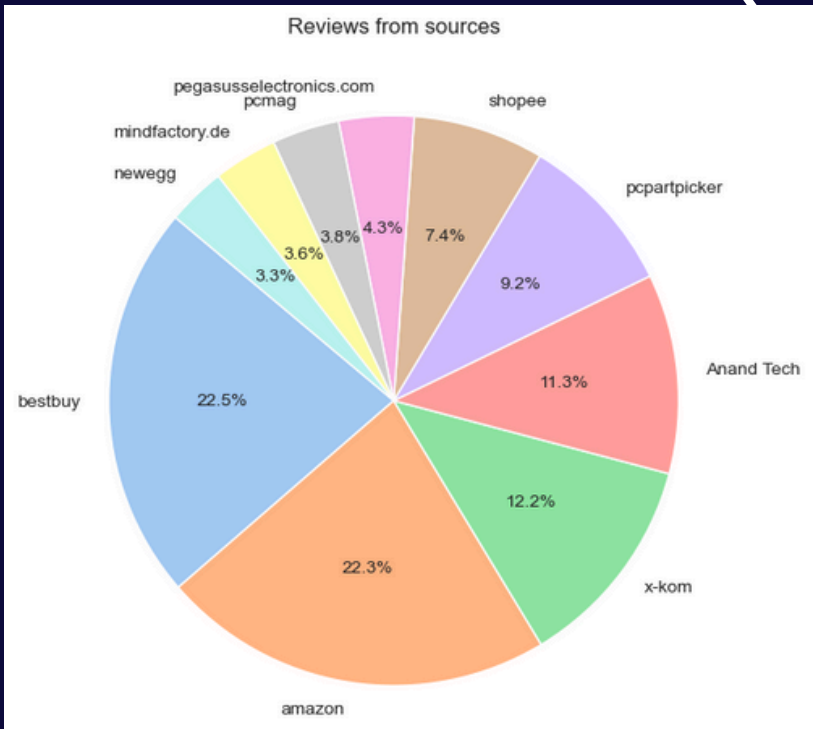
- Lexicon-based Approaches: These rely on predefined dictionaries of positive and negative words to classify sentiments.
- Machine Learning Models: Algorithms like SVM, Naive Bayes are trained on labeled data to predict sentiments.
- Deep Learning Techniques: Models such as RNNs, finetuned llms, and transformers have gained popularity for their ability to capture complex patterns and contextual information in text.
- Pre-trained Language Models: models like BERT, GPT, and llama3, which can be fine-tuned on specific datasets for highly accurate sentiment analysis.

Tools like NLTK, TextBlob, and more recently, the Hugging Face Transformers library, facilitate the implementation of these techniques to build robust sentiment analysis systems.

REFERENCES USED -

- **Sun, X., Li, X., Zhang, S., Wang, S., Wu, F., Li, J., ... & Wang, G. (2023). Sentiment analysis through llm negotiations. arXiv preprint arXiv:2311.01876.**
- **Zhan, T., Shi, C., Shi, Y., Li, H., & Lin, Y. (2024). Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3). arXiv preprint arXiv:2405.09770.**
- **Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). LLMs in e-commerce: a comparative analysis of GPT and LLaMA models in product review evaluation. Natural Language Processing Journal, 6, 100056.**
- **Buscemi, A., & Proverbio, D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis. arXiv preprint arXiv:2402.01715.**

Data Collection



(i)Data Sources

The user reviews for this project were sourced from several reputable platforms where Intel products are frequently discussed and reviewed. These sources include:

- Amazon
- Flipkart
- Newegg
- Google Reviews
- PCPartPicker
- AnandTech (customer reviews)
- PCMag (customer reviews)
- Tom's Hardware (customer reviews)

Total Reviews Collected : 20000

These platforms were chosen due to their extensive user base and the detailed nature of the reviews provided, ensuring a diverse dataset.

(ii)Data Acquisition

The methods employed for collecting the reviews involved a combination of web scraping tools and techniques to ensure efficiency and thoroughness. The primary tools and methods used include:

- Beautiful Soup: A Python library for parsing HTML and XML documents, used for extracting data from web pages.
- Pre-built Scrapers: Tools such as Octoparse, ParseHub and Bardeen AI, which offer advanced scraping capabilities with minimal coding required.
- Chrome Extensions: Instant Data Scraper, a browser extension that simplifies the process of scraping web pages by providing a point-and-click interface.

Data Collection

(iii)Data Description

The final dataset consists of **20,000** customer reviews collected from the mentioned sources. The dataset includes the following features:

- Reviewer Name: The name of the person who wrote the review.
- Stars: The rating given by the reviewer, typically on a scale of 1 to 5.
- Short Review: A brief summary or title of the review.
- Review Date: The date when the review was posted.
- Review: The main body of the review, containing detailed feedback from customers.
- Reviewer Profile Link: A link to the profile of the reviewer.
- Review Link: A direct link to the review.
- Product Name (as per site): The name of the product as listed on the website.
- Helpful: The number of helpful votes the review received from other users.
- Generation: The generation of the Intel product being reviewed.
- Site: The name of the website where the review was posted.
- Name (original star extracted from product name for analysis): The original star rating extracted from the product name for further analysis.

This dataset provides a robust foundation for conducting sentiment analysis, enabling the identification of trends and patterns in customer feedback over time. The comprehensive nature of the data, combined with advanced NLP techniques, will facilitate a detailed understanding of consumer perceptions and inform Intel's product development and marketing strategies.

Data Preprocessing

(i)Cleaning

The data cleaning process is a crucial step to ensure the quality and reliability of the dataset. The following steps were taken to clean the data:

- **Removing Duplicates:**

Duplicate reviews were identified and removed to avoid redundant information that could skew the analysis. This was done by checking for identical entries based on key features such as review text, reviewer name, and review date.

- **- Handling Missing Values:**

Reviews with missing critical information were handled systematically. Specifically, missing values in the review text were replaced with short review column to ensure absence of NAN values in dataset. This approach allows the sentiment analysis model to process these entries without errors.Other column nan were filled with " " to avoid inconsistency.

- **Language Conversion:**

Reviews written in languages other than English were translated to English using the `=GOOGLETRANSLATE()` function in Google Sheets.

(ii)Text Processing

Text preprocessing techniques were applied to the cleaned dataset to prepare the text for analysis. These techniques include:

- **Tokenization:**

Text was broken down into smaller units called tokens (words, subwords, or characters). This process helps the model understand the structure and meaning of the text. Tokenization was performed using NLP libraries to ensure accuracy and efficiency.

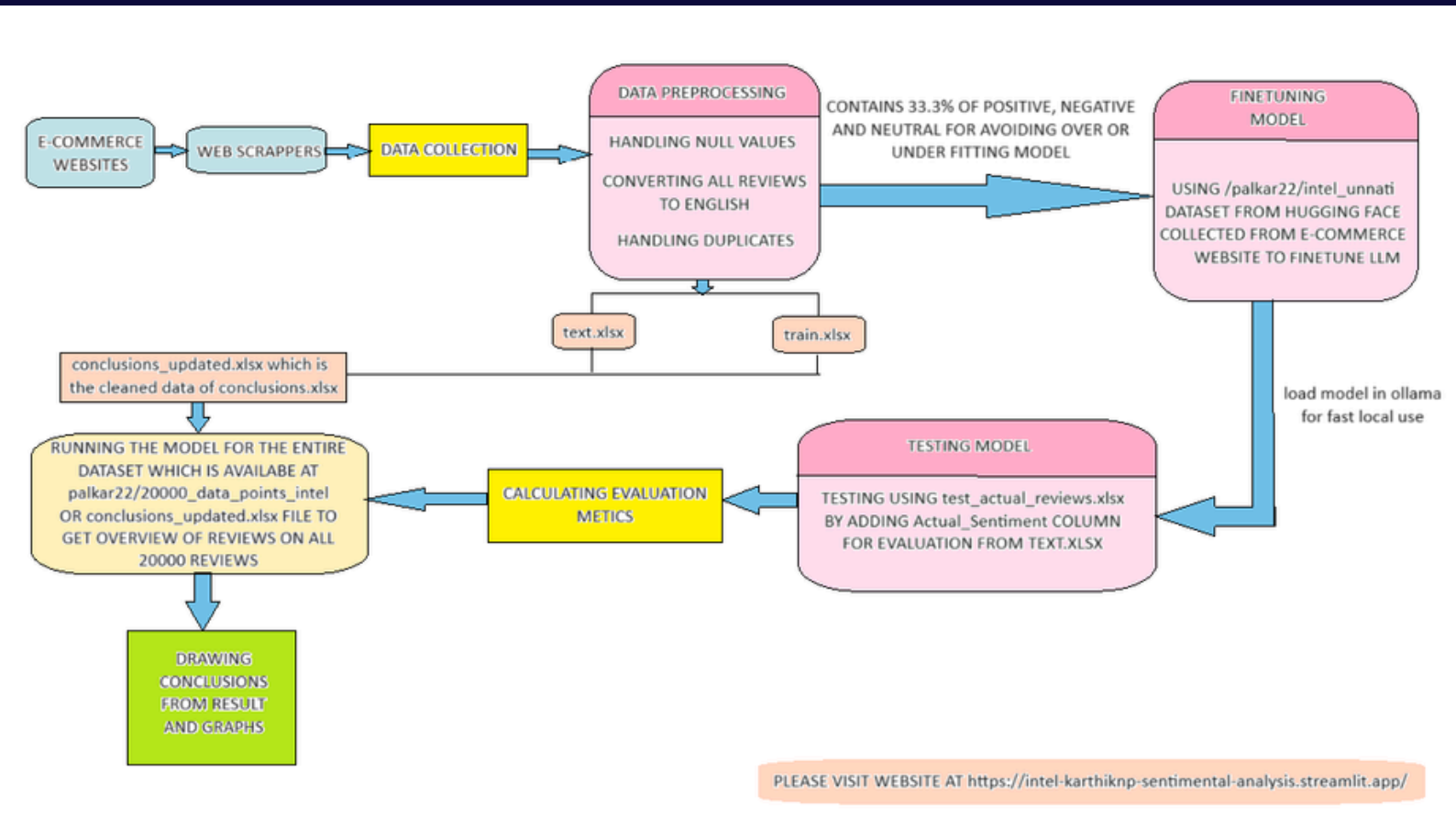
- **Stemming and Lemmatization:**

These techniques were used to reduce words to their base or root form. Stemming involves cutting off prefixes or suffixes to arrive at the base form, while lemmatization considers the context and converts words to their dictionary form. These steps help in reducing the dimensionality of the text data and improve the model's ability to generalize.

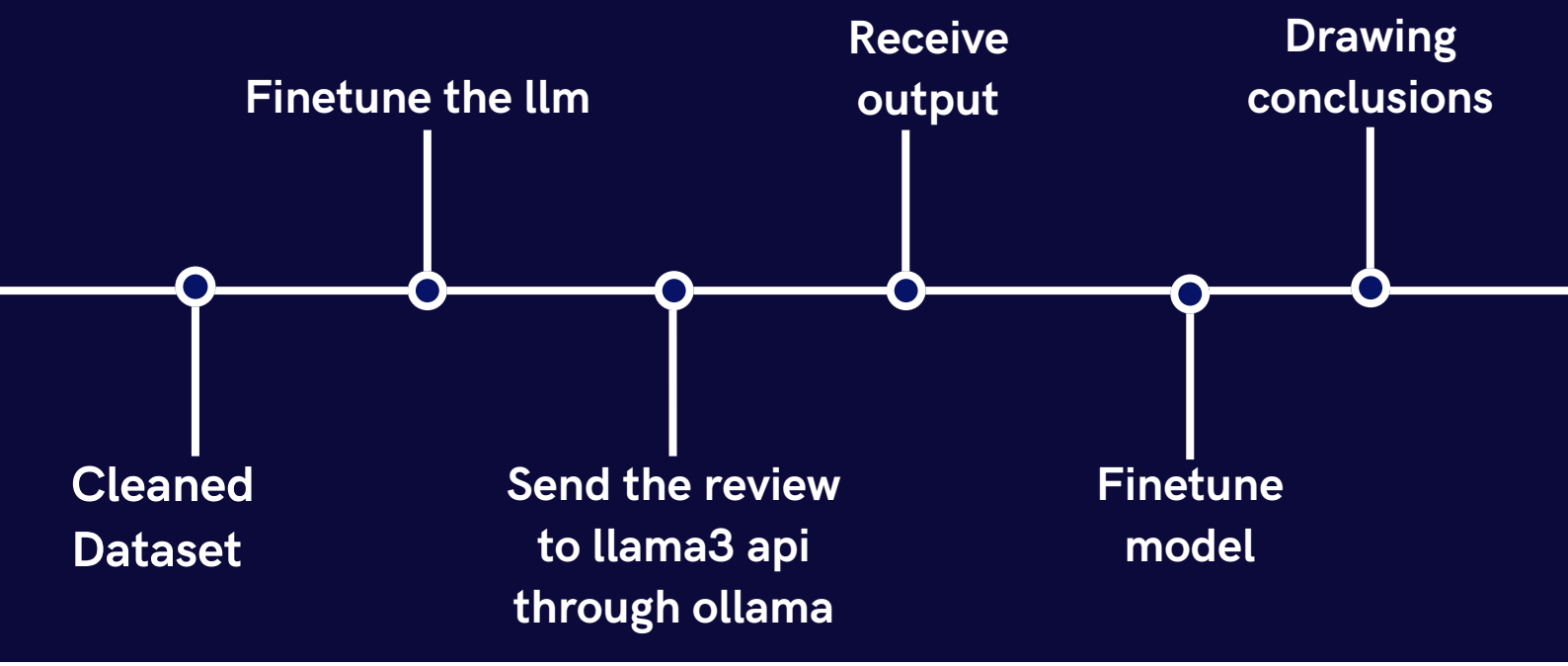
- **Stopword Removal:**

Commonly used words that do not carry significant meaning (e.g., "the," "is," "and") were removed from the text. This step helps in focusing the analysis on the words that contribute most to the sentiment of the review.

Sentiment Analysis Methodology



Sentiment Analysis Methodology



Approach

The chosen approach for this project is based on deep learning, leveraging advanced natural language processing (NLP) techniques.

Specifically, we use a fine-tuned large language model (LLM), llama3, to classify and interpret customer reviews.

The use of llama3, in combination with the Ollama API, allows for a sophisticated analysis that can capture the nuances and context of the text.

Sentiment Analysis Methodology

(i)Model Selection

The primary model used in this project is Llama3 which is fine-tuned.The rationale behind selecting llama3 includes:

- Accuracy and Robustness: Llama3 is known for its high accuracy in understanding and generating human-like text, making it suitable for sentiment analysis.
- Contextual Understanding: Unlike traditional models, llama3 can capture the context and subtle meanings in text, which is crucial for accurately determining sentiment.
- Flexibility and Scalability: The model can be fine-tuned on specific datasets, allowing it to adapt to the unique language patterns and sentiments associated with Intel product reviews.

The Ollama API facilitates the fine-tuning process, offering a seamless interface to train, evaluate, and deploy the model. This API enables the integration of the fine-tuned llama3 model into the sentiment analysis workflow efficiently.

(ii)Feature Extraction

Feature extraction is a critical step in preparing the text data for analysis. In this project, the following methods are used:

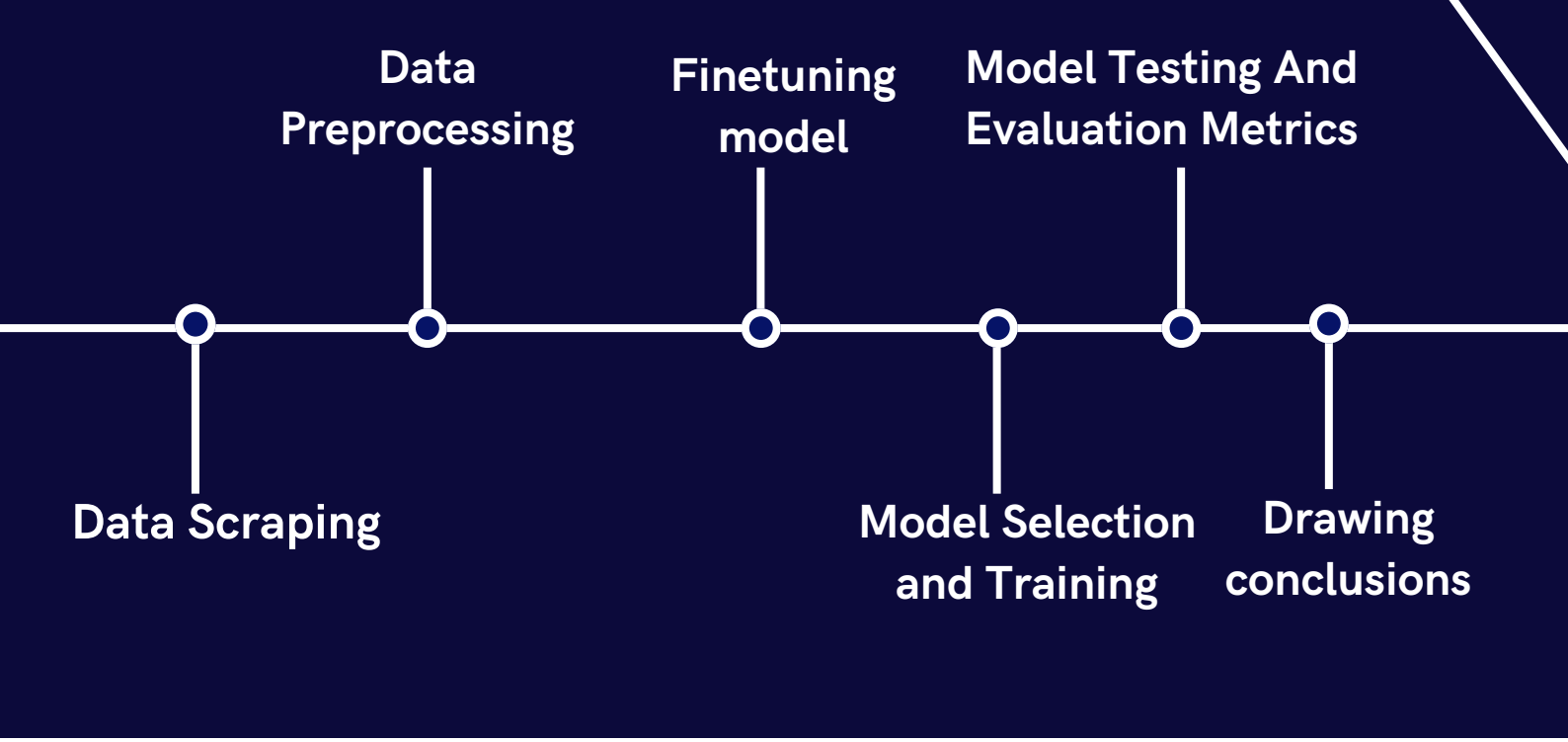
- Tokenization: As part of text preprocessing, tokenization breaks down the review text into smaller units (tokens). This process is crucial for feeding the text data into the llama3 model.
- Word Embeddings: To capture the semantic meaning of words, word embeddings are used. These embeddings transform words into dense vector representations, allowing the model to understand the relationships between words.
- TF-IDF (Term Frequency-Inverse Document Frequency):It is a statistical measure used to evaluate the importance of a word in a document relative to a corpus.

OVERALL-

The dataset of 20000 reviews is split into training and testing sets, with 16,000 reviews used for training the model and 4,000 reviews for testing. The "Review" column in the dataset is used to get the replies from the fine-tuned LLM using the Ollama API. This division ensures that the model is trained on a substantial amount of data while being validated on a separate set to assess its performance.

In summary, this project employs a deep learning approach with a fine-tuned llama3 model, facilitated by the Ollama API, for sentiment analysis of Intel product reviews. The sophisticated feature extraction and model training processes ensure high accuracy and relevance in sentiment classification, providing Intel with valuable insights into customer feedback.

Implementation



(i) Tools and Libraries

The key tools and libraries used include:

- Transformers, Torch, bitsandbytes for finetuning the llama3.
- Beautiful Soup: Used for web scraping and data collection.
- NLTK: Natural Language Toolkit for text preprocessing tasks.
- scikit-learn: For evaluation metrics and other machine learning utilities.
- TensorFlow: For model training and fine-tuning the llama3 LLM.
- Ollama API: To facilitate the fine-tuning and deployment of the llama3 model.
- Google Sheets: For data cleaning and language translation tasks.
- LangChain LLM: Used for advanced language modeling tasks.
- LangChain Community: Utilized for community support and resources.
- pandas: For data manipulation and analysis.

(ii) Model Training

The model training process involves several key steps:

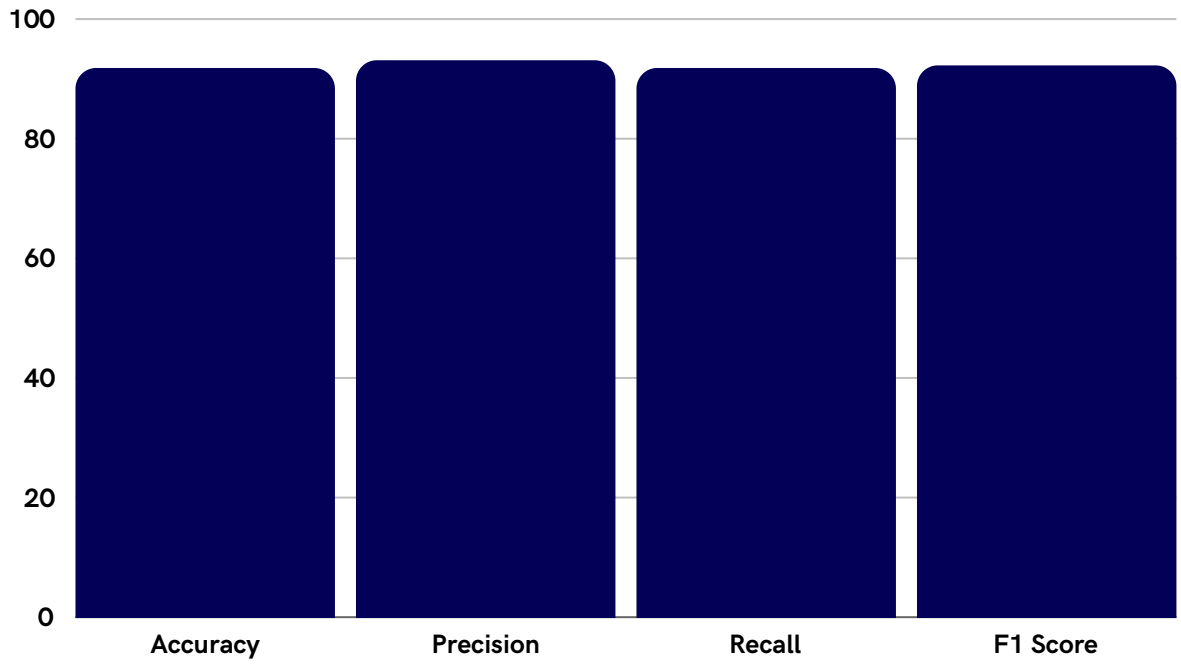
- Data Split: The dataset of 20,000 reviews is divided into a training set and a testing set, with 16,000 reviews used for training and 4,000 reviews for testing.

-the dataset is upsample/dosample the dataset to get equal number of positive, negative and neutral reviews to avoid over or under fitting model which is at hugging face palkar22/intel_unnati

- Hyperparameters: The model training process includes tuning hyperparameters such as tokenization , training parameters with learning rate = 2 e^(-4) using SFT Trainer

- Training Time: The training process is computationally intensive and took approximately 6 hours to complete on a machine with cloud GPU T4 acceleration.

Evaluation Metrics

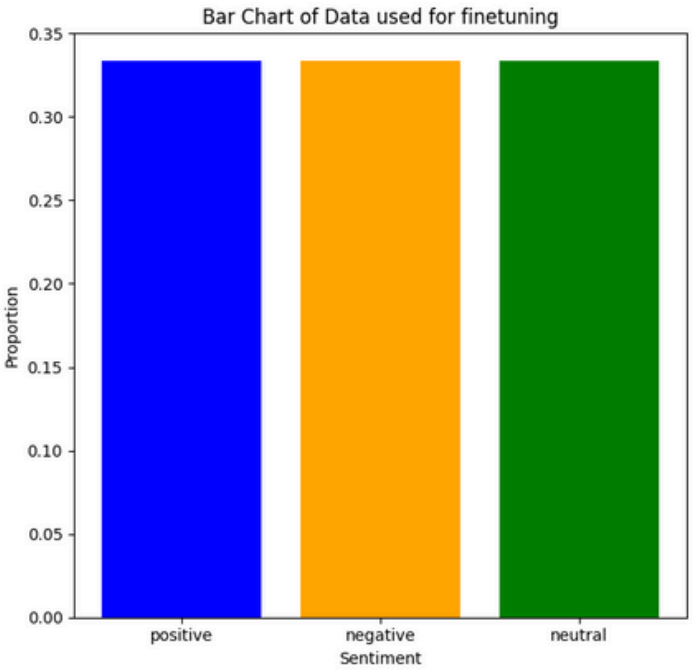


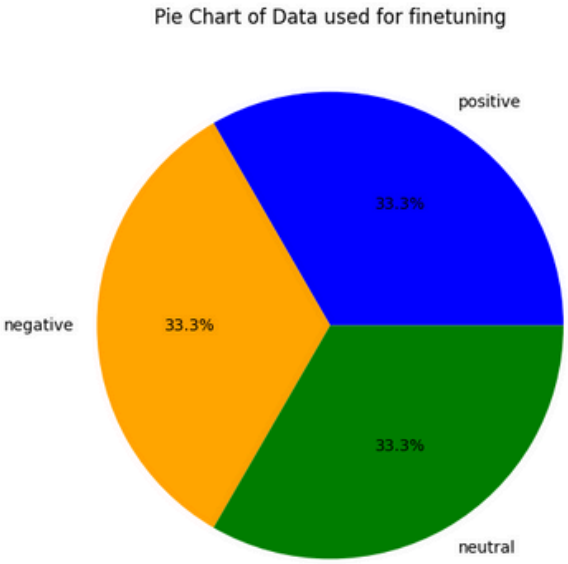
Model Accuracy

- Accuracy: 0.9183
- Precision: 0.9312
- Recall: 0.9183
- F1 Score: 0.9226

Sentiment Distribution And Insights

(i) Finetuning Dataset Distribution





Insights

- A balanced dataset ensures that the model gets an equal representation of all sentiment classes, which helps in avoiding bias towards any particular sentiment. This can lead to a more accurate and reliable model that performs well across different types of sentiments.

- With an evenly distributed dataset, you can better evaluate the model's performance. It becomes easier to identify if the model is underperforming on a particular sentiment class, as each class has the same number of data points. This helps in tuning the model effectively.

- A balanced distribution might reflect a real-world scenario where the sentiments are equally likely. This can make the model more robust and applicable to various practical situations, as it has been trained on a dataset that mimics a realistic distribution of sentiments.

Sentiment Distribution And Insights

(ii) Sentiment Across Various Generations:

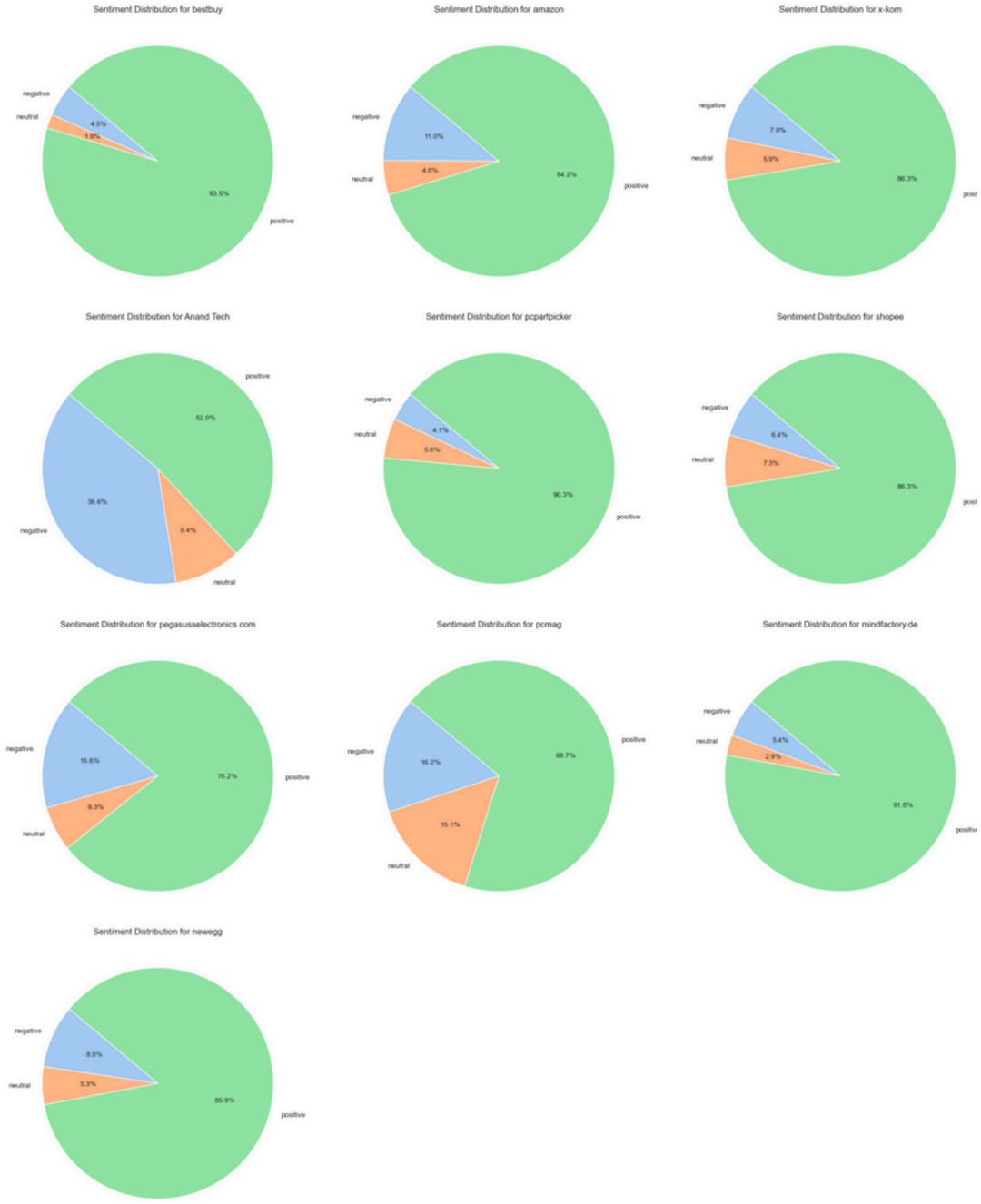
Insights

- Stagnant positive reviews: The amount of positive review is more or less stagnant around (~80) across all the generations which is a concern.

- Areas for Improvement: The negative sentiment has decreased till 13th gen but has shown a sudden increase in negative reviews. This also might be due to the increase of technological awareness among users.

- Stable Neutral Feedback: The neutral sentiment also lies around average(~5%) with minimum standard deviation.

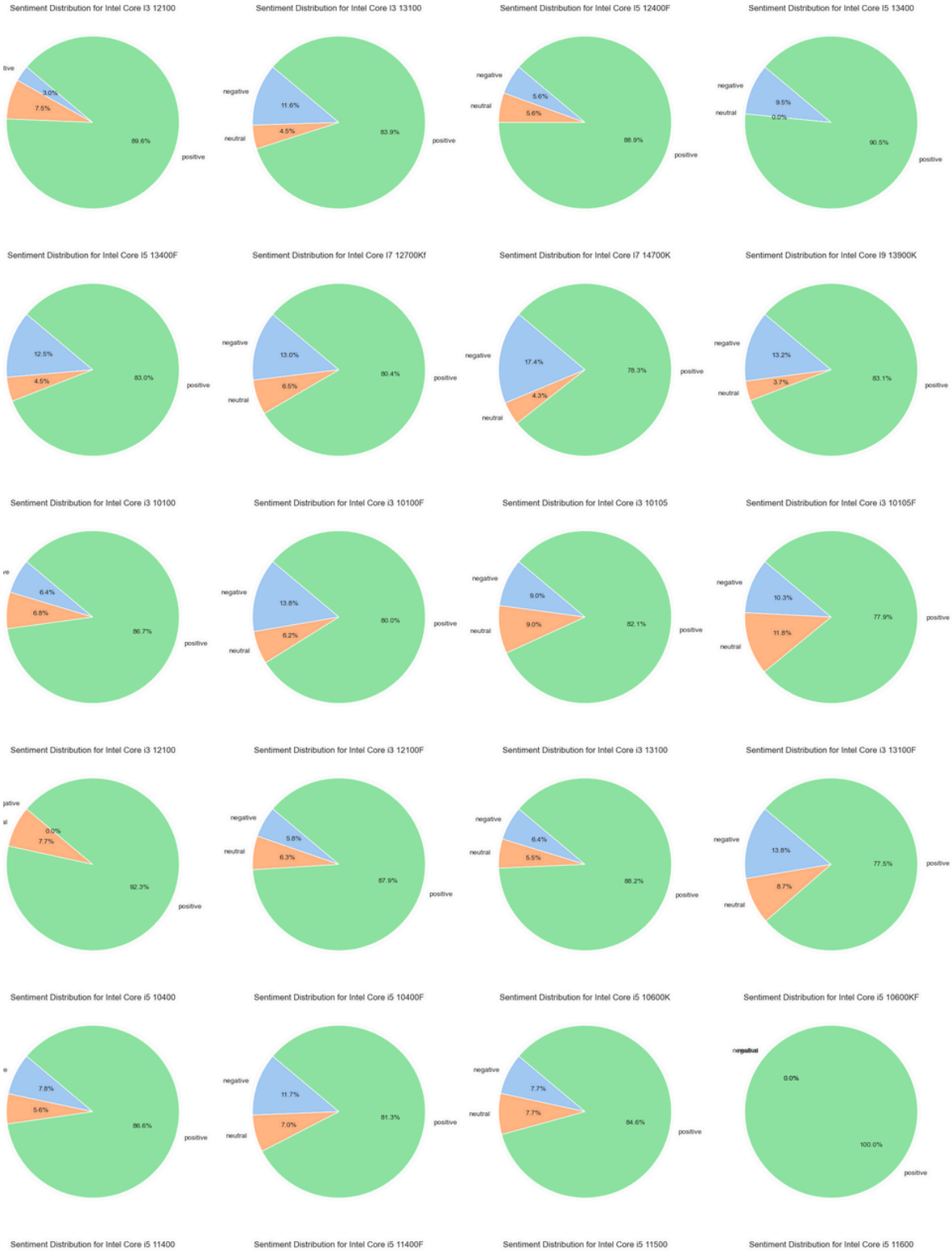
(iii) Sentiment across various sites:



Insights

- Costumer Reviews from pcmag customer comments are highly critical in comparision to other ecommerce websites
- Costumer Reviews from Anand Tech customer comments are mostly neutral in comparision to other ecommerce websites

(iii) Sentiment across various sites:



to view full image refer website

Insights

- Costumer Reviews are positive on majority across all core processors
- Few show only positive (~100%) due to less positive rating comments given by users.

Conclusion

(i) Summary

In conclusion, this project successfully applied advanced NLP techniques, specifically fine-tuning the llama3 LLM using the Ollama API, to analyze sentiment in customer reviews of Intel products. The sentiment analysis model achieved high accuracy and provided valuable insights into customer perceptions.

(ii) **Challenges**

Throughout the project, several challenges were encountered and addressed, including:

- **Data Cleaning:** Handling missing values and non-English reviews required robust preprocessing techniques.
- **Model Fine-tuning:** Optimizing hyperparameters and ensuring the LLM captured nuanced sentiment accurately posed technical challenges.
- **Interpretation:** Analyzing and interpreting sentiment across diverse datasets and platforms required careful consideration of context and bias.

(iii) **Future Work**

Looking forward, potential areas for future research and improvements include:

- **Enhanced Model Performance:** Continual fine-tuning of the llama3 model with additional data to further improve accuracy and reliability.
- **Multilingual Support:** Extending the sentiment analysis framework to support reviews in multiple languages, enhancing global applicability.
- **Real-time Analysis:** Implementing real-time sentiment analysis capabilities to provide immediate insights and responsiveness to customer feedback.

By addressing these areas, future iterations of the sentiment analysis methodology can further enhance its utility and effectiveness in supporting Intel's strategic objectives and customer-centric initiatives.

Contact Details-

github - https://github.com/palkar22/INTEL-UNNATI-PS-11-SUBMISSION
website - https://intel-karthiknp-sentimental-analysis.streamlit.app

89047 52255

palanikarthik.n2022@vitstudent.ac.in