|  |  |
| --- | --- |
| Internship Project Title | RIO-125: Automate Sentiment Analysis of Textual Comments and Feedbacks |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Mr. Debashis Roy |
| Name of the Institute | Vishwakarma University, Pune |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 19/06/2024 | 18/07/2024 | 128 Hrs. | Python | Jupyter Notebook -Pandas, NLTK (Natural Language Toolkit), TextBlob, Scikit-learn, Matplotlb, Seaborn, GridSearchCV, Piprline, CountVectorizer, LSTM, Dense, StandardScaler, Tokenizer, Sequential Embedding, etc. |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**ACKNOWLEDGEMENT**

I'm truly grateful for the unwavering support and guidance extended to me throughout my project, RIO-125: Automate Sentiment Analysis of Textual Comments and Feedbacks. I want to express my heartfelt appreciation to my industry mentor, Mr. Debashis Roy from TCS-iON, and my academic mentor, Prof. Shriprada Chaturbhuj from Vishwakarma University. Their constant motivation played a pivotal role in my journey.

Additionally, I extend my sincere thanks to TCS-iON and Vishwakarma University for granting me this invaluable opportunity, which has enriched my understanding of the industry landscape. I want to emphasize that I completed the project independently, without any external assistance.

**OBJECTIVE**

To develop advanced deep learning algorithms aimed at accurately detecting various types of sentiments expressed within English sentences or lengthy paragraphs, with the ultimate goal of precisely predicting the overall sentiment conveyed by the entire text.

**INTRODUCTION/DESCRIPTION OF THE INTERNSHIP**

This internship is an exciting journey into the realm of teaching computers to understand emotions conveyed in written text. We'll explore techniques in deep learning, empowering computers to discern whether a piece of writing evokes happiness, sadness, or anything in between. Our goal is to equip our computer counterparts with the ability to accurately interpret the overall emotional tone of any text, whether it's a brief message or a lengthy essay. It's akin to giving our computer buddies the superpower to read between the lines and grasp the nuances of human emotions expressed through words. Through this endeavor, we're delving deep into the fascinating world of emotional comprehension in written language, striving to enhance our computer companions' understanding beyond simple emotions to capture the diverse range of sentiments people convey through their writing.

**INTERNSHIP ACTIVITIES**

The internship activities include the following:

1. Research and study sentiment analysis and deep learning algorithms.
2. Collect diverse textual data, preprocess it, and prepare it for training.
3. Experiment with various deep learning architectures for sentiment analysis.
4. Train models, evaluate their performance, and iterate for improvement.
5. Explore hyperparameter tuning techniques for model optimization.
6. Validate models generalization ability with unseen data.
7. Document the process and prepare reports summarizing findings.
8. Engage in discussions, workshops, and team meetings for collaborative learning.

**APPROACH/METHODOLOGY**

1. **Text Preprocessing**:
   * Tokenization: Breaking down the text into individual words or tokens.
   * Part-of-Speech (POS) Tagging: Assigning grammatical information (like noun, verb, adjective) to each token.
   * Lemmatization: Reducing words to their base or dictionary form.
   * Removing Punctuation and Stopwords: Eliminating non-essential words and characters.
2. **Sentiment Analysis**:
   * Using the TextBlob library to analyze sentiment polarity, which determines whether the sentiment expressed in the text is positive, neutral, or negative.
3. **Regression Analysis**:
   * Using Linear Regression to analyze the relationship between sentiment polarity and product ratings.
4. **Hyperparameter Tuning**:
   * Utilizing GridSearchCV to search for the best hyperparameters for Linear Regression, optimizing the model's performance.
5. **N-gram Modeling**:
   * Including both unigrams and bigrams in the text vectorization process to capture more contextual information.
6. **Deep Learning**:
   * Implementing a Long Short-Term Memory (LSTM) model for sentiment classification.
   * Enhancing the model by combining LSTM with a Convolutional Neural Network (CNN) in an ensemble approach to improve performance.
7. **Visualization**:
   * Utilizing Matplotlib and Seaborn for visualizing the regression analysis results, such as the relationship between sentiment polarity and ratings.
   * Plotting the training and validation loss curves during the training of deep learning models to monitor model performance and prevent overfitting.

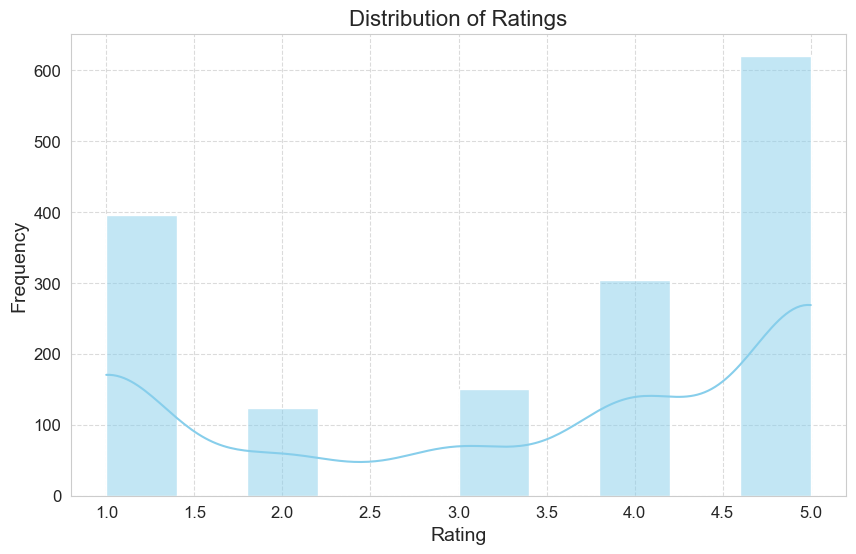
**ASSUMPTIONS**

1. **Homogeneity of Sentiment Analysis**:
   * The assumption that sentiment analysis can be accurately performed based solely on the polarity score generated by TextBlob. This assumption implies that the sentiment expressed in a review is sufficiently captured by a single numerical value, disregarding potential nuances and complexities in language use and context.
2. **Linearity in Regression Analysis**:
   * The assumption of linearity between sentiment polarity and product ratings in the regression analysis. This assumption suggests that the effect of sentiment on ratings follows a linear relationship, implying that a unit change in sentiment polarity leads to a consistent change in the rating score. Nonlinear relationships between sentiment and ratings may not be adequately captured by linear regression.
3. **Effectiveness of Hyperparameter Tuning**:
   * The assumption that hyperparameter tuning, particularly for the fit\_intercept parameter in linear regression, significantly impacts model performance. This assumption presupposes that optimizing hyperparameters can lead to substantial improvements in the model's ability to predict ratings based on sentiment polarity, potentially overlooking other factors contributing to model performance.
4. **Generalization of Deep Learning Models**:
   * The assumption that the ensemble of LSTM and CNN models will generalize well to unseen data and outperform individual models. This assumption relies on the premise that combining different neural network architectures can effectively capture diverse features and patterns in the text data, leading to improved sentiment classification performance across various review texts.
5. **Effectiveness of Text Preprocessing**:
   * The assumption that the text preprocessing steps, including tokenization, POS tagging, lemmatization, and stopwords removal, adequately capture the relevant linguistic information for sentiment analysis. This assumption implies that the preprocessing techniques used effectively transform raw text data into a format suitable for subsequent analysis, potentially overlooking the impact of preprocessing choices on model performance and interpretability.

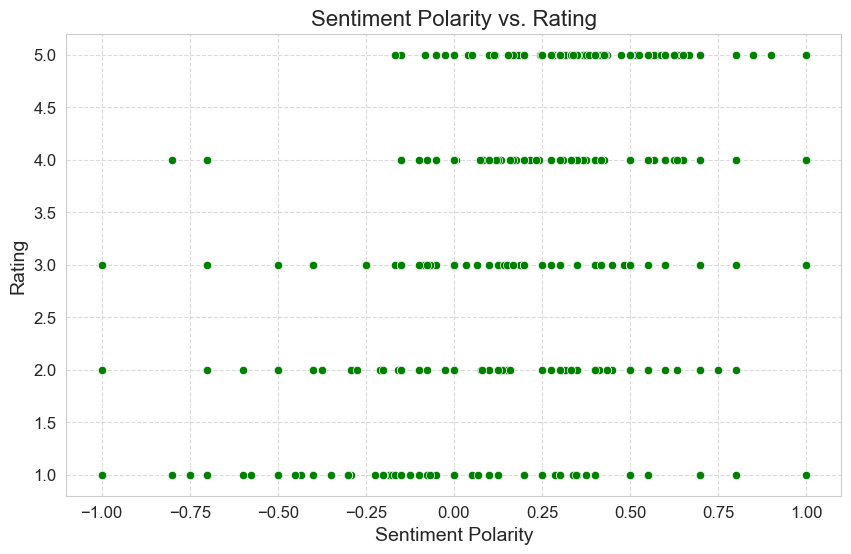
**EXCEPTIONS/EXCLUSIONS**

1. **Exceptions**:
   * **Assumption of Linearity**: Linear regression assumes a linear relationship between sentiment polarity and product ratings. However, real-world relationships may be nonlinear, leading to potential inaccuracies in predictions.
   * **Assumption of Homogeneous Sentiment Analysis**: Sentiment analysis based solely on polarity scores may overlook nuances in language and context, leading to misinterpretations of sentiment in certain cases.
2. **Exclusions**:
   * **Domain-Specific Factors**: The code doesn't consider domain-specific factors that may influence sentiment and ratings, such as product type, brand reputation, or user demographics. Ignoring these factors could limit the model's predictive accuracy and generalizability.
   * **Cultural and Contextual Variations**: The code doesn't account for cultural or contextual variations in language use and sentiment expression, which could affect the effectiveness of sentiment analysis across different regions or communities. Incorporating cultural and contextual insights could enhance the model's robustness and applicability in diverse contexts.

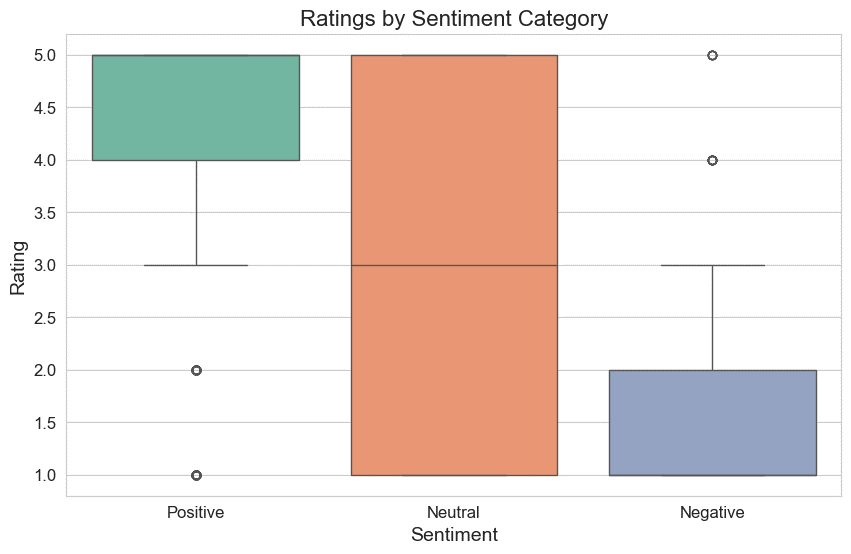
**CHARTS, TABLE AND DIAGRAMS**

****

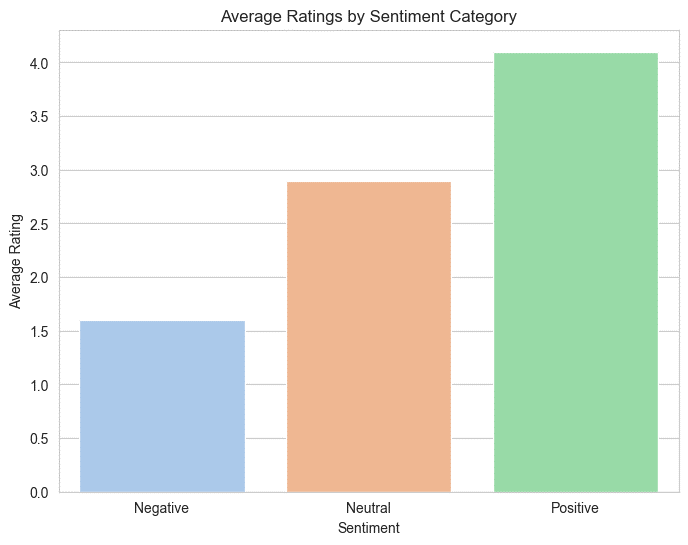
Histogram of customer ratings for various headphones on the online shopping platform. The graph displays ratings ranging from 1.0 to 5.0 on the x-axis and their corresponding frequencies on the y-axis. Notably, there is a scarcity of ratings between 2.0 and 3.0, with a slight increase noted at 3.0. Subsequently, there is a notable rise in frequency at 4.0, followed by a dip at 4.5, and a sharp peak at 5.0.



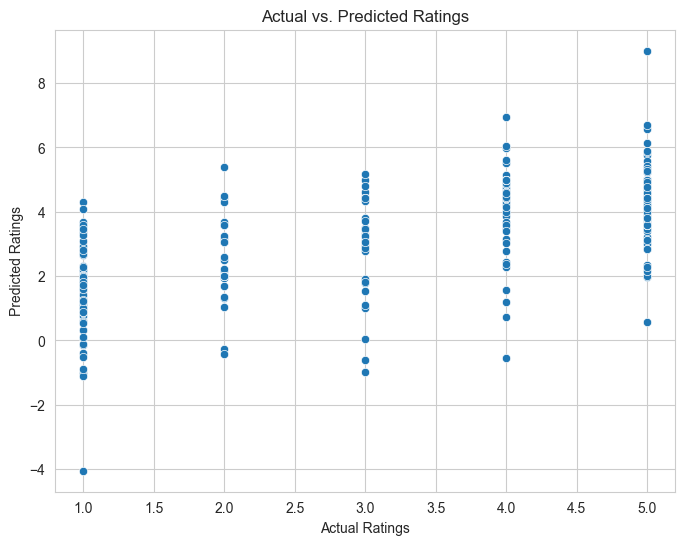
Scatter Plot titled "Sentiment Polarity vs. Rating" which illustrates the relationship between the sentiment polarity (ranging from -1.00 to 1.00) and the rating (ranging from 1.0 to 5.0). Green dots represent data points, indicating varying sentiment polarities and ratings. Data points cluster around integer rating levels, with denser concentrations near zero sentiment polarity. The highest concentration of points aligns with the top rating (5.0), predominantly displaying positive sentiment polarities. Lower ratings (1.0 to 2.0) exhibit fewer data points dispersed across both negative and positive sentiment polarities, suggesting a broader range of opinions when ratings are lower.



Boxplot titled "Ratings by Sentiment Category" which illustrates the average ratings assigned to Positive, Negative, and Neutral sentiment categories. Positive sentiment registers the highest average rating, followed by Neutral and Negative sentiments. Notably, outliers are visible within the Positive and Neutral categories, signifying substantial deviations from the average ratings within these groups.



Vertical Bar Chart titled "Average Ratings by Sentiment Category," depicting the average customer sentiments categorized as Positive, Negative, or Neutral. The chart highlights that Positive sentiment garners the highest average rating, suggesting that customers are notably satisfied with the products or services. Following Positive sentiment, Neutral sentiments range between 2.5 and 3.0, indicating moderate satisfaction. Negative sentiments, ranging between 1.5 and 2, represent the lowest average ratings, suggesting dissatisfaction among customers.



Scatterplot for "Actual Ratings" against "Predicted Ratings" using blue dots to denote individual data points. Actual ratings ranging from 1.0 to 5.0 are plotted on the x-axis, while predicted ratings are represented on the y-axis within the same range. Notably, for lower actual ratings, such as 1.0, the predicted ratings exhibit wide dispersion, ranging from -4 to over 6.0, indicating varying levels of accuracy in the prediction model. Conversely, as actual ratings increase, particularly for ratings of 1.0 and 3.0, predicted ratings cluster closer to the actual values with less variation. Overall, while the model demonstrates greater accuracy for higher actual ratings, it lacks consistency across all rating levels, suggesting room for improvement.

**ALGORITHMS**

1. **Tokenization**:
   * Algorithm: Word Tokenization using libraries like NLTK or spaCy.
   * Method: Splitting the text into individual words or tokens, preserving the semantic meaning of each word. For subword tokenization, algorithms like Byte Pair Encoding (BPE) or WordPiece can be used to handle rare or out-of-vocabulary words effectively.
2. **Lowercasing and Punctuation Removal**:
   * Algorithm: String Manipulation or Regular Expressions.
   * Method: Converting all text to lowercase ensures uniformity and reduces the vocabulary size. Removing punctuation marks eliminates non-alphanumeric characters that do not contribute to sentiment analysis.
3. **Stopwords Removal**:
   * Algorithm: Stopword Removal using predefined lists.
   * Method: Removing common stopwords helps in reducing noise and improving the quality of features used for sentiment analysis. Libraries like NLTK or spaCy provide built-in lists of stopwords for different languages.
4. **Stemming or Lemmatization**:
   * Algorithm: Porter Stemmer for stemming, WordNet Lemmatizer for lemmatization.
   * Method: Stemming reduces words to their base or root form by removing suffixes, while lemmatization reduces words to their canonical form based on a dictionary of word forms. Lemmatization is preferred for sentiment analysis tasks as it retains the semantic meaning of words.
5. **Model Architecture**:
   * Algorithm: Logistic Regression, Multinomial Naive Bayes, or Bernoulli Naive Bayes.
   * Method: Logistic Regression is a linear classification model that can handle binary or multiclass sentiment classification tasks effectively. Naive Bayes classifiers, such as Multinomial Naive Bayes and Bernoulli Naive Bayes, are probabilistic models based on Bayes' theorem and are well-suited for text classification tasks.
6. **Implementation using scikit-learn**:
   * Algorithm: Utilizing scikit-learn's **CountVectorizer** for feature extraction and **LogisticRegression**, **MultinomialNB**, or **BernoulliNB** for classification.
   * Method: Transforming preprocessed text data into numerical features using **CountVectorizer**, which converts text into a matrix of token counts. Then, training a classification model using **LogisticRegression** for binary classification or **MultinomialNB** / **BernoulliNB** for multinomial or binary Naive Bayes classification.

**CHALLENGES & OPPORTUNITY**

The different challenges and opportunities that were associated are as follows:

1. **Sentiment Analysis Basics**:
   * Challenges: Understanding people's feelings from text can be tricky because language is complex and can mean different things in different contexts.
   * Opportunities: Knowing how people feel about products, services, or topics can help businesses make better decisions and improve customer satisfaction.
2. **Different Sentiment Analysis Approaches**:
   * Challenges: Different methods of figuring out sentiment have their own problems. Some might miss important details, while others need a lot of data and time to work.
   * Opportunities: By combining different methods, we can get a more accurate picture of sentiment. New technologies are making sentiment analysis better all the time.
3. **Rule-Based Methods**:
   * Challenges: Rules for understanding sentiment might not cover every situation, and they can be hard to make and keep up-to-date.
   * Opportunities: These methods are straightforward and easy to understand, and they work well in some cases, especially where the rules are clear.
4. **Lexicon-Based Approaches**:
   * Challenges: Lists of words and their sentiment can't capture all the ways people express feelings, and they might miss the meaning of some words.
   * Opportunities: These approaches are quick and can give us a basic understanding of sentiment, especially when we have a lot of text to analyze.
5. **Machine Learning Methods**:
   * Challenges: Teaching computers to understand sentiment needs lots of examples, and sometimes they make mistakes or get confused.
   * Opportunities: With enough examples and some clever math, computers can learn to understand sentiment pretty well. They can handle lots of different situations and languages.
6. **Dataset Selection and Preprocessing**:
   * Challenges: Finding good examples of text with different feelings can be hard, and getting the text ready for analysis can take time.
   * Opportunities: Having a diverse collection of examples helps computers learn better, and cleaning up the text makes the analysis more accurate.

**RISKS vs REWARDS**

1. **Sentiment Analysis Basics**:
   * Risk: Misinterpreting sentiment due to language nuances and subjective viewpoints can result in inaccurate insights and flawed decisions.
   * Reward: Obtaining valuable insights into customer opinions, market trends, and brand perception can enhance decision-making and boost customer satisfaction.
2. **Different Sentiment Analysis Approaches**:
   * Risk: Relying on a single approach may yield incomplete or biased results, while combining multiple methods can increase complexity and require significant resources.
   * Reward: Employing various approaches allows for a more thorough understanding of sentiment, improving the accuracy and reliability of the analysis.
3. **Rule-Based Methods**:
   * Risk: Depending heavily on predefined rules can lead to oversimplified or inaccurate sentiment analysis, especially for nuanced or context-specific sentiments.
   * Reward: Rule-based methods are transparent and interpretable, providing clear guidelines for sentiment analysis and enabling straightforward implementation in specific domains.
4. **Lexicon-Based Approaches**:
   * Risk: Relying on sentiment lexicons may result in limited coverage of sentiment expressions, missing out on nuanced or domain-specific sentiments.
   * Reward: Lexicon-based approaches are efficient and can quickly provide insights into sentiment, serving as a useful starting point for sentiment analysis with limited resources.
5. **Machine Learning Methods**:
   * Risk: Machine learning models require large labeled datasets for training, and inaccuracies or biases in these datasets can lead to poor performance and unreliable predictions.
   * Reward: With proper training and validation, machine learning models can achieve high accuracy in sentiment analysis, offering scalable and adaptable solutions for diverse applications.
6. **Dataset Selection and Preprocessing**:
   * Risk: Choosing biased or inadequate datasets can result in models that generalize poorly or produce biased outcomes, while improper preprocessing can introduce noise and distort results.
   * Reward: Well-curated datasets and effective preprocessing techniques enhance the quality and reliability of sentiment analysis models, leading to more accurate insights and better-informed decisions.

**REFLECTION ON THE INTERNSHIP**

1. **Learning Opportunities**:
   * The internship offered an excellent chance to learn about various aspects of sentiment analysis, including different approaches, methodologies, and algorithms.
   * Researching sentiment analysis basics and exploring different approaches helped build a strong foundation in natural language processing concepts.
2. **Hands-on Experience**:
   * Engaging in practical tasks such as dataset selection, preprocessing, and model implementation using machine learning libraries like scikit-learn provided valuable hands-on experience.
   * Working on text data preprocessing and training sentiment analysis models provided insight into the complexities and challenges of real-world sentiment analysis projects.
3. **Problem-Solving Skills**:
   * Tackling challenges related to dataset selection, preprocessing, and model selection required critical thinking and problem-solving skills.
   * Addressing issues like dataset biases, model performance, and feature engineering enhanced analytical thinking and problem-solving abilities.
4. **Collaboration and Communication**:
   * Collaborating with mentors and colleagues to discuss ideas, seek feedback, and troubleshoot issues fostered teamwork and improved communication skills.
   * Presenting findings and insights from sentiment analysis tasks to stakeholders helped develop presentation and communication skills.
5. **Reflection on Challenges**:
   * Overcoming challenges related to dataset selection, preprocessing, and model implementation highlighted the importance of thorough planning, attention to detail, and adaptability.
   * Dealing with uncertainties and limitations in sentiment analysis methods emphasized the need for continuous learning and exploration of innovative techniques.
6. **Future Growth Opportunities**:
   * The internship experience laid a strong foundation for further exploration and specialization in sentiment analysis and natural language processing.
   * Building on the knowledge and skills gained during the internship, there are opportunities for future growth in research, industry projects, and advanced studies in related fields.

Top of Form

**RECOMMENDATIONS**

1. **Stay Updated** : Continuously follow the latest trends and techniques in sentiment analysis by reading research papers and taking online courses.
2. **Gain Practical Experience** : Engage in personal projects and participate in competitions to apply your knowledge in real-world scenarios.
3. **Network** : Connect with professionals and peers in the field to share ideas and seek mentorship opportunities.
4. **Develop a Portfolio** : Create a well-curated portfolio to showcase your projects and expertise in sentiment analysis.
5. **Specialize** : Focus on specific areas within sentiment analysis to deepen your expertise.
6. **Invest in Professional Development** : Attend workshops and seek opportunities for both personal and professional growth.

**OUTCOME/CONCLUSIONS**

1. **Understanding Sentiment Analysis Basics**:
   * Acquired foundational knowledge about sentiment analysis, its significance, and applications in natural language processing.
2. **Exploring Different Sentiment Analysis Approaches**:
   * Learned about rule-based, lexicon-based, machine learning, and deep learning methods for sentiment analysis, understanding their advantages and limitations.
3. **Diving Deeper into Rule-Based and Lexicon-Based Methods**:
   * Examined how rule-based methods use predefined rules and patterns, and how lexicon-based approaches utilize sentiment lexicons or dictionaries to assign sentiment scores to words.
4. **Studying Machine Learning Methods for Sentiment Analysis**:
   * Explored machine learning techniques such as logistic regression, Naive Bayes, and Support Vector Machines for sentiment analysis, focusing on their implementation and performance.
5. **Identifying Suitable Datasets and Preprocessing Data**:
   * Investigated the challenges and opportunities in dataset selection and preprocessing for sentiment analysis, including tokenization, stop word removal, and stemming/lemmatization.
6. **Choosing and Implementing Model Architecture**:
   * Selected basic model architectures like logistic regression or Naive Bayes for sentiment classification and implemented them using scikit-learn in Python.

**ENHANCEMENT SCOPE**

1. **Advanced Models** : Explore the use of RNNs, CNNs, or transformer-based models to achieve better performance in sentiment analysis.
2. **Domain-Specific Analysis** : Focus on analyzing sentiment within specific domains, such as healthcare or finance, for more tailored insights.
3. **Fine-Tuning Pretrained Models** : Use transfer learning to adapt pretrained models to specific sentiment analysis tasks.
4. **Aspect-Based Analysis**: Perform sentiment analysis at a more detailed level, concentrating on specific aspects or entities within the text.
5. **Multimodal Analysis**: Integrate text with other modalities, such as images or audio, for a more comprehensive sentiment analysis.
6. **Real-Time Analysis**: Create systems capable of analyzing sentiment in streaming data in real-time.
7. **Ethical Considerations**: Address biases and ethical concerns in sentiment analysis algorithms to ensure fairness and protect privacy.

**LINK TO THE EXECUTABLE FILE**

**Repository Link:** [**https://github.com/Tejas180408/RIO-125**](https://github.com/Tejas180408/RIO-125)

**RESEARCH QUESTIONS AND RESPONSES**

**Question:** How does sentiment analysis contribute to understanding customer feedback in the e-commerce industry?

**Responses:**

1. Sentiment analysis helps e-commerce businesses gain insights into customer opinions and emotions expressed in product reviews, allowing them to identify trends, strengths, and areas for improvement.
2. By analyzing sentiment, e-commerce companies can gauge customer satisfaction levels, identify common pain points, and tailor their products and services to meet customer expectations more effectively.
3. Sentiment analysis enables e-commerce platforms to automate the process of sorting and categorizing large volumes of customer feedback, making it easier to prioritize and address urgent issues.
4. Understanding sentiment in reviews allows e-commerce businesses to monitor brand reputation and identify potential PR crises before they escalate, enabling proactive reputation management strategies.
5. Sentiment analysis can also be used to personalize customer experiences by identifying individual preferences and tailoring recommendations and marketing messages accordingly.