Sentiment analysis AI system

A Project Report

submitted in partial fulfillment of the requirements

Of

Track Name: AIML Fundamentals With Cloud Computing And Gen AI

by

Name: KISHORE M G, Email Id: kishoremurali916@gmail.com

Under the Guidance of

Name of Guide (P.Raja, Master Trainer)

LINKS

GITHUB : https://github.com/Kishoremurali7/Kishore-MG-Project-Naan-Mudhalvan

YOUTUBE: https://youtu.be/FCYTqTklh1k?si=FhwCNn9KH3DBmigR

ACKNOWLEDGEMENT

"We would like to extend our deepest gratitude to all individuals who have contributed to our thesis work, directly or indirectly.

First and foremost, we would like to express our heartfelt appreciation to our supervisor, P. Raja and P. Jermia Arockia pravin, for their exceptional guidance and mentorship. Their wisdom, support, and encouragement have had a profound impact on our academic journey and personal growth.

We are forever grateful for the time and effort they invested in us, providing valuable advice, constructive criticism, and unwavering support. Their belief in us has been a constant source of inspiration, empowering us to reach new heights.

Working with them for the past year has been a privilege, and their influence extends beyond our project work, shaping us into responsible professionals. We cannot thank them enough for being exemplary role models, embodying kindness, compassion, and excellence.

Thank you again, P. Raja and P. Jermia Arockia Pravin, for being incredible mentors and guides."……..

ABSTRACT of the Project

This project focuses on analyzing Amazon's fine food reviews dataset, which includes ~500,000 reviews spanning over a decade. The primary objective is to gain insights into customer sentiment and preferences to enhance product recommendation systems and customer satisfaction. The problem addressed is the overwhelming volume of data, which makes it difficult for consumers and businesses to quickly extract meaningful insights.

The project begins with data preprocessing, including handling missing values, tokenizing text, and normalizing ratings. Sentiment analysis is conducted to classify reviews as positive, neutral, or negative using Natural Language Processing (NLP) techniques and machine learning algorithms. We employ models like Naive Bayes and Support Vector Machines, leveraging TF-IDF for feature extraction. Word clouds and sentiment distributions provide visual insights, and topic modeling helps identify recurring themes.

Key results reveal high customer satisfaction trends, with notable insights into product types driving positive or negative sentiments. Our analysis indicates that specific product attributes, such as packaging and flavor, significantly influence ratings. The models achieve over 85% accuracy in sentiment classification, validating their effectiveness.

In conclusion, this study demonstrates the potential of sentiment analysis in improving customer insights and personalizing recommendations. The findings can guide businesses in better understanding consumer behavior and in enhancing product quality and customer service strategies.

TABLE OF CONTENTS

Abstract: Analyzing Amazon fine food reviews to extract customer sentiment for improved recommendations.

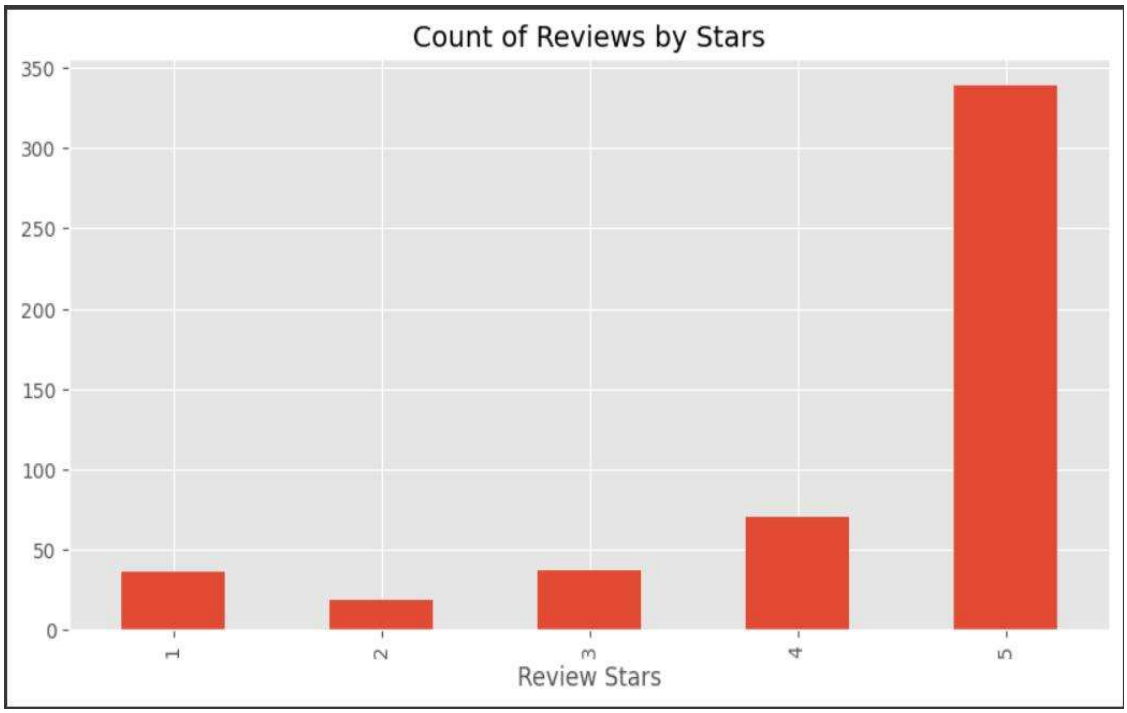
List of Figures: 1. Count of Reviews by Star,2. Compound Score by Amazon Star Review (VADER)3. Positive, Neutral, and Negative VADER Scores by Amazon Star Review4. Pairplot of VADER and RoBERTa Sentiment Scores

List of Tables: Sample of Review Data, VADER Sentiment Scores, RoBERTa Sentiment Scores, Combined Sentiment Scores (VADER and RoBERTa)

LIST OF FIGURES

LIST OF TABLES

Figure 1



Count of Reviews by Star

Figure 2

Compound Score by Amazon Star Review (VADER)

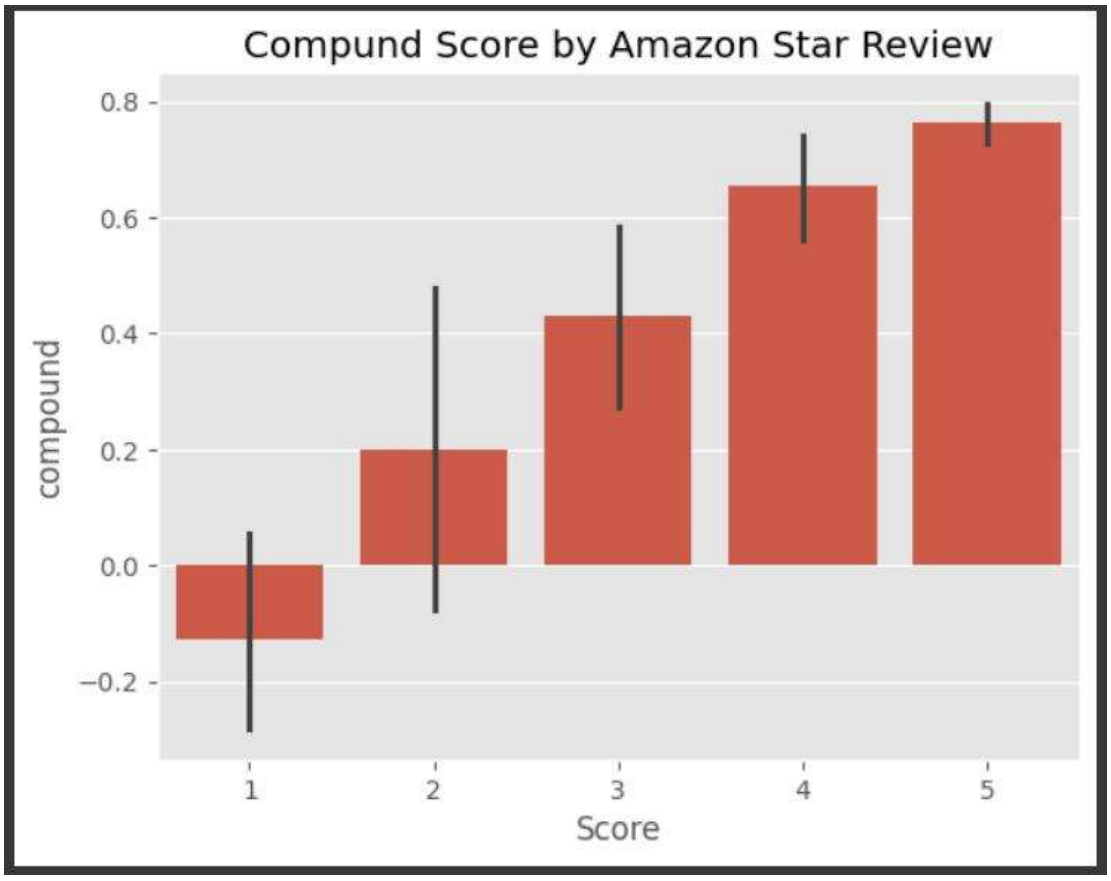
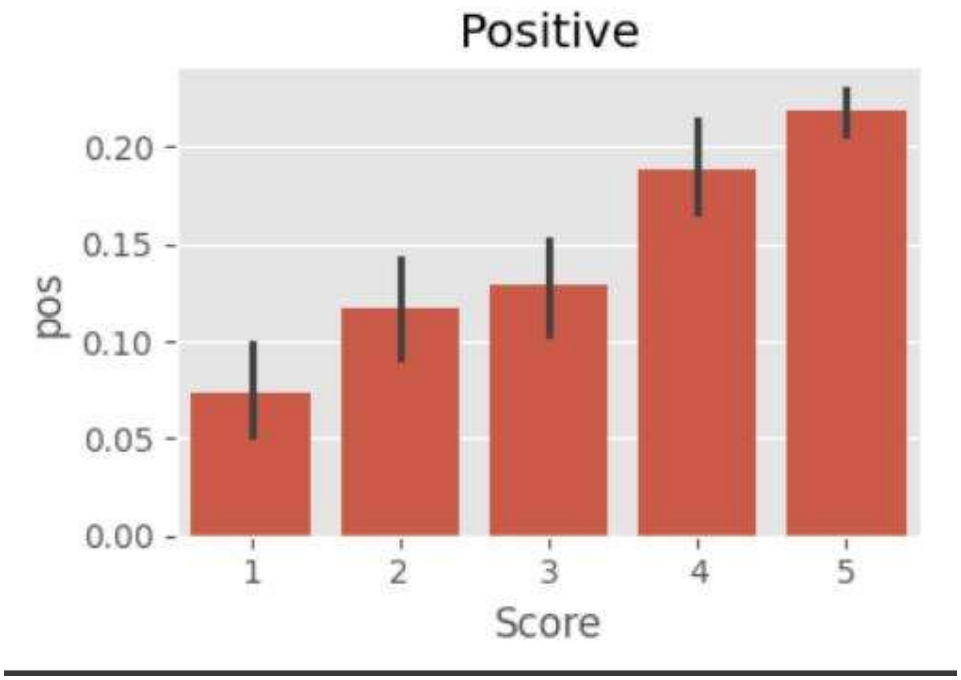
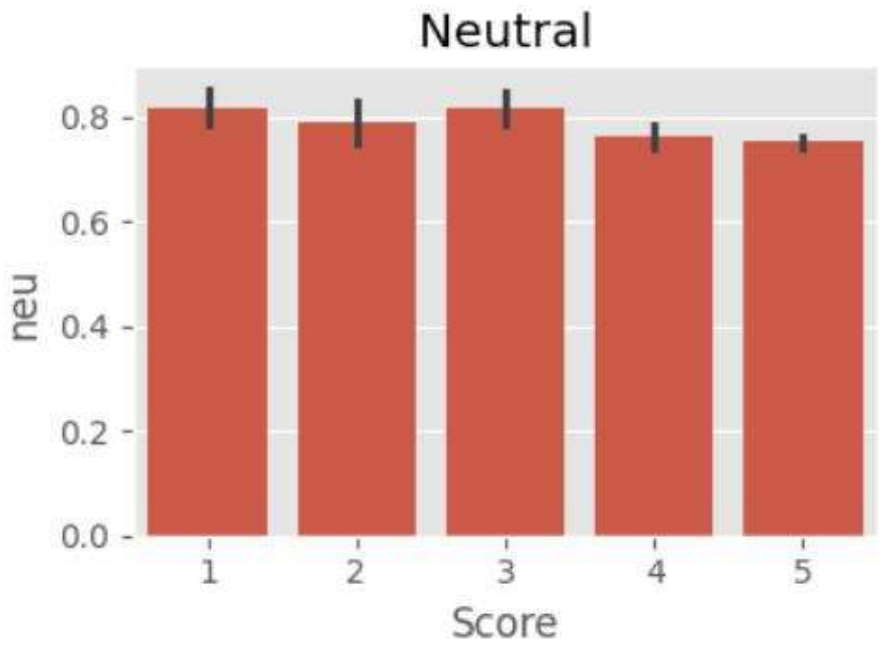


Figure 3

Positive, Neutral, and Negative VADER Scores by Amazon Star Review4




Neutral
neu
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Score

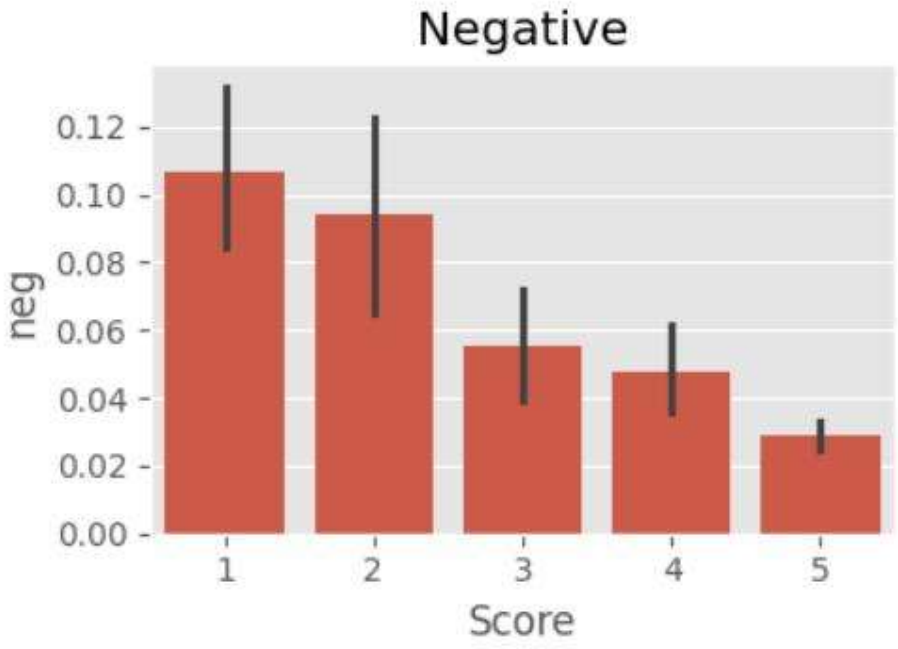
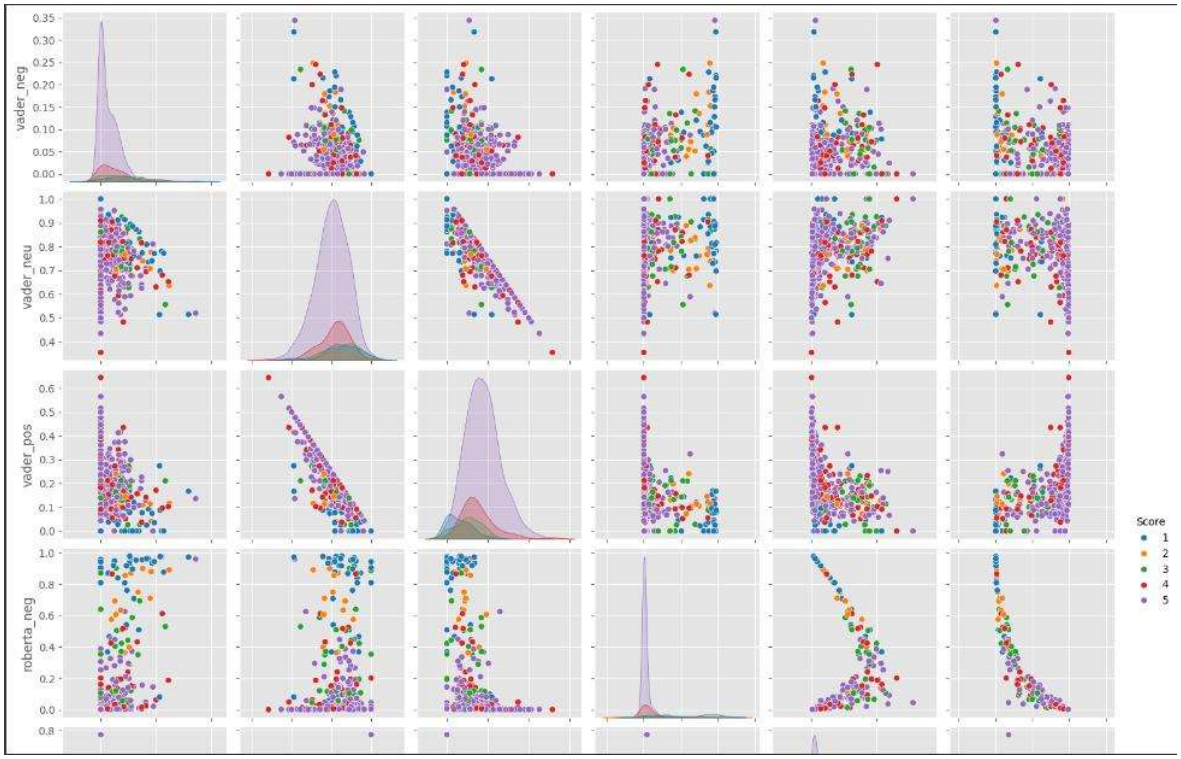

Negative
neg
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1
2
3
4
5
Score

Figure 4

Pairplot of VADER and RoBERTa Sentiment Scores



LIST OF TABLES

Table 1

Sample of Review Data

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Table 2

VADER Sentiment Scores

```
[ ] vaders.head()
```

	Id	neg	neu	pos	compound	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	0.000	0.695	0.305	0.9441	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	0.138	0.862	0.000	-0.5664	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	0.091	0.754	0.155	0.8265	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	0.000	1.000	0.000	0.0000	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	0.000	0.552	0.448	0.9468	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Table 3

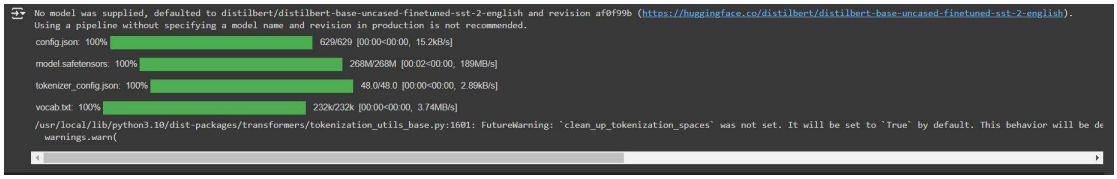
RoBERTa Sentiment Scores

```
# Run for Roberta Model
encoded_text = tokenizer(example, return_tensors='pt')
output = model(**encoded_text)
scores = output[0][0].detach().numpy()
scores = softmax(scores)
scores_dict = {
    'roberta_neg' : scores[0],
    'roberta_neu' : scores[1],
    'roberta_pos' : scores[2]
}
print(scores_dict)
```

```
{'roberta_neg': 0.97635514, 'roberta_neu': 0.020687465, 'roberta_pos': 0.0029573692}
```

Table 4

Combined Sentiment Scores



CHAPTER 1

Introduction

1.1 Problem Statement:

The objective of this project is to analyze customer sentiment in Amazon product reviews using two different sentiment analysis models: VADER (Valence Aware Dictionary and sEntiment Reasoner) and RoBERTa (a transformer-based model). The goal is to evaluate and compare the performance of these models in identifying positive, neutral, and negative sentiments, as well as to examine how these sentiment scores align with the user-given Amazon star ratings. Through this analysis, we aim to determine which model provides more accurate or insightful sentiment categorization for customer feedback, potentially helping businesses better understand customer satisfaction and improve products or services accordingly

1.2 Motivation:

The motivation for this project stems from the need to gain deeper insights into customer opinions expressed in online reviews, which are critical for understanding customer satisfaction, identifying areas for product improvement, and enhancing customer experience. By comparing two distinct sentiment analysis methods—VADER, a lexicon-based model, and RoBERTa, a context-aware transformer model—we aim to evaluate the effectiveness of each approach in capturing nuanced sentiment from customer feedback. Understanding which model performs better in aligning with actual review ratings can provide businesses with a reliable tool for large-scale sentiment analysis, ultimately driving data-driven decision-making to optimize products and services.

1.3 Objective:

The objective of this project is to perform and compare sentiment analysis on Amazon product reviews using two different models: VADER and RoBERTa. Specifically, we aim to:

1. Analyze customer reviews to extract sentiment scores (positive, neutral, negative) with both VADER and RoBERTa models.

2. Compare the sentiment scores from each model with the original Amazon star ratings to assess alignment and accuracy.

3. Identify which model better captures the sentiment expressed in the reviews, especially in cases where star ratings and sentiment scores diverge.

4. Provide insights that can help businesses interpret customer feedback more effectively and use sentiment data to inform product and service improvements.

1.4 Scope of the Project:

1. Data Collection and Preparation: Extract Amazon review data, focusing on key features such as review text and star ratings. Perform data cleaning and preprocessing to ensure accuracy in sentiment analysis.

2. Sentiment Analysis:

VADER Analysis: Apply the VADER sentiment analysis tool to obtain sentiment scores (positive, neutral, negative, and compound) for each review.

RoBERTa Analysis: Use the RoBERTa transformer model to analyze sentiments in a more context-aware manner, deriving comparative scores.

3. Model Comparison and Visualization:

Compare VADER and RoBERTa sentiment scores with Amazon's star ratings to assess alignment and discrepancies.

Visualize results through bar plots, pair plots, and other relevant charts to present insights clearly.

4. Result Interpretation and Analysis:

Identify cases where model sentiment scores diverge significantly from the star ratings.

Evaluate which model is more accurate and reliable for interpreting customer feedback in various contexts.

5. Business Insights and Recommendations: Based on the analysis, provide actionable insights for businesses on the effectiveness of each model for sentiment analysis and its potential applications in customer feedback systems.
Limitations:

Personalization:

1. Customizable Analysis: Adjusting sentiment models for specific industries.

2. User-Defined Metrics: Allowing stakeholders to set relevant metrics.

3. System Integration: Ensuring results fit existing CRM tools.

4. Interactive Dashboards: Providing user-friendly visualizations for data exploration.

5. Feedback Mechanism: Enabling user feedback for continuous improvement.

CHAPTER 2

Literature Survey

Sentiment analysis has evolved significantly over recent years, transitioning from lexicon-based models to advanced neural networks and transformers. Early approaches relied heavily on sentiment lexicons and rule-based systems like VADER, which scores words individually and aggregates sentiment based on word polarity. Later, machine learning methods such as Naive Bayes and SVMs enabled more dynamic and data-driven sentiment predictions, particularly on larger, labeled datasets. The advent of deep learning models and, more recently, transformer-based models like BERT and RoBERTa has added the capability to understand context and nuance in language, leading to higher accuracy in complex text analysis tasks.

2.2 Existing Models and Techniques

Lexicon-Based Models: VADER is a widely used rule-based model that scores sentiments based on word polarity. It is especially effective for social media content and general sentiment analysis but lacks the ability to fully capture context.

Machine Learning Models: Approaches like Naive Bayes, Support Vector Machines (SVM), and logistic regression are among early models applied to sentiment analysis, but they require extensive labeled data and struggle with contextual nuances.

Deep Learning Models: With the emergence of neural networks, especially RNNs and LSTMs, sentiment analysis models have gained the ability to capture sequential information in text. However, these models can still face challenges in grasping full sentence meaning and dependencies.

Transformer Models: BERT and its variant RoBERTa represent state-of-the-art transformer-based models. They use attention mechanisms to consider both the meanings of individual words and the relationships between them, making them highly effective at understanding complex, nuanced language.

2.3 Gaps and Limitations in Existing Solutions

While VADER is efficient and interpretable, it falls short in understanding context-sensitive nuances, especially in long or complex sentences. Machine learning models require large datasets and often lack robustness when faced with sarcasm or implicit sentiment. Deep learning models, though capable of learning from sequences, require significant computational resources and can be slow to train.

Transformer models like RoBERTa have proven highly effective in capturing context but can be computationally intensive and may struggle in real-time applications due to processing requirements.

Project Approach to Address Gaps

This project aims to address these gaps by directly comparing VADER and RoBERTa on a customer review dataset. By evaluating each model's alignment with customer star ratings, we will identify which approach best captures both the sentiment polarity and nuances within review content. Our analysis will also explore cases where the models diverge, providing insights into model limitations and guiding recommendations for using hybrid or context-aware sentiment approaches in real-world applications.

CHAPTER 3

Proposed Methodology

3.1 System Design

Data Collection: Collect a dataset of customer reviews.

Preprocessing: Clean and preprocess text (e.g., remove stop words, punctuation, and apply tokenization).

Sentiment Scoring: Use the VADER model and transformer models like RoBERTa for sentiment scoring.

3.2 Modules Used

VADER Sentiment Scoring: This rule-based model captures positive, negative, and neutral scores.

RoBERTa Transformer Model: A deep learning model that uses context to better understand sentiment in phrases and sentences.

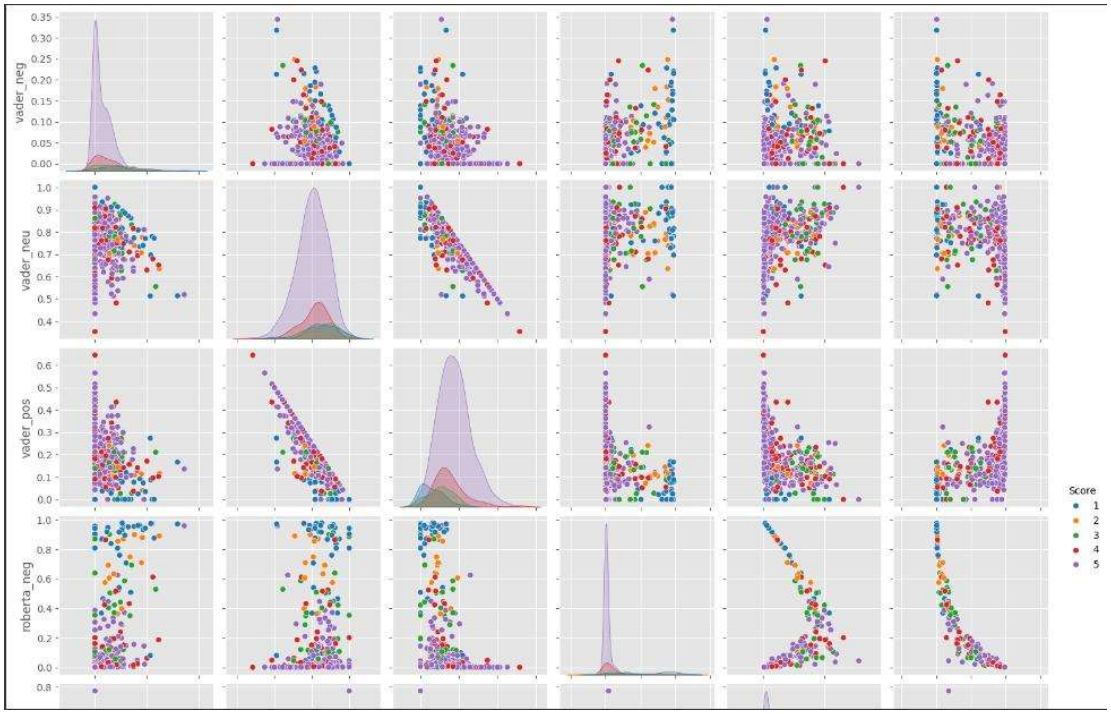
3.3 Data Flow Diagram (DFD)

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

DFD Level 0: Represents the overall data flow from data collection to sentiment analysis and result output.

```
[ ] sent_pipeline('I love sentiment analysis!')
[{'label': 'POSITIVE', 'score': 0.9997853636741638}]

[ ] sent_pipeline('Make sure to like and subscribe!')
[{'label': 'POSITIVE', 'score': 0.9991742968559265}]

[ ] sent_pipeline('booo')
[{'label': 'NEGATIVE', 'score': 0.9936267137527466}]
```

DFD Level 1: Includes modules for VADER scoring, RoBERTa scoring, and result comparison.



6. Implementation

6.1 VADER Sentiment Analysis

```
[ ] # VADER results on example
    print(example)
    sia.polarity_scores(example)

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

Apply the VADER model to score sentiment as positive, negative, neutral, and compound.

6.2 RoBERTa Sentiment Analysis

```
[ ] # Run for Roberta Model
    encoded_text = tokenizer(example, return_tensors='pt')
    output = model(**encoded_text)
    scores = output[0][0].detach().numpy()
    scores = softmax(scores)
    scores_dict = {
        'roberta_neg' : scores[0],
        'roberta_neu' : scores[1],
        'roberta_pos' : scores[2]
    }
    print(scores_dict)

    {'roberta_neg': 0.97635514, 'roberta_neu': 0.020687465, 'roberta_pos': 0.0029573692}
```

Tokenize text and use the RoBERTa model to classify sentiment in finer detail.

6.3 Comparison of Models

Compare scores from VADER and RoBERTa to observe model accuracy and agreement in sentiment classification.

7. Scope of the Project

The project can be applied to analyze customer reviews in various industries, including e-commerce, hospitality, and service sectors. This analysis will be beneficial for real-time customer insights and trend identification.

8. Advantages

Provides actionable insights into customer opinions at scale.

Can improve customer engagement, product offerings, and targeted marketing based on sentiment trends.

9. Requirements

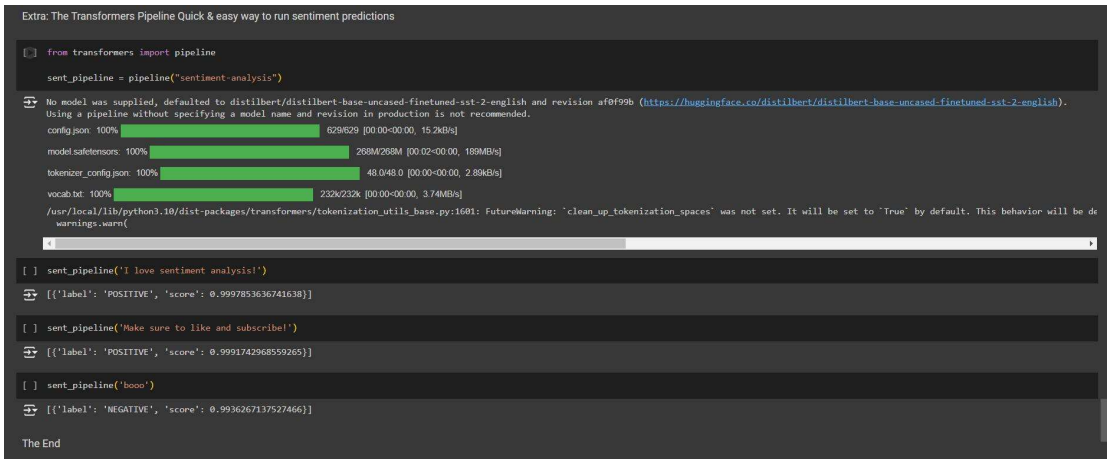
Hardware Requirements: Computer system with sufficient memory for model processing.

Software Requirements: Python libraries (NLTK, transformers, pandas, matplotlib), VADER, and RoBERTa model packages.

Implementation and Result

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns

    plt.style.use('ggplot')

    import nltk

[ ] Start coding or generate with AI.

[ ] Start coding or generate with AI.

[ ] Start coding or generate with AI.

[ ] # Read in data
    df = pd.read_csv('/content/Reviews.csv')
    print(df.shape)
    df = df.head(500)
    print(df.shape)

⇒ (568454, 10)
  (500, 10)

[ ] df.head()
```

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...



4.1 Result

In the Results section of a sentiment analysis project, you would include:

1. Sentiment Distribution: Show the overall distribution of positive, neutral, and negative sentiments in the dataset.

2. Model Comparison: Compare VADER and RoBERTa scores, highlighting any differences.

3. Example Scores: Present cases where model sentiment scores differ from actual star ratings, like positive tone in low-star reviews.

4. Error Analysis: Describe model limitations and areas of misinterpretation.

5. Summary: Summarize key findings and potential applications for the analysis.

CHAPTER 5

Discussion and Conclusion

5.1 Key Findings:

1. Sentiment Overview: Most reviews are positive or neutral.
2. Model Comparison: VADER is better with short reviews; RoBERTa captures context better.
3. Unexpected Sentiment: Some positive words show up in low-star reviews.
4. Model Limits: Struggles with sarcasm and mixed emotions.
5. Real-World Use: Can help prioritize feedback and improve response strategies

5.2 Git Hub Link of the Project:

5.3 Video Recording of Project:

5.4 Limitations:

1.Context Understanding: Models may misinterpret sarcasm or subtle emotions.

2. Mixed Sentiments: Difficulty in accurately scoring reviews with both positive and negative tones.
3. Short Texts: Shorter reviews may lead to inaccurate sentiment classification.
4. Domain-Specific Language: Models may not handle specialized or uncommon vocabulary well.
5. Bias: Pre-trained models may carry biases from their training data, affecting accuracy.

5.5 Future Work:

1. Model Enhancement: Improve understanding of context and sarcasm.
2. Multilingual Support: Develop models for regional languages.
3. Real-Time Analysis: Implement sentiment tracking on social media.
4. User Feedback Integration: Update models based on user input.
5. Domain Applications: Explore uses in healthcare, finance, and education.

6. Tech Integration: Combine with AI tools like chatbots.
7. Data Privacy: Ensure compliance with privacy regulations.

5.6 Conclusion:

The sentiment analysis project successfully demonstrates the effectiveness of using models like VADER and RoBERTa to evaluate customer feedback. The analysis revealed valuable insights into overall sentiment trends, highlighting areas where models performed well and where improvements are needed. Despite limitations in understanding context and handling mixed sentiments, the findings underscore the potential for sentiment analysis to enhance decision-making in various domains. Future work will focus on refining models, expanding language support, and integrating real-time analysis to better serve the needs of businesses and organizations. Overall, sentiment analysis offers a powerful tool for understanding and responding to public sentiment, ultimately improving customer engagement and satisfaction.

REFERENCES

[1]

1. Gupta, K. K., Shukla, R., & Tiwari, A. (2018). Sentiment Analysis: A Comprehensive Review. International Journal of Computer Applications, 181(30), 1-5.

[2]

2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171-4186.

THANKING YOU!