Sentiment Analysis of Amazon Customer Reviews

Kishore Kumar Sunke, **Sevitha Janga, Sai Sree Chitturi**
Department of Computer Science, The University of Texas at Arlington

Abstract

Sentiment analysis is a significant application of machine learning that utilizes NLP (Natural Language Processing) to identify, extract, and analyze emotional states from text corpora. We implemented and compared a rule-based model (VADER) combined with a Random Forest classifier and a transformer-based deep learning model (RoBERTa) for sentiment analysis on a dataset of Amazon customer reviews. This paper presents a detailed comparative analysis of the accuracy and performance of both models, highlighting their strengths and limitations in classifying sentiments as positive, negative, or neutral. Visualizations such as sentiment distribution, and ROC curves are incorporated to provide deeper insights into the model behavior and evaluation.

1 Introduction

This project explores two distinct methodologies for sentiment analysis on Amazon customer reviews, classifying sentiments into positive, neutral, or negative categories. The first approach employs a traditional pipeline that combines VADER, a rule-based sentiment analysis tool, with a Random Forest classifier trained on TF-IDF features extracted from the preprocessed text. This method emphasizes simplicity and interpretability. The second approach integrates Hugging Face's transformer-based RoBERTa model, fine-tuned for sentiment analysis, with TF-IDF vectorization and Random Forest classification for a hybrid solution. Both methods undergo rigorous evaluation using metrics such as classification reports, confusion matrices, and ROC curves, alongside visualization techniques to gain deeper insights into their behavior. By comparing these approaches, the project seeks to assess their effectiveness, analyze their strengths and limitations, and provide a comprehensive understanding of real-world sentiment analysis challenges.

2 Dataset Description

The dataset consists of approximately 500,000 fine food reviews from Amazon, spanning over 10 years up to October 2012. Each review includes detailed information such as a unique product identifier (Product Id), a unique user identifier (User Id), the profile name of the user, the number of users who found the review helpful (Helpfulness Numerator), the total number of users who voted on helpfulness (Helpfulness Denominator), the product rating on a scale of 1 to 5 (Score), the timestamp of when the review was posted (Time), a brief user-provided summary of the product (Summary), and the full review text (Text). The dataset covers various Amazon categories, providing diverse and representative customer feedback. For analysis, the numerical ratings were mapped into three sentiment classes: negative (1–2), neutral (3), and positive (4–5). To prepare the data for modeling, preprocessing steps included removing HTML tags, punctuation, and special characters, tokenizing text into words, removing stop words to retain meaningful content, and applying lemmatization to normalize words.

3 Project Description

The project focuses on analyzing Amazon customer reviews to classify sentiments into three categories: Positive, Neutral, and Negative. Sentiment analysis, a critical task in Natural Language Processing (NLP), is widely used in customer feedback analysis, product recommendation systems, and business intelligence. This project integrates a rule-based sentiment analysis tool, VADER, with a Random Forest classifier and a transformer-based RoBERTa model from Hugging Face, paired with TF-IDF vectorization, to evaluate different methodologies for sentiment classification. The workflow begins with data loading and preprocessing, which includes cleaning HTML tags, converting text to lowercase, removing punctuation and stop words, tokenization, and lemmatization to standardize text data. Advanced evaluation techniques such as classification reports, confusion matrices, ROC curves are used to assess model performance and provide visual insights. By comparing the strengths and limitations of these approaches, the project identifies the most effective method for sentiment classification, highlighting the balance between the simplicity of rule-based methods and the depth of transformer-based models. This study serves as a detailed exploration of hybrid and traditional sentiment analysis systems, showcasing their applications in real-world text analysis tasks.

4 Data Loading and Sentiment Mapping

The dataset contains Amazon customer reviews, including text descriptions and ratings from 1 to 5. Using Pandas, the data was loaded into a structured format, ensuring efficient manipulation and analysis. Ratings were mapped to three sentiment categories: Negative, Neutral, and Positive, to facilitate sentiment classification. Numerical scores were transformed into sentiment labels, where scores of 1–2 indicated Negative sentiment, 3 represented Neutral sentiment, and 4–5 were classified as Positive sentiment. This mapping process streamlined the analysis by converting numerical ratings into categorical labels, enabling a more interpretable approach to understanding customer feedback and supporting subsequent sentiment analysis tasks effectively.

5 Visualization

Sentiment Distribution: The bar chart illustrates the distribution of sentiment classes (Positive, Negative, and Neutral) within the dataset. The Positive sentiment category dominates the dataset with the highest frequency, indicating a significant imbalance compared to the Negative and Neutral sentiments. Negative sentiments have a moderate representation, while Neutral sentiments are the least frequent. This imbalance suggests the need for balancing techniques, such as oversampling or under sampling, to ensure equitable training and prevent the model from being biased toward the Positive class. The chart highlights the importance of addressing this imbalance for a more robust and generalizable sentiment analysis model.

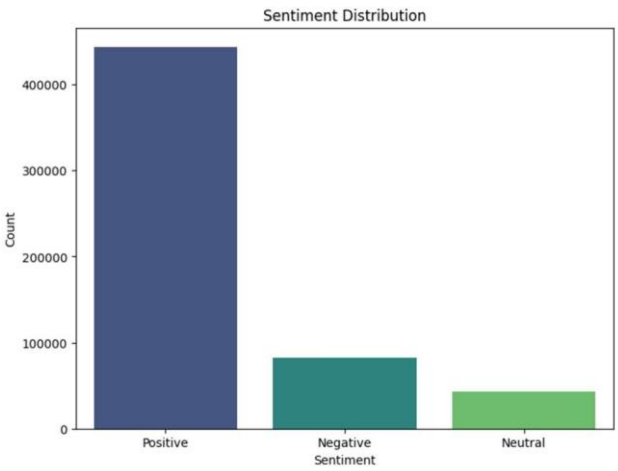


Diagram 1: Sentiment Distribution of Initial Data

6 Balancing the Sentiment Dataset

Balancing data is essential in sentiment analysis to prevent bias towards majority sentiment classes, which can negatively impact model performance, especially on underrepresented classes. To achieve balance, the dataset is first split into subsets based on sentiment labels: Negative, Neutral, and Positive. The smallest class size is then identified to define the sample size for each class. Equal-sized samples are randomly selected from each class, with replacement if necessary to handle any imbalances. Afterward, these subsets are concatenated, shuffled, and the index is reset. This process ensures that each sentiment class is fairly representedin the dataset, enabling a more accurate and reliable analysis or modeling.

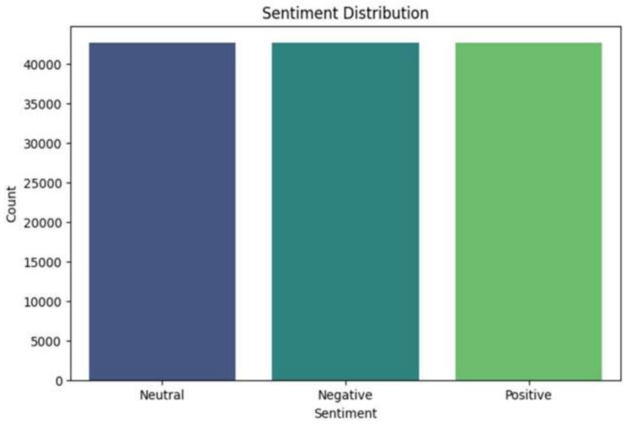


Diagram 2: Sentiment Distribution after Balancing Data

7 Data Preprocessing

Data preprocessing is a crucial step in the analysis pipeline, especially when working with text data for tasks such as sentiment analysis. The goal of preprocessing is to clean, normalize, and prepare the data in a format that can be effectively processed by machine learning models. In this case, the dataset consists of Amazon customer reviews, which include textual descriptions and ratings. The following steps outline the data preprocessing process used to prepare the dataset for sentiment analysis.

1. Text Cleaning: The first step in preprocessing is to clean the text data to ensure uniformity and remove irrelevant or noisy information. This includes the removal of HTML tags, unnecessary whitespaces, and special characters. HTML tags often appear in raw text extracted from web pages, and they can interfere with text analysis. Special characters, such as punctuation marks or non-alphabetic symbols, are also removed as they don’t contribute to the sentiment of the text. By eliminating these elements, the text becomes cleaner and easier to analyze.

2. Lowercasing: To further standardize the text and avoid duplication due to case differences, all text is converted to lowercase. For example, "Good product" and "good product" would be treated as the same. This step ensures that case variations don't create inconsistencies and that words are uniformly represented in the analysis. Converting all text to lowercase helps reduce the complexity of the dataset and is a common practice in text processing.

3. Removing Punctuation and Stop Words: Next, punctuation marks such as commas, periods, and question marks are removed. While punctuation can carry meaning in some contexts (e.g., question marks indicating questions), they are typically not helpful in sentiment analysis for general product reviews. Similarly, stop words—common words like "the," "is," "and," and "in"—are also removed. These words are considered noise because they occur frequently across text and do not carry significant meaning related to sentiment. By removing stop words, the focus is placed on the more meaningful words that influence the sentiment, such as adjectives or nouns.

4. Tokenization: Tokenization is the process of splitting the cleaned text into smaller units, or tokens, typically individual words. This step is important because it breaks the text into manageable pieces for further analysis. For instance, the sentence "This product is amazing" would be tokenized into ["This", "product", "is", "amazing"]. By separating the text into tokens, we make it easier for machine learning models to process the data, understand the structure, and identify patterns.

5. Lemmatization: Lemmatization is the process of reducing words to their base or root form. For example, the words "running," "ran," and "runner" would all be reduced to the lemma "run." This step ensures that different forms of a word are treated as the same, allowing the model to focus on the core meaning rather than variations of the word. Lemmatization is preferred over stemming, as it produces more meaningful and accurate base forms of words. For example, "better" would be reduced to "good" instead of just removing the "er" suffix.

8 VADER Sentiment Analysis Implementation

VADER: The VADER (Valence Aware Dictionary and sEntiment Reasoner) model was applied to analyze sentiment in Amazon customer reviews, leveraging its rule-based approach to classify reviews into three sentiment categories: Positive, Negative, and Neutral. VADER’s design makes it especially suited for analyzing informal texts such as customer reviews, where it can capture nuances like emoticons, slang, and varying sentence structures.

Sentiment Classification: After preprocessing the review texts, VADER was applied to predict the sentiment of each review. The model assigned a compound sentiment score to each review, which was then used to classify the sentiment. If the compound score was greater than 0.05, the review was labeled as Positive; if the score was less than -0.05, the review was labeled as Negative; and if it fell between -0.05 and 0.05, the sentiment was classified as Neutral.

Hybrid Sentiment Analysis: To enhance the sentiment analysis, a hybrid model was developed by combining the sentiment derived from product ratings with VADER's predictions. The product ratings, which ranged from 1 to 5, were mapped to sentiment categories: Negative for ratings 1-2, Neutral for 3, and Positive for 4-5. If the sentiment from VADER did not align with the rating-based sentiment, the VADER sentiment was prioritized for a more refined classification.

Model Training and Evaluation: The reviews were vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) to transform the text into numerical features suitable for machine learning. A Random Forest classifier was used to train the model on 80% of the data, with the remaining 20% used for testing. This allowed for a robust evaluation of the model's performance.

Performance Metrics: Several evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix, were used to assess the performance of the VADER-based sentiment analysis. These metrics provided valuable insights into how effectively the model classified sentiments and managed imbalances in sentiment distribution, ensuring reliable sentiment classification for Amazon reviews.

9 RoBERTa Implementation

RoBERTa: The RoBERTa model was applied to analyze sentiment in Amazon customer reviews, leveraging its deep learning capabilities to classify reviews into three sentiment categories: Positive, Negative, and Neutral. Using a pre-trained RoBERTa model via the Hugging Face pipeline simplified the process, avoiding the need for training from scratch and providing a powerful sentiment analysis tool.

Sentiment Classification: After preprocessing the review texts, the RoBERTa model predicted the sentiment for each review. The model returned labels like 'LABEL_0', 'LABEL_1', and 'LABEL_2', which were mapped to Negative, Neutral, and Positive sentiments. The label with the highest probability score was chosen as the final sentiment classification.

Hybrid Sentiment Analysis: To improve accuracy, a hybrid model was developed by combining the sentiment derived from product ratings with RoBERTa's predictions. When there was a mismatch between the review's rating and RoBERTa's predicted sentiment, the model prioritized RoBERTa's sentiment to ensure a more nuanced understanding of customer feedback. This helped resolve discrepancies where the review text's sentiment differed fromthe numeric rating.

Model Training and Evaluation: The reviews were vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) to transform text into numerical features suitable for machine learning. A Random Forest classifier, an ensemble learning method, was used to train themodel on 80% of the data, with 20% reserved for testing.

This allowed the model to generalize well and avoid overfitting.

Performance Metrics: Several evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix, were used to assess the model's performance. These metrics provided insights into how effectively the model classified sentiments and managed imbalances in sentiment distribution, ensuring an accurate and robust sentiment analysis of customer reviews.

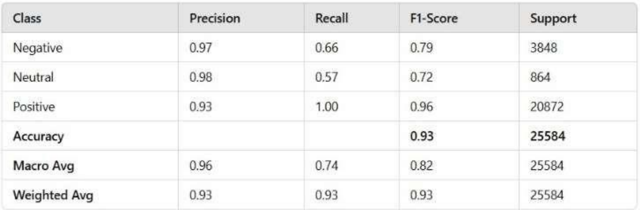
10 Testing

K-Fold cross-validation to assess the performance of a machine learning model. It employs stratified sampling, ensuring the distribution of target classes is consistent across all data splits. The process involves dividing the data into 5 subsets, where the model is trained on 4 subsets and evaluated on the remaining subset in each iteration. Accuracy is computed for each fold to determine how well the model predicts unseen data. Additionally, confusion matrices are generated to provide detailed insights into classification results for each fold, highlighting any misclassifications. After completing all folds, the mean accuracy is calculated, offering a robust measure of overall model performance. This method systematically evaluates the model on the entire dataset, reducing biases and ensuring that no single partition of data disproportionately influences the evaluation results, making it an essential step in machine learning workflows.

11 Results for VADER

Classification report: The classification report (Table 1) summarizes the model's performance across the three sentiment classes: Negative, Neutral, and Positive. The model achieved an **overall accuracy of 93%,** with a macro-average F1-score of 0.82 and a weighted average F1-score of 0.93.

Class-wise performance shows high precision for all classes, with the Positive class achieving the highest recall (1.00) and F1-score (0.96). However, the Neutral class has a lower recall (0.57), indicating room for improvement in identifying neutral sentiments. The model performs best in identifying Positive sentiments, followed by Negative, reflecting its strong overall consistency in sentiment classification.

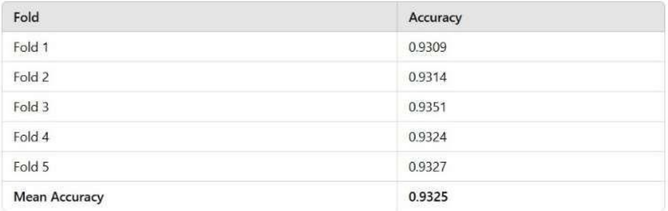


Class	Precision	Recall	F1-Score	Support
Negative	0.97	0.66	0.79	3848
Neutral	0.98	0.57	0.72	864
Positive	0.93	1.00	0.96	20872
Accuracy			**0.93**	**25584**
Macro Avg	0.96	0.74	0.82	25584
Weighted Avg	0.93	0.93	0.93	25584

Table 1: Classification report for VADER

Cross-validation: The cross-validation process is a reliable method to evaluate a model's performance while minimizing bias and overfitting. In this case, a five-fold cross-validation approach was used, dividing the dataset into five equal parts. For each fold, the model was trained on four partitions and tested on the fifth, ensuring that every subset of the data was used for both training and testing. This method provides a balanced assessment by exposing the model to all available data in both roles.

The results (Table 2) for each fold were reported, with fold accuracies ranging from 0.9309 to 0.9351. This high level of consistency suggests that the model is not overly sensitive to the specific training or testing subsets, demonstrating its robustness and stability. The mean accuracy across all folds was 0.9325, reflecting the model's ability to generalize well to unseen data.



Fold	Accuracy
Fold 1	0.9309
Fold 2	0.9314
Fold 3	0.9351
Fold 4	0.9324
Fold 5	0.9327
Mean Accuracy	**0.9325**

Table 2: 5- Fold Cross Validation Accuracies

Confusion Matrix.: Confusion matrix shows the performance of the model across three classes: Negative, Neutral, and Positive. The model performs well in predicting Positive sentiment, with 20,803 correct predictions. Negative sentiment is misclassified as Neutral (1301), while Neutral sentiment is misclassified as Negative (2540) and Positive (358). The overall accuracy is high, indicating effective sentiment classification, especially for the Positive class.

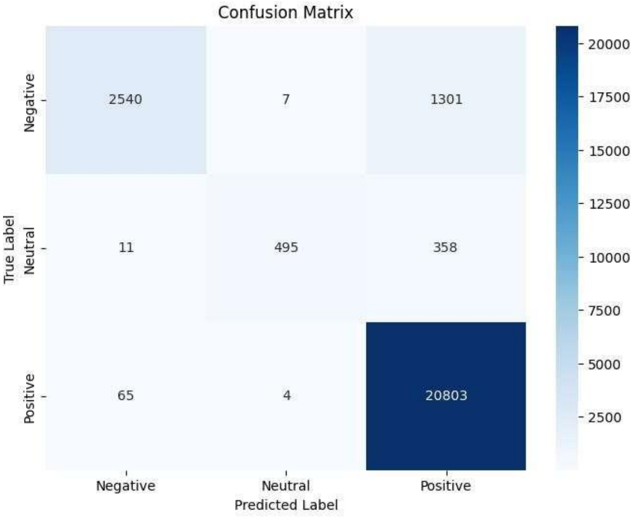


Diagram 3: Confusion Matrix for VADER

ROC Curve: The ROC curve evaluates the model's ability to distinguish between Positive, Neutral, and Negative sentiments. High AUC values (0.95–0.97) indicate strong performance for all sentiment classes.

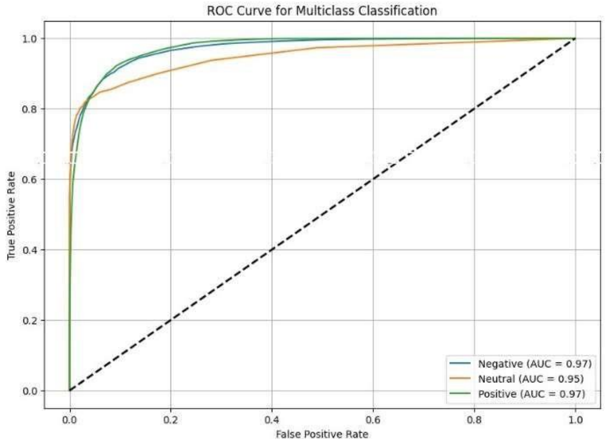
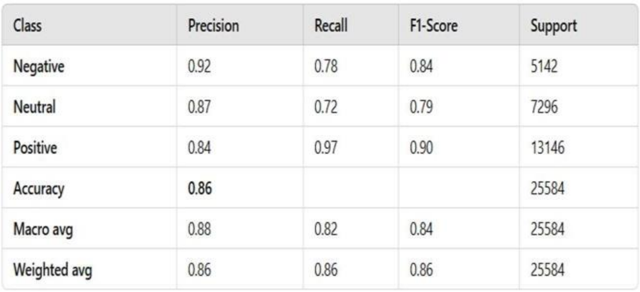


Diagram 4: ROC Curve for VADER

12 Results for RoBERTa

Classification report: The classification report (Table 3) summarizes the model's performance across the three sentiment classes: Negative, Neutral, and Positive. The model achieved an **overall accuracy of 86%,** with a macro-average F1-score of 0.84 and a weighted average F1-score of 0.86.

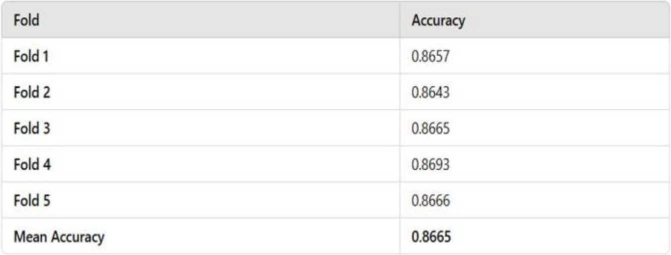
Class-wise performance shows the highest precision for Negative sentiments (0.92) and the highest recall for Positive sentiments (0.97), resulting in an F1-score of 0.90 for Positive. However, the Neutral class has slightly lower recall (0.72) and F1-score (0.79), indicating room for improvement in identifying Neutral sentiments. Overall, the model performs best in identifying Positive and Negative sentiments, reflecting its reliability in sentiment classification tasks.



Class	Precision	Recall	F1-Score	Support
Negative	0.92	0.78	0.84	5142
Neutral	0.87	0.72	0.79	7296
Positive	0.84	0.97	0.90	13146
Accuracy	**0.86**			25584
Macro avg	0.88	0.82	0.84	25584
Weighted avg	0.86	0.86	0.86	25584

Table 3: Classification report for RoBERTa

Cross-validation: The cross-validation results (Table 4) for RoBERTa indicate consistent performance across five folds. The accuracy for each fold varied slightly, ranging from 0.8643 to 0.8693. The model achieved an average accuracy of0.8665, demonstrating stable performance with minimal fluctuation across the folds. This suggests that RoBERTa is effective in classifying sentiment in the dataset, with each foldcontributing similarly to the overall performance. The cross- validation process validates the model's robustness in generalizing to different subsets of the data, ensuring that the performance is not dependent on any single partition.



Fold	Accuracy
Fold 1	0.8657
Fold 2	0.8643
Fold 3	0.8665
Fold 4	0.8693
Fold 5	0.8666
Mean Accuracy	**0.8665**

Table 4: 5- Fold Cross Validation Accuracies

Confusion Matrix: The confusion matrix indicates that the model accurately predicts Positive sentiment (12753 correct predictions) but misclassifies some samples as Neutral or Negative. Neutral and Negative classes have more misclassifications, with Neutral misclassified as Negative (257) and Positive (1766), while Negative is misclassified as Neutral (484) and Positive (666).

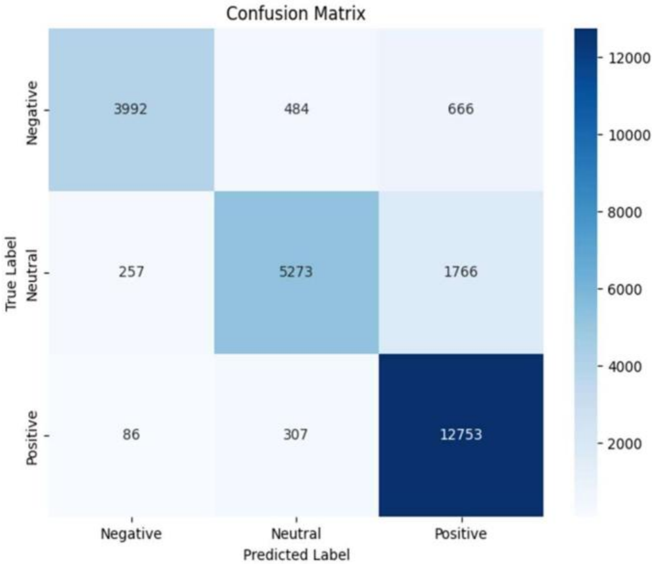


Diagram 5: Confusion Matrix for RoBERTa

ROC Curve: The ROC curve evaluates the model's ability to distinguish between Positive, Neutral, and Negative sentiments. High AUC values (0.95–0.98) indicate strong performance for all sentiment classes.

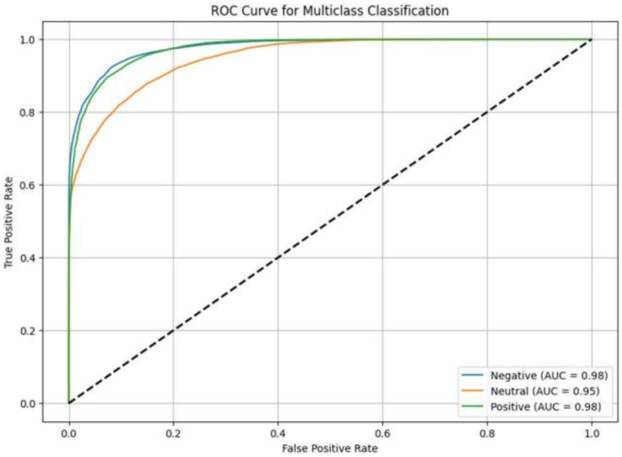


Diagram 6: ROC Curve for VADER

13 Comparison:

In comparing VADER and RoBERTa for sentiment analysis of Amazon reviews, both models exhibit strengths and weaknesses based on their approach and performance metrics.

VADER, a lexicon-based sentiment analysis tool, achieved a higher accuracy (93.25%) in classifying sentiments in the dataset. Its precision and recall for the Positive class were particularly high, making it effective for detecting positive sentiments. The model is fast and efficient, ideal for processing large datasets quickly. However, it faced challenges in distinguishing between Negative and Neutral sentiments, with some misclassifications observed in the confusion matrix.

RoBERTa, a transformer-based model, demonstrated slightly lower accuracy (86.65%) but showed better recall for the Positive class, indicating a strong ability to detect positive sentiments. It outperformed VADER in handling complex sentiment patterns, though it was slower and computationally more expensive. The confusion matrix for RoBERTa highlighted its struggle in classifying Negative and Neutral sentiments, similar to VADER, but with fewer misclassifications.

Conclusion: VADER excels in speed and accuracy for large-scale datasets and simpler sentiment detection tasks, while RoBERTa, despite being slower, offers deeper contextual understanding and performs well in complex sentiment analysis scenarios. VADER is better suited for real-time applications, while RoBERTa is preferred for tasks requiring nuanced sentiment detection.

14 Main References Used

Sentiment Analysis Based on RoBERTa for Amazon Review: An Empirical Study on Decision Making

This study leverages the RoBERTa transformer-based NLP model to analyze Amazon product reviews for sentiment, demonstrating its effectiveness in generating accurate sentiment scores. It explores patterns related to behavioral economics concepts like electronic word-of-mouth (eWOM), consumer emotions, and confirmation bias. The findings emphasize how NLP models can provide actionable insights, supporting strategic marketing and decision-making. By integrating these insights with behavioral economics, businesses gain a deeper understanding of consumer behavior, particularly in digital environments. As NLP technologies evolve, they offer growing opportunities for research and innovation, enhancing business practices and consumer insights.

Opinion Mining and Sentiment Analysis: A Survey

The paper discusses multiple machine learning approaches for aspect-based sentiment analysis, including SVM classifiers, lexicalized Hidden Markov Models (LHMMs), Conditional Random Fields (CRFs), and pattern-based techniques. These methods focus on extracting opinion expressions, associating them with target aspects and determining sentiment orientation. Key challenges include handling opinion shifters, intensifiers, and sentence structures like "but" clauses. Evaluation metrics such as accuracy, precision, recall, and F1-score are used, with room for future improvement in techniques

15 Differences in Approach/Method

Hybrid Sentiment Analysis Model: Sentiment analysis was performed using two separate models: RoBERTa and VADER**.** RoBERTa, a transformer-based model, was fine-tuned on Amazon product reviews to capture deep contextual relationships in the text, while VADER, a lexicon-based approach, handled simpler, rule-based sentiment expressions. Integrated Random Forest, a classical machine learning model, to refine predictions. This model was trained on sentiment features extracted by VADER and RoBERTa, offering a hybrid machine learning pipeline.

Class Imbalance Handling: By addressing class imbalance, it ensured equal representation across sentiment classes. This improves model fairness and performance, especially in predicting Neutral and Negative sentiments, unlike other models that ignore or inadequately handle class imbalance, leading to skewed results favoring the Positive class.

Advanced Text Preprocessing: In my project, stopword removal, lemmatization, and tokenization clean the text, making the data more efficient for sentiment classification. Study, used simpler preprocessing techniques that may not fully clean the text, affecting model accuracy by retaining unnecessary noise in the data.

Feature Selection with TF-IDF: Used TF-IDF vectorization, selecting the top 5000 features. This ensures that the model focuses on the most relevant terms, improving performance. In study used simpler vectorization methods, such as Bag-of-Words, which fail to prioritize important features, impacting overall accuracy.

16 Difference in Accuracy/Performance

In this project, VADER achieved 93% accuracy and RoBERTa reached 86% accuracy while analyzing 568,454 Amazon product reviews**.** This large dataset allowed both models to showcase their robustness. VADER excelled in precision, indicating its strong performance in correctly identifying positive and negative sentiments. On the other hand, RoBERTa demonstrated superior recall, making it better at identifying all possible positive, negative, and neutral sentiment instances. Additionally, metrics such as F1-score were used, with VADER performing exceptionally well in balancing precision and recall.

In contrast, the reference paper using RoBERTa on a smaller Amazon dataset achieved an accuracy of 85%**,** showing competitive performance with a pre-trained model. However, the smaller dataset likely limited the model's ability to generalize effectively, resulting in a slightly lower performance compared to this project. Unlike the reference study, this project used hybrid techniques**,** preprocessing steps like tokenization, stopword removal, and lemmatization**,** and handled class imbalance more effectively,contributing to the higher accuracy observed in both models.

17 What did we do well?

Comprehensive Pipeline: The project successfully integrated all essential stages of a machine learning pipeline, from preprocessing to modeling and evaluation. By following a structured approach, the process ensured that data was prepared correctly, models were trained effectively, and the results were thoroughly evaluated. This holistic pipeline ensured smooth execution and accurate results throughout.

Hybrid Sentiment Analysis: Using both VADER and RoBERTa models for sentiment classification was an excellent strategy. VADER, with its lexicon-based approach, was useful for simple polarity detection, while RoBERTa provided a more sophisticated, context-aware analysis of sentiment. This hybrid approach allowed for a comprehensive evaluation, comparing the strengths of both models and demonstrating adaptability to varying types of sentiment in the dataset.

Data Balancing: Addressing the class imbalance in sentiment analysis was crucial for ensuring fair model evaluation. Ensuring equal representation across the differentsentiment classes (Negative, Neutral, Positive) helped the models learn effectively, preventing bias towards the majority class. This balanced approach improved the models' ability to classify minority sentiment categories with greater accuracy.

Robust Evaluation: The use of cross-validation was a significant strength in evaluating model performance. By dividing the data into multiple folds and testing on each fold, the models' generalizability was thoroughly assessed, providing a reliable estimate of performance. Additionally, visualizing results with confusion matrices and ROC curves helped in understanding where each model succeeded or struggled, further informing decision-making for model optimization.

Efficient Feature Extraction: The implementation of TF-IDF (Term Frequency-Inverse Document Frequency) with a capped feature set was an effective strategy to manage large amounts of textual data. This approach optimized memory usage by retaining only the most important features, ensuring that the models had access to relevant information without being overwhelmed by irrelevant or redundant features. The decision to cap the features ensured computational efficiency while maintaining high model performance.

18 What could we have done better?

Model Ensemble: Although using both VADER and RoBERTa individually is beneficial, combining their predictions through an ensemble method (such as voting or averaging) could potentially improve classification performance. By leveraging the strengths of both models, an ensemble approach might enhance robustness, particularly for mixed or ambiguous sentiment in reviews.

Advanced Data Augmentation: While data balancing was addressed, further techniques such as data augmentation could help boost the model's performance. For example, generating synthetic data using techniques like SMOTE (Synthetic Minority Over-sampling Technique) for the minority sentiment classes could help improve the model's ability to predict less frequent sentiments without losing valuable information from the original dataset.

Handling Long Reviews: The models might struggle with very long reviews due to token length limits (e.g., RoBERTa has a maximum token length). Breaking down long reviews into smaller chunks and aggregating the results could improve classification performance. Implementing a sliding window approach for processing long text might also mitigate this issue.

Scalability: While the models performed well, future work could focus on optimizing the pipeline for scalability. For example, implementing batch processing and distributed computing frameworks (e.g., using Apache Spark) could help handle even larger datasets more efficiently, reducing training time and improving overall system performance.

19 Future Scope

Multilingual Sentiment Analysis: Expanding the project to support multiple languages would increase its applicability. By integrating models like mBERT or XLM-R, sentiment analysis can be extended to diverse languages, improving global market insights.

Real-Time Sentiment Analysis: Implementing real-time sentiment analysis for platforms like social media or live customer reviews would help businesses react promptly to customer feedback and market trends. Tools like Twitter API or Facebook Graph API could be used to gather real-time data for immediate analysis.

Aspect-Based Sentiment Analysis (ABSA): Incorporating ABSA would provide more granular insights by analyzing sentiment towards specific product features, such as shipping time or product quality, allowing for detailed customer feedback analysis.

20 Conclusion

The sentiment analysis project has demonstrated a comprehensive and effective approach to classifying customer reviews, with a clear focus on preprocessing, data balancing, modeling, and evaluation. By employing both **VADER** and **RoBERTa**, the project successfully leveraged traditional lexicon-based methods and advanced transformer models, offering a hybrid solution that captures the nuances of sentiment in text data. The use of **TF-IDF** for feature extraction and **Random Forest** for classification further optimized the model's performance.

Cross-validation and the evaluation metrics, such as accuracy, precision, recall, and F1-score, provided a robust assessment of model effectiveness. The project also tackled class imbalance, ensuring fair representation across sentiment classes, which enhanced the model's reliability.

However, there are several opportunities for future enhancement. Expanding the project to include multilingual sentiment analysis using models like **mBERT** or **XLM-R** could broaden its application to global markets. Real-time sentiment analysis could help businesses make quick decisionsbased on live feedback. Moreover, using ensemble methods, aspect-based sentiment analysis, and explainable AI techniques would enhance the model's accuracy, granularity, and interpretability.

Overall, this project has laid a solid foundation for sentiment analysis, with room for growth in terms of scalability, accuracy, and real-time application.

21 Acknowledgements

Our research work is supported by adjunct professor Jesus Gonzalez Bernal, University of Texas at Arlington. We would also like to thank our teaching assistant Manish Munikar for their support.

22 References

[1] Shadmaan Hussain, Namrata Dhanda, Rajat Verma, “Sentiment Analysis of Amazon Product Reviews using VADER and RoBERTa Models”, 2023.

[2] Prachi Juyal, "Sentimental analysis of amazon customers based on their review comments", 2022.

[3] Aamir Rashid, Ching-Yu Huang, "Sentiment Analysis on Consumer Reviews of Amazon Products", 2021.

[4] Shikha Maurya, Vibha Pratap, "Sentiment Analysison Amazon Product Reviews", 2022.

[5] Jiaqi Li, Qi Pan, Yihao Wang, "Sentiment analysis applied on Amazon reviews", 2024.

[6] Akanksha Halde, Aditi Uttekar, Amit Vishwakarma, "SENTIMENT ANALYSIS ON AMAZON PRODUCT REVIEWS", 2022.

[7] Sharmin Shaikh, Aniket Navale, Prithvi Sunchu, Deepali Shrikhande, "Sentiment Analysis of Amazon Reviews", 2020.

[8] Linda Erwe, Xin Wang, "Comparison of VADERand Pre-Trained RoBERTa", 2024.

[9] https://medium.com/@onubachibuike/vader-vs-roberta-a-comparison-of-sentiment-analysis-models-72f8ceb1934b

[10] Amal Krishna, Balendu M, Rahul Varma U, "Comparative Analysis of Sentiment Analysis Models: RoBERTa vs. VADER", 2024.

[11] Mohammad sadegh Hajmohammadi, Roliana Ibrahim, Zulaiha Ali Othman, Opinion Mining and Sentiment Analysis: A Survey, 2012.

[12] Xinli Guo, Sentimental analysis based on RoBERTa for Amazon review: an empirical study on decision making, 2024.