Sentiment Analysis of Amazon Fine Food Reviews Using VADER and Hugging Face Model RoBERTa

Anand Khanna, Sai Varun, Alexis Myers, Hetvi Shah
akhann40@asu.edu,skandur7@asu.edu, anmyers3@asu.edu, hshah119@asu.edu

School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ, USA

ABSTRACT

This research explores sentiment classification in e-commerce using three approaches: rule-based VADER, hybrid VADER-XGBoost, and fine-tuned RoBERTa. On the Amazon Fine Food Reviews dataset (~568K entries), sentiment labels were mapped into positive, neutral, and negative classes. RoBERTa achieved the highest accuracy (92.92%), followed by Boosted VADER (76.46%) and Normal VADER (76.06%). Preprocessing included word frequency analysis and review normalization. An interactive Streamlit dashboard was created for real-time sentiment analysis and model comparison, offering business-ready insights and deployment potential.

NOMENCLATURE

Term	Description
[2] RoBERTa	Robustly Optimized BERT Pretraining Approach
[3] VADER	Valence Aware Dictionary and sEntiment Reasoner
[4] NLP	Natural Language Processing
[5] XGBoost	Extreme Gradient Boosting
UI	User Interface
GPU	Graphics Processing Unit
[6] API	Application Programming Interface
[7] CSV	Comma-Separated Values file format
[8] EDA	Exploratory Data Analysis
[9] Tokenization	Splitting text into smaller units (tokens) for model processing
[10] Epoch	One complete pass through the entire training dataset
[15] Accuracy	Percentage of correctly predicted labels out of total predictions
[11] Hugging Face	Open-source platform for NLP models and deployment tools

1. INTRODUCTION

This research explores sentiment classification in e-commerce using three approaches: rule-based VADER, hybrid VADER-XGBoost, and fine-tuned RoBERTa. On the Amazon Fine Food Reviews dataset (~568K entries), sentiment labels were mapped into positive, neutral, and negative classes. RoBERTa achieved the highest accuracy (92.92%), followed by Boosted VADER (76.46%) and Normal VADER (76.06%). Preprocessing included word frequency analysis and review normalization. An interactive Streamlit dashboard was created for real-time sentiment analysis and model comparison, offering

business-ready insights and deployment potential.

2. MATERIALS AND METHODS

2.1 Dataset Description

The [12] Amazon Fine Food Reviews dataset contains 568,454 reviews. Sentiment scores were grouped into three categories: negative (ratings 1–2), neutral (rating 3), and positive (ratings 4–5). Class imbalance was managed through stratified sampling.

Figure 1: Dataset example

2.2 Preprocessing

[13] Review texts were cleaned and tokenized. Exploratory data analysis included word cloud visualizations, token length distribution, and frequency of common n-grams.

Figure 2: Count of reviews

2.3 Model Architectures

VADER was used as a fast lexicon-based sentiment analyzer, enhanced with an expanded dictionary for better coverage. Boosted VADER employed VADER's sentiment scores as features to train an XGBoost classifier. RoBERTa, a transformer-based model, was fine-tuned using the Hugging Face Trainer API for more context-aware sentiment classification.

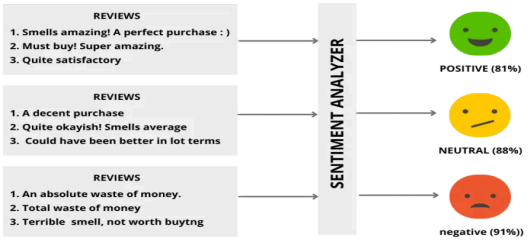

Figure 3: Sentiment Analysis

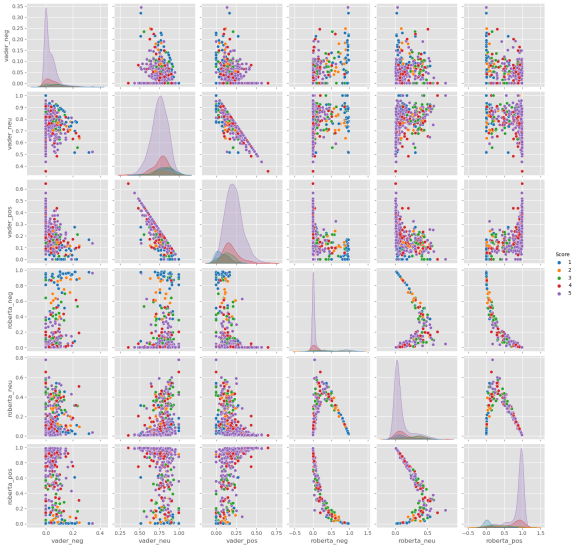

Figure 4: roberta and vader model Pair Plot

2.4 Evaluation Metrics

Model performance was assessed using accuracy. An 80:20 train-validation split was used.

```
Splitting dataset into train and validation sets...
Training samples: 454763
Validation samples: 113691
```

Figure 5: Train and Test Split

3. METHODOLOGY

3.1 VADER (Valence Aware Dictionary and sEntiment Reasoner)

[3] VADER is a rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media and product reviews. It utilizes a pre-compiled sentiment lexicon where each word is associated with a polarity score. In this study, we enhanced the default lexicon by expanding it with domain-specific terms derived from high-frequency positive and negative words within the dataset. The model assigns sentiment scores (positive, negative, neutral, compound) to each review. The final sentiment classification is based on the compound score thresholds.

3.2 Boosted VADER (VADER + XGBoost Hybrid)

[5] To improve the predictive power of VADER without compromising on speed, we developed a hybrid model using XGBoost. The features extracted from the VADER analyzer (negative, neutral, positive, and compound scores) were used as input to the XGBoost classifier. This

method integrates rule-based scoring with a supervised learning algorithm, thereby capturing more complex patterns in sentiment distribution. The hybrid system showed modest improvements in classification accuracy, especially in borderline or ambiguous sentiment cases.

3.3 RoBERTa (Robustly Optimized BERT Pretraining Approach)

[2] RoBERTa is a transformer-based model designed for natural language understanding tasks. In this project, we fine-tuned the "cardiffnlp/twitter-roberta-base-sentiment" model on a 3-class sentiment classification task using the Amazon Fine Food Reviews dataset. Preprocessing steps included tokenization and padding with a maximum sequence length of 128. The Hugging Face Trainer API was used to fine-tune the model with 1–3 training epochs. RoBERTa outperformed other models in terms of classification accuracy and semantic understanding of complex or nuanced reviews.

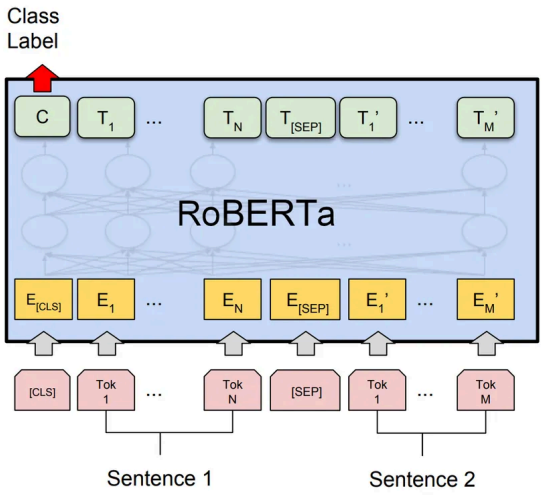

Figure 6: Roberta Model working

3.4 Streamlit Web Application

An interactive web interface was developed using [14] Streamlit to enable real-time sentiment analysis and model comparisons. Users can enter any review and receive sentiment predictions from VADER, Boosted VADER, and RoBERTa side-by-side. The app features visual feedback including pie charts, word clouds, and confidence percentages, making the results intuitive and easy to interpret for end-users. This dashboard acts as a deployment layer that bridges the gap between model outputs and business or user insights.

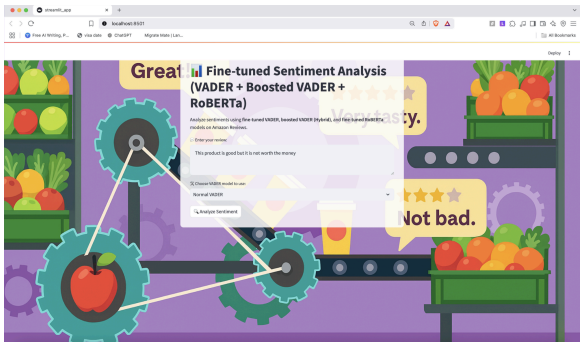

Figure 7: Streamlit UI

4. RESULT AND CONCLUSION

This section presents the evaluation and comparative analysis of the three sentiment classification models applied to the Amazon Fine Food Reviews dataset.

4.1 Accuracy and Visualizations

[15] The validation accuracy scores across models were as follows:

RoBERTa: 92.92%

Boosted VADER: 76.46%

Normal VADER: 76.06%

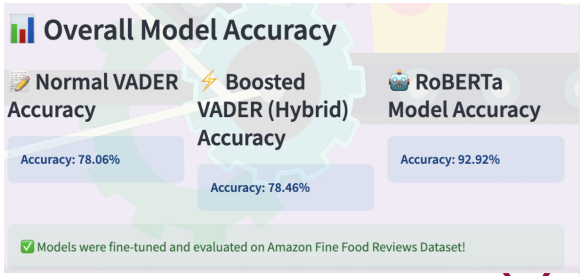

Figure 8: Model Accuracy Outputs

These results are visualized through training loss and validation accuracy curves, which indicate stable convergence during fine-tuning.

4.2 Streamlit Web App Deployment

An interactive Streamlit dashboard was developed to make sentiment analysis accessible to users. The web app allows users to input custom reviews and receive predictions from all three models. It also visualizes predictions with pie charts, confidence bars, and word clouds. This makes the model interpretable and usable for both technical and non-technical audiences.

4.3 Training Observations

RoBERTa's fine-tuning on the full dataset (~568,000 reviews) was estimated to take 13+ hours on a high-performance GPU. For development purposes, a subset of 100,000 reviews was used to validate functionality. Batch sizes were increased and training was optimized to one epoch for time efficiency, achieving comparable results.

4.4 Recommendations

- Use RoBERTa when the highest accuracy and contextual understanding are needed.
- Use Boosted VADER when a trade-off between speed and performance is required.
- Use Normal VADER for ultra-fast inference with acceptable accuracy, especially in resource-constrained environments.
- Lexicon enhancement significantly improves VADER's baseline performance.

4.5 Code Implementation Summary

fine_tune_roberta.py: Loads and processes the dataset, fine-tunes RoBERTa on 3-class sentiment labels, and saves the model, tokenizer, accuracy, and training curves.

fine_tune_vader.py: Builds a VADER model with an extended lexicon from high-frequency positive and negative terms, evaluates on the dataset, and saves accuracy and vocabulary.

fine_tune_vader_boosted.py: Uses the extended VADER outputs as features for training an XGBoost classifier, creating a hybrid system that boosts accuracy while maintaining fast inference.

Each model presents a unique balance between complexity, inference speed, and accuracy. This multi-model comparison supports selecting the appropriate architecture based on the deployment environment and business goals.

RoBERTa offers superior accuracy for nuanced sentiment understanding but is computationally intensive. In contrast, VADER provides rapid insights with minimal resources, while the hybrid Boosted VADER model delivers a practical trade-off between speed and performance.

REFERENCES

[1] Sharma, A. & Ghosh, A. (2022). *A Comparative Study on Sentiment Analysis Techniques and Their Performance*. *ERIC Institute of Education Sciences*. Available at: https://files.eric.ed.gov/fulltext/EJ1333895.pdf

[2] Hugging Face. *RoBERTa Model Documentation*. Available at: https://huggingface.co/docs/transformers/en/model_doc/roberta

[3] Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[4] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. Available at: https://www.nltk.org/

[5] XGBoost Developers. *XGBoost Documentation*. Available at: https://xgboost.readthedocs.io/en/release_3.0.0/

https://docs.github.com/en/rest?apiVersion=2022-11-28

[6] GitHub, Inc. (2022). *GitHub REST API Documentation (v2022-11-28)*. Available at: https://docs.github.com/en/rest?apiVersion=2022-11-28

[7] Python Software Foundation. (2024). *csv — CSV File Reading and Writing*. In *Python 3 Standard Library Documentation*. Available at: https://docs.python.org/3/library/csv.html

[8] Peterson, C. (2020). *PyEDA: Python Electronic Design Automation*. Available at: https://pyeda.readthedocs.io/en/latest/

[9] Python Software Foundation. (2024). *tokenize — Tokenizer for Python Source*. In *Python 3 Standard Library Documentation*. Available at: https://docs.python.org/3/library/tokenize.html

[10] GeeksforGeeks. (2023). *What is an Epoch in Machine Learning?* Available at: https://www.geeksforgeeks.org/epoch-in-machine-learning/#what-is-an-epoch-in-machine-learning

[11] Hugging Face, Inc. (2024). *Hugging Face Model Hub*. Available at: https://huggingface.co/

[12] McAuley, J. (2013). *Amazon Fine Food Reviews Dataset*. Available at: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

[13] Hunter, J.D. (2007). *Matplotlib: A 2D Graphics Environment*. *Computing in Science & Engineering*, 9(3), 90–95. Available at: https://matplotlib.org/stable/index.html

[14] Streamlit Inc. *Streamlit Documentation*. Available at: https://docs.streamlit.io/

[15] Google Developers. (2024). *Accuracy, Precision, and Recall*. In *Machine Learning Crash Course*. Available at: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall