University of New Haven

Team: Data Ops

Team Members:

- Team Lead & Data Scientist: Naga Swetha Nallaninnall10@unh.newhaven.edu
- Data Engineer: Govardhan Namala – gnama2@unh.newhaven.edu
- Data analyst: Amani Ponnam – aponn6@unh.newhaven.edu

Research Question:

How do consumer sentiments in product reviews influence brand loyalty and perception of Apple's product ecosystem?

The Dataset we chose is the Apple Product Review Dataset which is available in Kaggle at this link

Github Repository:

A GitHub repository has been created for this final project and all the source code used for this phase -6 is uploaded in a different branch named Phase-6

You can access the repository at this link. The code can be seen in the said branch.

Data Processing and Splitting:

Divided data into 80% training and 20% testing sets. Ensured no data leakage by applying preprocessing independently on both splits**.**

Data Mining Techniques Used:

1) Logistic Regression
2) Naive Bayes
3) Support Vector Machine
4) XG Boost
5) Ensemble Method (Naive Bayes + Logistic Regression)
6) CNN Model

Key parameters for each model:

1) Logistic Regression: C=1.0, solver='liblinear'
2) Naive Bayes: alpha=1.0
3) SVM: kernel='rbf', C=2.0, gamma=0.1
4) XGBoost: learning_rate=0.1, max_depth=6, n_estimators=100
5) Ensemble Method: TF-IDF vectorization (ngram_range=(1,2), max_df=1.0, min_df=2), GridSearchCV for Naive Bayes (alpha=[0.1, 0.5, 1.0], cv=5, scoring='accuracy'), Logistic Regression (max_iter=1000, random_state=42), and soft voting in the ensemble model.
6) CNN: Embedding (128-dim), Conv1D (128 filters, kernel size=5, activation='relu'), GlobalMaxPooling1D, Dense (64 units, activation='relu'), Dropout (rate=0.5), output Dense (3 units, activation='softmax'), compiled with 'categorical_crossentropy', 'adam', and 'accuracy', trained for 10 epochs with batch_size=32

Hyper Parameters used to tune each model are:

1) Logistic Regression: Increased regularization strength by setting C from 1.0 to 1.5
2) Naive Bayes: Reduced smoothing parameter alpha from 1.0 to 0.5.
3) Support Vector Machine: RandomizedSearchCV explored 50 combinations of C and gamma across a broader range C searched in the range [0.1, 10] and gamma searched in the range [0.001, 1]
4) XG Boost:
   learning_rate: Explores smaller increments for finer control over step size.
   max_depth: Includes a range of depths for flexibility.
   subsample and colsample_bytree: Regularization through sampling, preventing overfitting.
   gamma, reg_alpha, and reg_lambda: Regularization terms to balance model complexity.
   RandomizedSearchCV:
   Efficiently explores a larger parameter space with fewer iterations than Grid Search
5) Ensemble Method (Naive Bayes + Logistic Regression): GridSearchCV-tuned Logistic Regression for C values and liblinear solver, combined with Naive Bayes
6) CNN Model:
7) Architecture: Embedding layer → SpatialDropout1D (0.3) → Conv1D (128 filters, kernel size 5, ReLU, L2 regularization) → Additional Conv1D (128 filters, kernel size 3, ReLU, L2 regularization) → GlobalMaxPooling → Dense (128 units, L2 regularization) → Dropout (0.6) → Dense output layer (softmax for 3 classes). Optimizer: Adam with a reduced learning rate of 0.0001

Optimization Techniques: Enhancements from Different Perspectives

Performance Metrics:

- Logistic Regression and Naive Bayes benefitted from **parameter tuning**, improving their ability to generalize.
- SVM and XGBoost showed significant improvements in **AUC**, indicating better discrimination capabilities for sentiment classification.
- All optimization methods (ensemble approaches and CNN tuning) showed consistent accuracy improvements, with the Enhanced CNN achieving the highest accuracy at **88.56%**.
- Optimization techniques significantly enhanced AUC values, with the Enhanced CNN achieving **0.97**, reflecting superior classification performance.

Error Reduction:

- Logistic Regression and Naive Bayes reduced misclassification in neutral classes, with optimized regularization and smoothing.
- SVM and XGBoost minimized false negatives in positive sentiments, improving precision and recall.
- The Enhanced CNN reduced errors in challenging classes (neutral and slightly negative sentiments), as observed in its confusion matrix.
- Ensemble methods mitigated individual model weaknesses, particularly addressing false positives and false negatives in the positive class.

Model Efficiency:

- While Logistic Regression and Naive Bayes were computationally efficient, SVM and XGBoost required more resources but yielded higher accuracy after optimization.
- Optimization methods like **RandomizedSearchCV** efficiently explored hyperparameter spaces for SVM and XGBoost, yielding substantial performance improvements without excessive computational cost.
- The enhanced CNN architecture, although computationally intensive, demonstrated the payoff in achieving state-of-the-art results.

Scalability:

- XGBoost proved highly scalable, benefiting from parallel computation and subsampling techniques during optimization.
- SVM required careful parameter tuning to handle larger datasets efficiently.

Model Stability:

- Regularization techniques in the Enhanced CNN (L2 regularization and dropout layers) ensured stability, reducing overfitting tendencies seen in the initial CNN.
- The tuned Logistic Regression provided a more stable decision boundary for the ensemble models.

Visualization Techniques

To communicate our findings effectively, we employed several visualization techniques:

- **ROC Curves**: Generated for each model, allowing us to compare their ability to distinguish between positive, negative, and neutral sentiments. The ROC curve for the enhanced CNN model showed a significant improvement, indicating a better balance between sensitivity and specificity.
- **Confusion Matrices**: Displayed the distribution of correctly and incorrectly classified sentiments for each model. The CNN model's confusion matrix showed fewer misclassifications, particularly in distinguishing between neutral and negative sentiments, compared to other models.
- **Consolidated Performance Table**: Summarized all metrics for each model, providing a clear side-by-side comparison that highlighted the CNN's superior performance. This table allowed for quick evaluation of each model's effectiveness in meeting the project's objectives.
- Histogram for all Accuracies: Summarized all accuracies for each model and visulaised in a histogram.

Conclusion:

Optimization significantly improved our ability to answer how consumer sentiments impact Apple's brand loyalty. Baseline models provided moderate accuracy, with Naive Bayes achieving **68% accuracy** and an AUC of **0.45**, but struggled with nuanced sentiments like neutral or slightly negative. Through parameter tuning and ensemble methods, models like SVM and XGBoost showed notable improvements, and the **Enhanced CNN** achieved **88.56% accuracy** and an **AUC of 0.97**, capturing complex sentiment patterns effectively. These refinements transformed our results from basic insights to actionable understanding, showcasing the value of optimization in sentiment analysis.

Our findings suggest that while traditional models can provide foundational insights, deep learning methods like CNNs are better suited to the intricate nature of sentiment analysis in product reviews. Future research could build upon these results by implementing even more advanced NLP techniques, such as transformer-based models, which may yield deeper insights given appropriate computational resources.