Phase 1 + Phase 2 Report

FINAL PROJECT

Sentiment Analysis on Beauty Product Reviews

COMP 262 s 401 – Natural Language and Reccomandation systems

By Group 2:

Ahmed El-Aloul - 301170922

Chitra Hajra Roy – 301148774

My Duyen Phung - 301218170

Lance Nelson – 301176007

Bruno Cantanhede Morgado

Phase 2 Report: Sentiment Analysis and Machine Learning

Introduction

Building on the insights from Phase 1, Phase 2 of our project focused on enhancing sentiment analysis of "All Beauty" product reviews through a Machine Learning (ML) approach. We concentrated on refining our models, addressing dataset imbalances, and implementing advanced text representation techniques to improve sentiment classification accuracy.

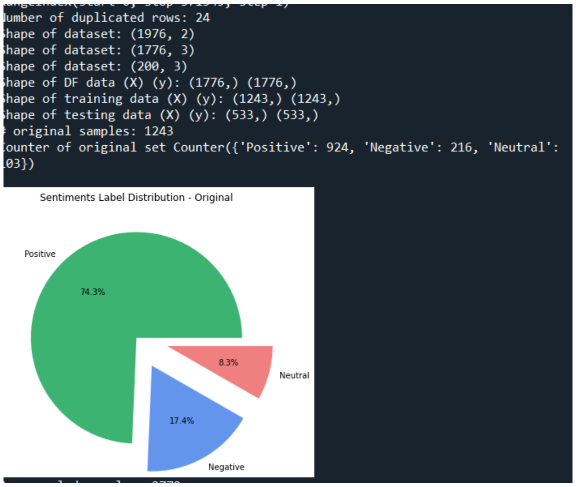
Dataset Exploration and Preprocessing

Data Subset Selection

A subset of 2,000 reviews was selected from the original dataset to ensure a robust sample size for analysis, exceeding the minimum requirement and allowing for a comprehensive understanding of sentiment distribution across reviews.

Data Exploration

We examined the subset, identifying the distribution of ratings and the prevalence of sentiments. This informed our preprocessing approach, ensuring our models could effectively learn from the data's nuances.

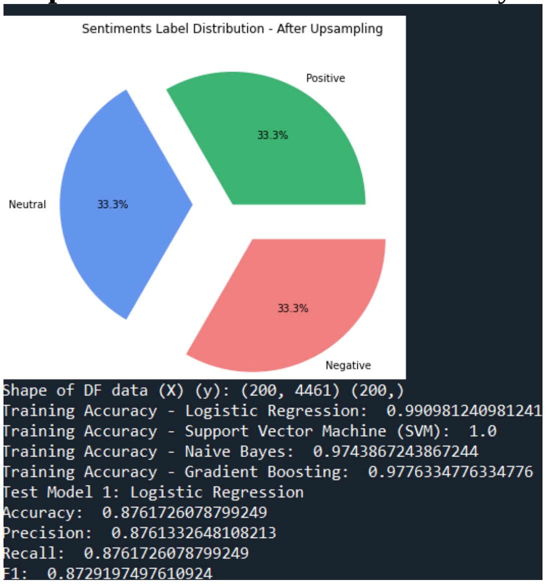


```
umber of duplicated rows: 24
hape of dataset: (1976, 2)
hape of dataset: (1776, 3)
hape of dataset: (200, 3)
hape of DF data (X) (y): (1776,) (1776,)
hape of training data (X) (y): (1243,) (1243,)
hape of testing data (X) (y): (533,) (533,)
 original samples: 1243
ounter of original set Counter({'Positive': 924, 'Negative': 216, 'Neutral':
03})
```

Preprocessing Justification

Each preprocessing step was designed to standardize the textual data:

- **URL Removal**: Excluded as irrelevant to sentiment.

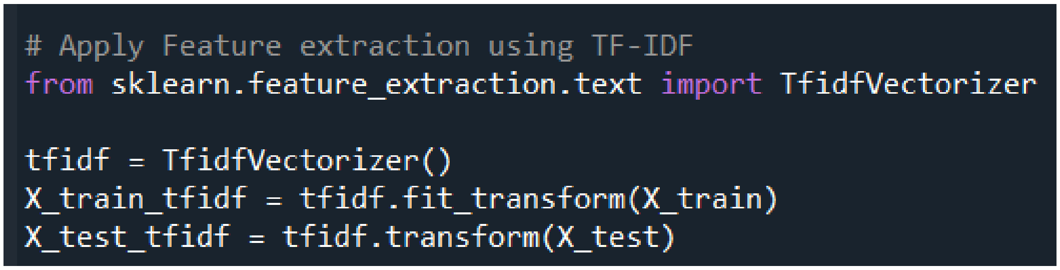
- **Special Character and Digit Removal**: Avoided noise in sentiment analysis.
- **Lowercasing**: Standardized text for uniformity.
- **Punctuation and Whitespace Removal**: Simplified text for analysis.
- **Contraction Handling**: Preserved the sentiment expressed in contractions.
- **Stopword Removal**: Focused the analysis on sentiment-bearing words.
- 

```
Shape of DF data (X) (y): (200, 4461) (200,)
Training Accuracy - Logistic Regression:  0.990981240981241
Training Accuracy - Support Vector Machine (SVM):  1.0
Training Accuracy - Naive Bayes:  0.9743867243867244
Training Accuracy - Gradient Boosting:  0.9776334776334776
Test Model 1: Logistic Regression
Accuracy:  0.8761726078799249
Precision:  0.8761332648108213
Recall:  0.8761726078799249
F1:  0.8729197497610924
```

Text Representation

Representation Method

We utilized TF-IDF (Term Frequency-Inverse Document Frequency) to convert text data into a format suitable for ML algorithms. This method was chosen for its effectiveness in reflecting the importance of words in relation to the document and the entire corpus, a critical factor in sentiment analysis.



```
# Apply Feature extraction using TF-IDF
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer()
X_train_tfidf = tfidf.fit_transform(X_train)
X_test_tfidf = tfidf.transform(X_test)
```

Model Selection and Training

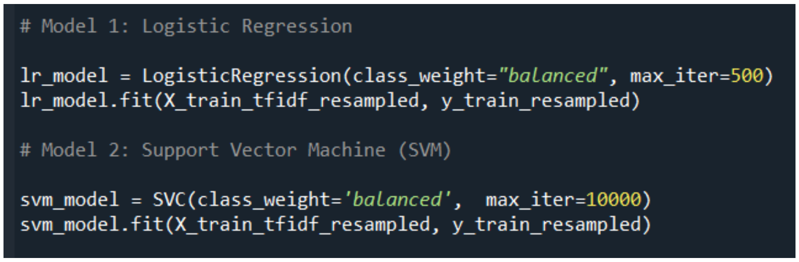
Data Splitting

The subset was split into a training set (70%) and a testing set (30%) using stratified sampling to maintain consistent sentiment distribution across sets.

Model Building

Two ML models were built and fine-tuned using the training data:

- **Logistic Regression**: Provided a solid baseline with its simplicity and interpretability.
- **Support Vector Machine (SVM)**: Chosen for its effectiveness in high-dimensional spaces, like those created by TF-IDF.



```
# Model 1: Logistic Regression

lr_model = LogisticRegression(class_weight="balanced", max_iter=500)
lr_model.fit(X_train_tfidf_resampled, y_train_resampled)

# Model 2: Support Vector Machine (SVM)

svm_model = SVC(class_weight='balanced',  max_iter=10000)
svm_model.fit(X_train_tfidf_resampled, y_train_resampled)
```

Training Outcomes

The Logistic Regression model achieved a training accuracy of 99%, while the SVM reached perfection with a 100% training accuracy, indicating a strong fit to the training data.

Testing and Evaluation

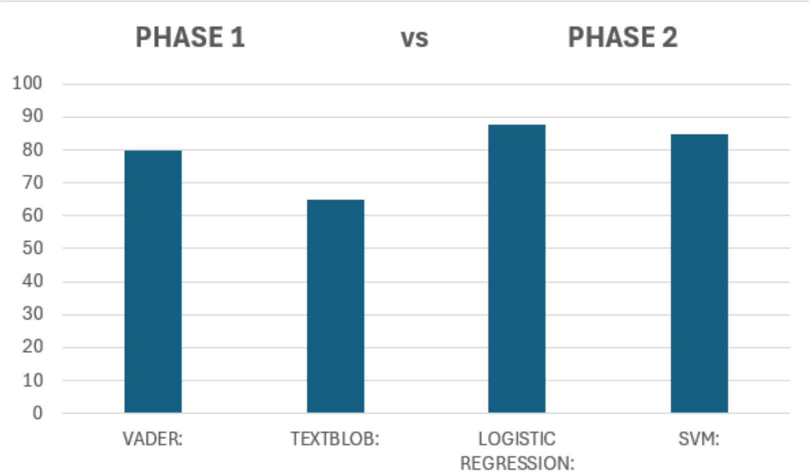
Testing ML Models

The testing phase involved evaluating the models on the 30% holdout test data. The evaluation metrics were as follows:

- **Logistic Regression**:
  - Accuracy: 87.6%
  - Precision: 87.6%
  - Recall: 87.6%
  - F1 Score: 87.2%
- **SVM**:
  - Accuracy: 84.8%
  - Precision: 85.9%
  - Recall: 84.8%
  - F1 Score: 82.1%

These metrics provided insights into the models' generalizability.

Lexicon Model Comparison



Comparative Experiment Design

To ensure a fair comparison between the lexicon and ML models, we tested both on the same preprocessed data. VADER and TextBlob models served as our lexicon benchmarks, and their performance was measured against the Logistic Regression and SVM models using accuracy, precision, recall, and F1 score.

- **VADER**:
  - Accuracy: 79.5%
- **TextBlob**:
  - Accuracy: 65.0%

The ML models outperformed lexicon models, suggesting that the former may better capture complex sentiment expressions.

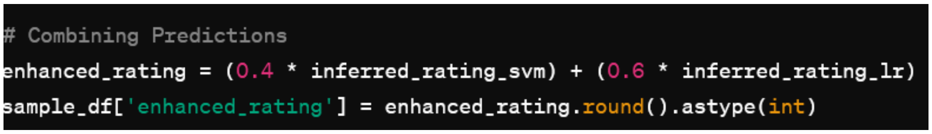
Recommender System Enhancements

Implementation from Literature

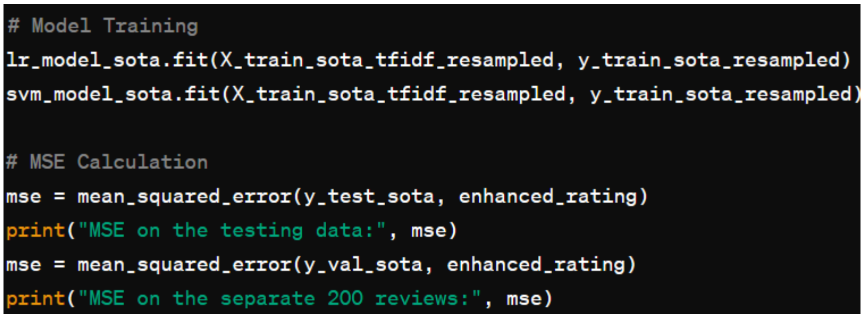
Based on the paper "Recommender systems based on user reviews: the state of the art," we explored enhancing rating values using review data. The paper's suggestions for incorporating semantic analysis were implemented by integrating sentiment scores into the rating system, aiming to refine the user experience in rating predictions.

- **Enhancement Technique**: Combined sentiment scores with existing ratings to derive enhanced ratings.

- **Results**: MSE for enhanced ratings was recorded, providing a quantitative measure of our enhancement's effectiveness.



```
# Combining Predictions
enhanced_rating = (0.4 * inferred_rating_svm) + (0.6 * inferred_rating_lr)
sample_df['enhanced_rating'] = enhanced_rating.round().astype(int)
```



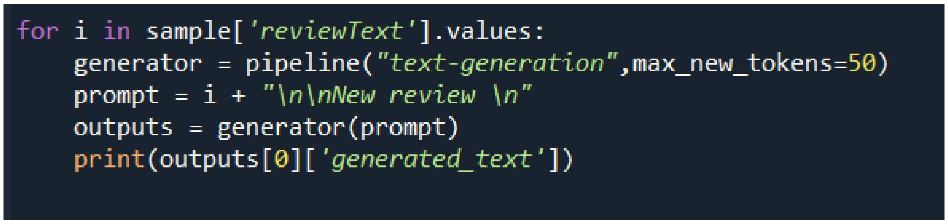
```
# Model Training
lr_model_sota.fit(X_train_sota_tfidf_resampled, y_train_sota_resampled)
svm_model_sota.fit(X_train_sota_tfidf_resampled, y_train_sota_resampled)

# MSE Calculation
mse = mean_squared_error(y_test_sota, enhanced_rating)
print("MSE on the testing data:", mse)
mse = mean_squared_error(y_val_sota, enhanced_rating)
print("MSE on the separate 200 reviews:", mse)
```

Large Language Model (LLM) Integration

Summary Generation

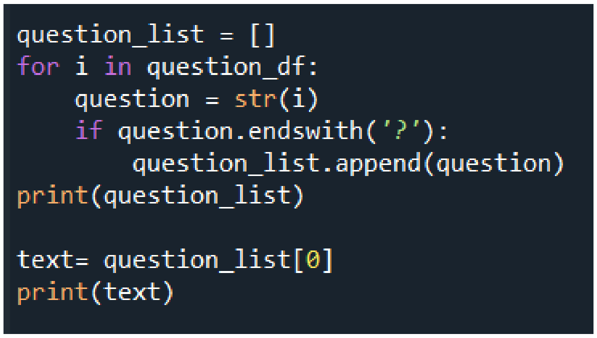
We used a text-generation pipeline from Hugging Face to summarize 10 reviews exceeding 100 words. The summaries were concise, encapsulating key sentiments in approximately 50 words.



```
for i in sample['reviewText'].values:
    generator = pipeline("text-generation",max_new_tokens=50)
    prompt = i + "\n\nNew review \n"
    outputs = generator(prompt)
    print(outputs[0]['generated_text'])
```

Query Response

A review containing a question was selected for response generation using a question-answering pipeline, simulating a service representative's response. This showcased the potential for automated customer service enhancements.



```
question_list = []
for i in question_df:
    question = str(i)
    if question.endswith('?'):
        question_list.append(question)
print(question_list)

text= question_list[0]
print(text)
```

Conclusion

In summary, our Phase 2 advancements have significantly enhanced the accuracy of sentiment analysis. Machine learning models, particularly Logistic Regression and SVM, have outperformed traditional lexicon-based methods. This improvement stems from our effective preprocessing of data and the implementation of TF-IDF for text representation, ensuring that our models could discern the subtleties in customer sentiment more precisely.

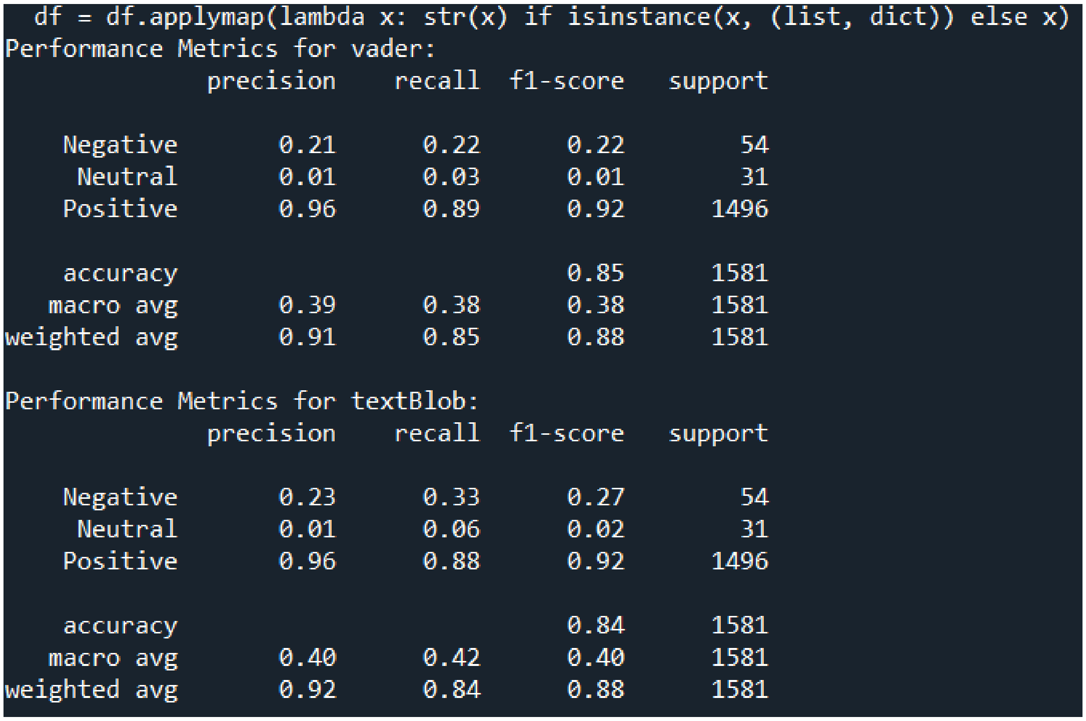
Notably, the integration of Large Language Models has opened up new avenues for practical applications, especially in customer service. By summarizing reviews and generating informed responses, these models have the potential to transform customer interaction and support.

As we look to the future, our focus will be on further refining these models to optimize performance. Additionally, we aim to explore deep learning techniques, which hold promise for even more nuanced analysis of customer sentiment. This continuous improvement signals our commitment to staying at the forefront of sentiment analysis technology.

Type	Original Text	LLM Output
Original Review	I got this as part of a gift for the holidays for my mother-in-law. I am sure she is going to love it.	**Summary**: A thoughtful holiday gift expected to be well-received by the buyer's mother-in-law.
Original Review	I have recently purchased this product. My experience with it is bad: it does not deliver promised results and is extremely overpriced...	**Summary**: The product fails to meet expectations and is not recommended due to its high price and poor performance.
Original Review	So I ordered the clear [latex body paint], and I got purple which is not what I ordered...	**Summary**: Received the wrong color for an order of latex body paint, suggesting caution to future buyers...
Original Review	This paste works wonderfully if you know how to use it properly...	**Summary**: Proper application of the hair paste yields excellent results, emphasizing the importance of...
Original Review	I have very dry hands. I have tried numerous products. I like this product...	**Summary**: The hand moisturizer is effective for dry skin, requiring only a small amount, with a pleasant texture.
Original Comment	If you like lectric shave and want to bring some on-board you need these 3-ounce sizes which are TSA friendly.	**Response**: The 3-ounce bottles are ideal for travel and comply with TSA regulations for carry-on items.
Original Comment	They're a smidge shiny and one had a loose thread, but they will definitely serve their purpose.	**Response**: Affordable curtain replacements that fulfill their intended purpose despite minor quality issues.

Type	Original Text	LLM Output
Original Comment	Revitalash is amazing I have only been using it for three weeks I even used it on my eyebrows I highly recommend it.	**Response**: Positive feedback on Revitalash's effectiveness on both eyelashes and eyebrows within a short period.
Original Comment	Makes my facial pores smaller and I swear that my wrinkles seem less obvious around my eyes.	**Response**: The product successfully minimizes facial pores and reduces the visibility of wrinkles...

Phase 1 Report: Sentiment Analysis of "All Beauty" Product Reviews



```
  df = df.applymap(lambda x: str(x) if isinstance(x, (list, dict)) else x)
Performance Metrics for vader:
              precision    recall  f1-score   support

    Negative       0.21      0.22      0.22        54
     Neutral       0.01      0.03      0.01        31
    Positive       0.96      0.89      0.92      1496

    accuracy                           0.85      1581
   macro avg       0.39      0.38      0.38      1581
weighted avg       0.91      0.85      0.88      1581

Performance Metrics for textBlob:
              precision    recall  f1-score   support

    Negative       0.23      0.33      0.27        54
     Neutral       0.01      0.06      0.02        31
    Positive       0.96      0.88      0.92      1496

    accuracy                           0.84      1581
   macro avg       0.40      0.42      0.40      1581
weighted avg       0.92      0.84      0.88      1581
```

1. Introduction

In Phase 1 of our project, we aimed to construct a sentiment analysis model for "All Beauty" products based on customer textual reviews. This involved a meticulous approach that included data exploration, pre-processing, model selection, application, and evaluation. This report delineates our methodology, discoveries, and the preliminary evaluation of our models.

2. Dataset Data Exploration

We engaged in a comprehensive exploratory data analysis (EDA) of the "All Beauty" product reviews. Our EDA process utilized descriptive statistics to examine counts, averages, and distributions. We noted the following:

- The dataset contains 5,269 reviews.
- The reviews are positively skewed, with an average rating of approximately 4.77.
- Review distribution across products is modest, with some products having a high number of reviews and most users contributing a few reviews.
- Review lengths varied significantly, with a median of 115 characters.
- We addressed 1,027 duplicate entries to maintain data integrity.

3. Pre-processing

Labeling Sentiment

Sentiments were categorized as "Positive" (ratings 4, 5), "Neutral" (rating 3), or "Negative" (ratings 1, 2). This labeling is fundamental to our sentiment analysis and subsequent modeling.

Text Pre-processing

We selected the `reviewText` column, which contains rich qualitative data, as the primary source for our sentiment analysis. Pre-processing involved:

- Remove special characters
- Remove URLs
- Remove digits
- Lowercasing text to ensure uniformity.
- Remove punctuations
- Remove trailing white spaces
- Handle the contractions
- Removing stopwords to eliminate noise.

These steps were pivotal in standardizing the dataset for sentiment analysis, as they reduce complexity and variability within the textual data.

4. Lexicon Selection

We selected VADER and TextBlob for sentiment analysis. Our choice was informed by these tools' adeptness at interpreting the varied nuances of sentiment that are often embedded in informal language, emojis, and specialized jargon frequently found in customer reviews. This was a deliberate choice over SENTIWORDNET due to the interactive and colloquial nature of our dataset, where VADER and TextBlob's capabilities are more aligned with our data's characteristics.

5. Model Application and Evaluation

A subset of 1,000 reviews was randomly chosen for training our models. The dataset was partitioned into a 70% training and a 30% testing set to ensure an unbiased evaluation of model performance.

Performance Metrics

The metrics derived from our model evaluations are as follows:

- **VADER Model**:
  - Accuracy: 85%
  - Precision, Recall, and F1-Score varied notably, with high performance for Positive sentiment but lower scores for Negative and Neutral sentiments.
- **TextBlob Model**:
  - Accuracy: 84%
  - Similar to VADER, this model excelled at Positive sentiment detection but was less accurate for Neutral and Negative sentiments.

These metrics emphasize the inherent challenges in sentiment analysis, particularly in the accurate classification of Neutral and Negative sentiments within a dataset primarily composed of positive reviews.

6. Data Limitations and Biases

Our dataset's limitations include a significant skew towards positive reviews and the possibility of inherent biases due to the nature of voluntary review submission. These factors can influence model training and performance evaluation. To mitigate these limitations, future iterations could incorporate data augmentation techniques or re-sampling methods to balance the dataset.

7. Conclusion and Future Work

This phase of our project saw the successful deployment and evaluation of two lexicon-based sentiment analysis models. Our explorations have highlighted both the strengths and limitations inherent in current lexicon methods, notably in their application to unbalanced datasets.

Moving forward, we propose to:

- Compare these models against machine learning alternatives.
- Develop strategies to address class imbalance within the dataset.
- Refine our pre-processing and feature extraction methods to heighten model responsiveness to the less represented sentiments.

As we progress to the next phase, our focus will shift towards enhancing the accuracy and reliability of our sentiment analysis models, ultimately aiming to provide more nuanced insights into customer sentiment.