# Report on Sentiment Analysis Using Vader, Flair, and TextBlob on Yelp Restaurant Reviews Dataset

This report evaluates the performance of three sentiment analysis tools: **Vader**, **TextBlob**, and **Flair**, applied to a large Yelp restaurant reviews dataset containing approximately 4.7 million records. Each tool's results are compared based on accuracy, Matthews Correlation Coefficient (MCC), classification reports, and confusion matrices.

## Overview of Tools

1. **Vader**
    a. A lexicon-based sentiment analysis tool
    b. Suitable for social media or informal text
    c. Outputs polarity scores (compound scores) used to classify sentiments into positive, neutral, or negative
2. **TextBlob**
    a. A simple lexicon-based sentiment analysis tool
    b. Uses a predefined polarity scale (-1 to +1) to determine sentiment categories
3. **Flair**
    a. A neural-network-based sentiment analysis model.
    b. Pre-trained on various datasets for binary sentiment classification (positive/negative)

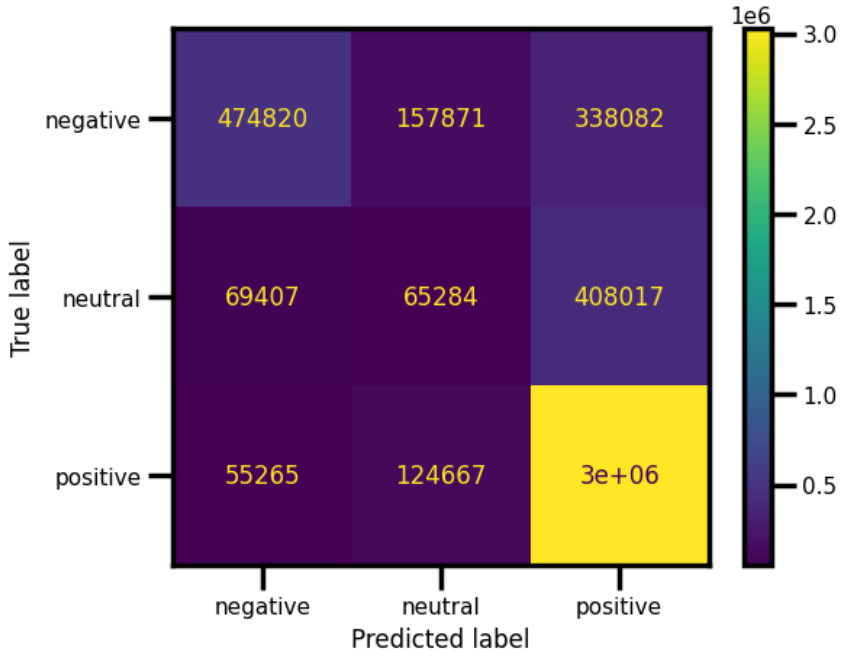
## Summary of Results

### Vader Sentiment Analysis

**Key Results:**

- **Accuracy**: 75.56%
- **Classification Report**:
    - Negative:
        - Precision: 0.79
        - Recall: 0.49

        - F1-Score: 0.60
    - Neutral:
        - Precision: 0.19
        - Recall: 0.12
        - F1-Score: 0.15
    - Positive:
        - Precision: 0.80
        - Recall: 0.94
        - F1-Score: 0.87
- **Confusion Matrix**




Vader performed well on predicting positive sentiments achieving a high precision score of 80% and a recall score of 94%, resulting in an F1-score of 87%. This indicates that most of the positive reviews were correctly predicted. However, a poor performance on neutral reviews was recorded, With a recall score of only 12%. This indicates that Vader had trouble with correctly identifying neutral sentiments. From the confusion matrix above, we see that most misclassifications involved neutral sentiments being classified as either positive or negative.

For classifying negative reviews, vader recorded a precision score of 79%. Precision indicates the proportion of reviews classified as negative that were truly negative. With a precision of 79%, Vader shows relatively strong performance in ensuring that reviews labeled as negative are indeed negative. However, this still means about 21% of reviews classified as negative may be

false positives. With only 49% recall, Vader misses more than half of the truly negative reviews, this low recall indicates that the model struggles to capture all negative sentiment effectively.

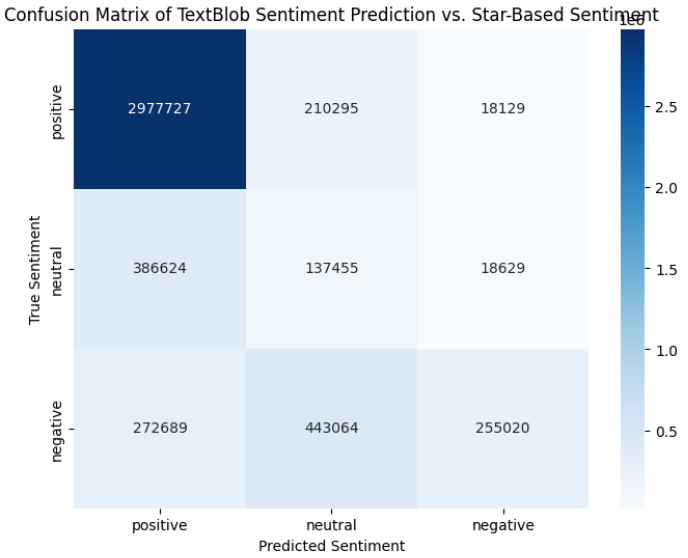
The F1-score of 60% reflects the imbalance between good precision and poor recall. While Vader avoids labeling non-negative reviews as negative, it fails to identify many actual negative reviews.

From these results, one may infer that Vader is best suited for polar sentiments (positive/negative) but struggles with neutrality. It's reliance on a lexicon-based approach may limit its ability to handle nuanced reviews.

## TextBlob Sentiment Analysis

**Key Results**

- **Accuracy:** 71%
- **Matthews Correlation Coefficient (MCC):** 0.36
- **Classification Report:**
    - Positive:
        - Precision: 0.86
        - Recall: 0.26
        - F1-Score: 0.40
    - Neutral:
        - Precision: 0.17
        - Recall: 0.23
        - F1-Score: 0.20
    - Negative:
        - Precision: 0.81
        - Recall: 0.93
        - F1-Score: 0.86
    - **Confusion Matrix:**




TextBlob's performance on negative reviews is good, with a high recall score of 93% and a precision score of 81% indicating reliability in detecting negative sentiments and ensuring relatively few false positives for negative reviews. In classifying positive sentiments, TextBlob recorded a low recall of 26% which means most positive reviews were misclassified. A low precision and recall score was recorded for the neutral class indicating a significant misclassification of neutral reviews.

For predicting positive sentiment, a precision score of 86% indicates that 86% of the reviews classified as positive by TextBlob were truly positive. A high precision score suggests that TextBlob performs well at avoiding false positives, meaning most of the reviews it identifies as positive are accurate. The low recall score shows that TextBlob struggles to identify all positive reviews, capturing only 26% of the actual positive instances. This suggests that many positive reviews are being misclassified as neutral or negative.

A MCC score of 0.36 indicates a weak-to-moderate positive correlation between the predicted and true labels. The low score here suggests that while TextBlob performs well on certain classes (e.g., negative sentiment), it struggles with others (e.g., positive and neutral), leading to an overall mixed performance. The MCC indicates that TextBlob's polarity-based approach is not robust for handling nuanced sentiments, especially in distinguishing between positive and neutral classes.

TextBlob's reliance on a simple polarity score to determine sentiment may fail to capture the nuances of positive sentiments, particularly in ambiguous or complex review texts.

## Flair Key Results

- **Accuracy**: 90%
- **Matthews Correlation Coefficient (MCC)**: 0.78
- **Classification Report**:
    - Positive:
        - Precision: 0.81
        - Recall: 0.89
        - F1-Score: 0.85
    - Negative:
        - Precision: 0.95
        - Recall: 0.90
        - F1-Score: 0.92

Flair achieved a 90% accuracy score, likely due to its neural network-based approach. High precision and recall for the negative class (Precision: 95%, Recall: 90%) ensure minimal false negatives and false positives. Flair also excels with the positive class, achieving an F1-score of 85%, which indicates reliable performance for favorable reviews. However, it struggled slightly more with positive reviews than negative reviews, as indicated by the slightly lower precision (81%). The distribution of misclassifications is significantly smaller compared to TextBlob, suggesting Flair's deep learning model captures nuanced sentiment expressions better.

The Matthews Correlation Coefficient (MCC) for Flair sentiment analysis is 0.78, a significant improvement over TextBlob's MCC score of 0.36. This indicates that Flair has strong predictive power and is highly effective in distinguishing between positive and negative classes. It also reflects strong agreement between the predicted and true labels. This shows that Flair is effective in capturing sentiment nuances in the Yelp restaurant reviews dataset.

| ***Metric*** | **Vader** | **TextBlob** | **Flair** |
|---|---|---|---|
| ***Accuracy*** | 75.56% | 71% | 90% |
| ***MCC*** | - | 0.36 | 0.78 |
| ***Positive Precision*** | 80% | 86% | 81% |
| ***Positive Recall*** | 94% | 26% | 89% |
| ***Negative Precision*** | 79% | 81% | 95% |

| | | | |
|---|---|---|---|
| ***Negative Recall*** | 49% | 93% | 90% |
| ***Neutral F1-Score*** | 15% | 20% | - |

# Report on Sentiment Analysis Using Logistic Regression, Multinomial naïve Bayes, Decision Trees, and Random Forest on Yelp Restaurant Reviews Dataset

The models were implemented using a balanced sample of the Yelp reviews Dataset.

## Logistic Regression

| | Model Name | Train Accuracy | Validation Accuracy | Accuracy Difference | MCC (Validation) |
|---|---|---|---|---|---|
| **0** | Logistic Regression | 0.756715 | 0.756054 | 0.000661 | 0.629038 |

```
The best hyperparameters are:  {'C': 0.01, 'penalty': 'l2'}

Predicted values: [2 2 0 ... 0 2 2]
True values: [2 2 0 ... 0 2 2]
MCC (Validation): 0.62903752520419
```

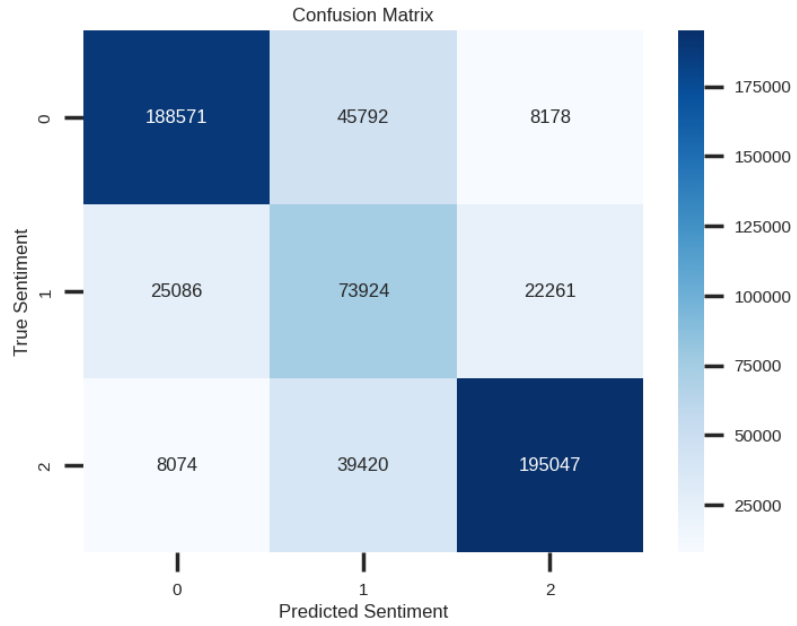
| | 0 | 1 | 2 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|---|
| **precision** | 0.852124 | 0.465828 | 0.866378 | 0.756054 | 0.728110 | 0.780680 |
| **recall** | 0.777710 | 0.612122 | 0.806230 | 0.756054 | 0.732021 | 0.756054 |
| **f1-score** | 0.813218 | 0.529048 | 0.835223 | 0.756054 | 0.725830 | 0.765272 |
| **support** | 169796.000000 | 84769.000000 | 169882.000000 | 0.756054 | 424447.000000 | 424447.000000 |

```
Test Set Metrics:
Accuracy: 0.7546
Matthews Correlation Coefficient (MCC): 0.6268
Classification Report:
                      0              1              2  accuracy
precision      0.850449       0.464533       0.865007   0.75458
recall         0.777481       0.609577       0.804182   0.75458
f1-score       0.812330       0.527262       0.833486   0.75458
support    242541.000000  121271.000000  242541.000000   0.75458

               macro avg   weighted avg
precision       0.726663       0.779089
recall          0.730413       0.754580
f1-score        0.724359       0.763779
support    606353.000000  606353.000000
```

| | 0 | 1 | 2 | accuracy | macro avg |
|---|---|---|---|---|---|
| **precision** | 0.850449 | 0.464533 | 0.865007 | 0.75458 | 0.726663 |
| **recall** | 0.777481 | 0.609577 | 0.804182 | 0.75458 | 0.730413 |
| **f1-score** | 0.812330 | 0.527262 | 0.833486 | 0.75458 | 0.724359 |

```
Confusion Matrix:
[[188571  45792   8178]
 [ 25086  73924  22261]
 [  8074  39420 195047]]
```




- **Train Accuracy:** 75.67%
- **Validation Accuracy:** 75.61%

- **Test Accuracy:** 75.46%
- **MCC (Validation):** 0.629
- **MCC (Test):** 0.627
- **Classification Report (Test):**
    - Positive (Class 2): Precision 86.5%, Recall 80.4%, F1-score 83.35%
    - Neutral (Class 1): Precision 46.45%, Recall 60.96%, F1-score 52.73%
    - Negative (Class 0): Precision 85.04%, Recall 77.75%, F1-score 81.23%

Logistic Regression performs well, with balanced precision and recall across positive and negative classes. It struggles with the neutral class, misclassifying many as either positive or negative. Minimal overfitting is evident due to a small accuracy difference between training and validation.

## Multinomial Naïve Bayes

| | Model Name | Train Accuracy | Validation Accuracy | Accuracy Difference | MCC (Validation) |
|---|---|---|---|---|---|
| **0** | Multinomial Naive Bayes | 0.712524 | 0.711448 | 0.001076 | 0.564875 |

```
The best hyperparameters are:  {'alpha': 1.0}

Predicted values: [2 0 0 ... 0 2 2]
True values: [2 2 0 ... 0 2 2]
MCC (Validation): 0.5648747838682471
```

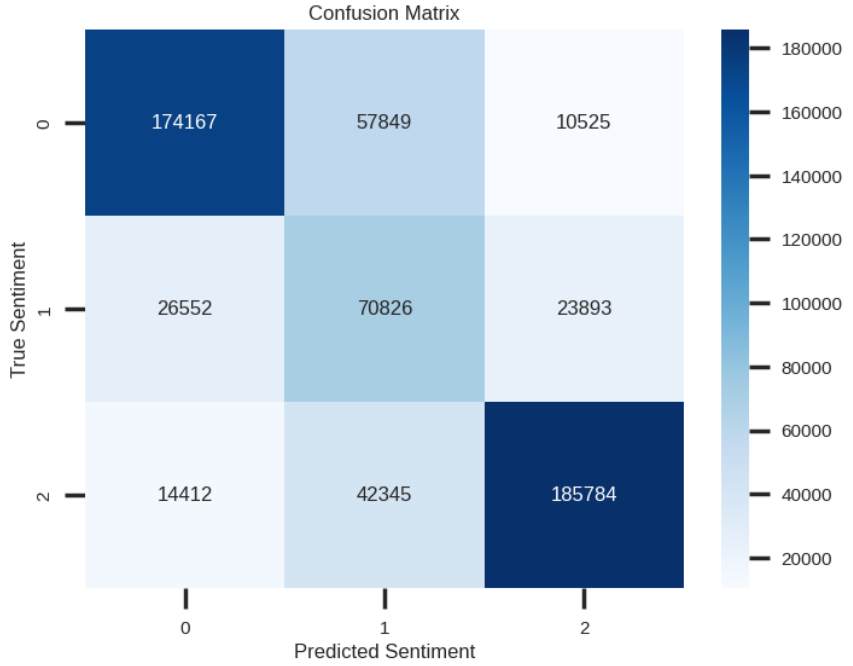
| | 0 | 1 | 2 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|---|
| **precision** | 0.810660 | 0.414397 | 0.845215 | 0.711448 | 0.690090 | 0.745350 |
| **recall** | 0.718315 | 0.585261 | 0.767550 | 0.711448 | 0.690375 | 0.711448 |
| **f1-score** | 0.761699 | 0.485227 | 0.804513 | 0.711448 | 0.683813 | 0.723619 |
| **support** | 169796.000000 | 84769.000000 | 169882.000000 | 0.711448 | 424447.000000 | 424447.000000 |

```
Test Set Metrics:
Accuracy: 0.7104
Matthews Correlation Coefficient (MCC): 0.5633
Classification Report:
                     0             1             2  accuracy
precision     0.809586      0.414139      0.843698  0.710439
recall        0.718093      0.584031      0.765990  0.710439
f1-score      0.761100      0.484627      0.802968  0.710439
support    242541.000000 121271.000000 242541.000000  0.710439

              macro avg   weighted avg
precision      0.689141       0.744141
recall         0.689371       0.710439
f1-score       0.682898       0.722552
support    606353.000000  606353.000000
```

| | 0 | 1 | 2 | accuracy | macro avg |
|---|---|---|---|---|---|
| **precision** | 0.809586 | 0.414139 | 0.843698 | 0.710439 | 0.689141 |
| **recall** | 0.718093 | 0.584031 | 0.765990 | 0.710439 | 0.689371 |
| **f1-score** | 0.761100 | 0.484627 | 0.802968 | 0.710439 | 0.682898 |

```
Confusion Matrix:
[[174167  57849  10525]
 [ 26552  70826  23893]
 [ 14412  42345 185784]]
```




- **Train Accuracy**: 71.25%
- **Validation Accuracy**: 71.14%

- **Test Accuracy**: 71.04%
- **MCC (Validation)**: 0.565
- **MCC (Test)**: 0.563
- **Classification Report (Test)**:
    - Positive (Class 2): Precision 84.37%, Recall 76.60%, F1-score 80.30%
    - Neutral (Class 1): Precision 41.41%, Recall 58.40%, F1-score 48.46%
    - Negative (Class 0): Precision 80.96%, Recall 71.81%, F1-score 76.11%

Multinomial Naïve Bayes shows strong performance for polar sentiments (positive and negative). Neutral sentiments remain a challenge, with low precision and recall. Excellent generalization due to minimal accuracy difference between training and validation.

## Decision Tree

| | Model Name | Train Accuracy | Validation Accuracy | Accuracy Difference | MCC (Validation) |
|---|---|---|---|---|---|
| **0** | Decision Tree | 0.596846 | 0.591808 | 0.005038 | 0.377757 |

```
The best hyperparameters are:  {'max_depth': 10, 'min_samples_leaf': 3}

Predicted values: [2 2 2 ... 0 0 2]
True values: [2 2 0 ... 0 2 2]
MCC (Validation): 0.3777574375107083
```

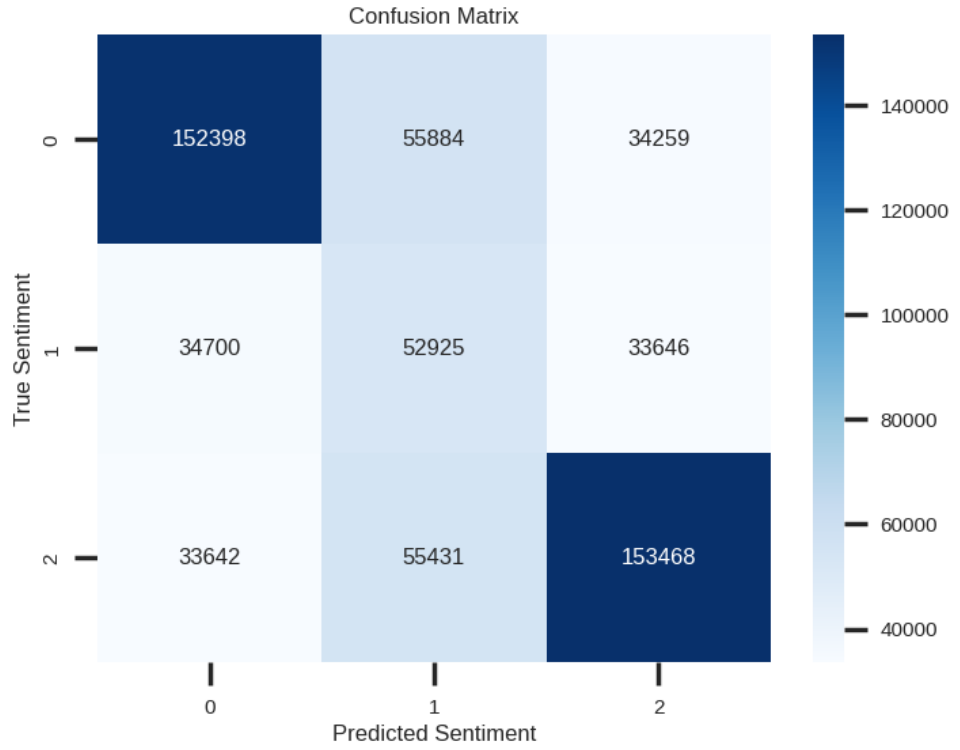
| | 0 | 1 | 2 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|---|
| **precision** | 0.689781 | 0.320820 | 0.693254 | 0.591808 | 0.567951 | 0.617483 |
| **recall** | 0.630262 | 0.432080 | 0.633075 | 0.591808 | 0.565139 | 0.591808 |
| **f1-score** | 0.658680 | 0.368229 | 0.661799 | 0.591808 | 0.562903 | 0.601920 |
| **support** | 169796.000000 | 84769.000000 | 169882.000000 | 0.591808 | 424447.000000 | 424447.000000 |

```
Test Set Metrics:
Accuracy: 0.5917
Matthews Correlation Coefficient (MCC): 0.3781
Classification Report:
                     0              1              2  accuracy
precision     0.690396       0.322242       0.693255   0.59172
recall        0.628339       0.436419       0.632751   0.59172
f1-score      0.657907       0.370739       0.661623   0.59172
support  242541.000000  121271.000000  242541.000000   0.59172

              macro avg   weighted avg
precision       0.568631       0.617909
recall          0.565836       0.591720
f1-score        0.563423       0.601960
support    606353.000000  606353.000000
```

| | 0 | 1 | 2 | accuracy | macro avg |
|---|---|---|---|---|---|
| **precision** | 0.690396 | 0.322242 | 0.693255 | 0.59172 | 0.568631 |
| **recall** | 0.628339 | 0.436419 | 0.632751 | 0.59172 | 0.565836 |
| **f1-score** | 0.657907 | 0.370739 | 0.661623 | 0.59172 | 0.563423 |

```
Confusion Matrix:
[[152398  55884  34259]
 [ 34700  52925  33646]
 [ 33642  55431 153468]]
```



- **Train Accuracy**: 59.68%
- **Validation Accuracy**: 59.18%
- **Test Accuracy**: 59.17%
- **MCC (Validation)**: 0.378
- **MCC (Test)**: 0.378
- **Classification Report (Test)**:
    - Positive (Class 2): Precision 69.33%, Recall 63.28%, F1-score 66.16%
    - Neutral (Class 1): Precision 32.22%, Recall 43.64%, F1-score 37.07%
    - Negative (Class 0): Precision 69.04%, Recall 62.83%, F1-score 65.79%

Decision Tree underperforms compared to other models, with lower precision, recall, and F1-scores. While it is interpretable and avoids overfitting, its inability to handle complex patterns limits its effectiveness. Neutral sentiments remain its weakest point, with the lowest scores among all classes.

## Random Forest

| | Model Name | Train Accuracy | Validation Accuracy | Accuracy Difference | MCC (Validation) |
|---|---|---|---|---|---|
| **0** | Random Forest | 0.722053 | 0.690256 | 0.031797 | 0.525738 |

```
The best hyperparameters are:  {'max_depth': 20, 'min_samples_leaf': 3}

Predicted values: [2 2 2 ... 0 2 2]
True values: [2 2 0 ... 0 2 2]
MCC (Validation): 0.5257381792145353
```

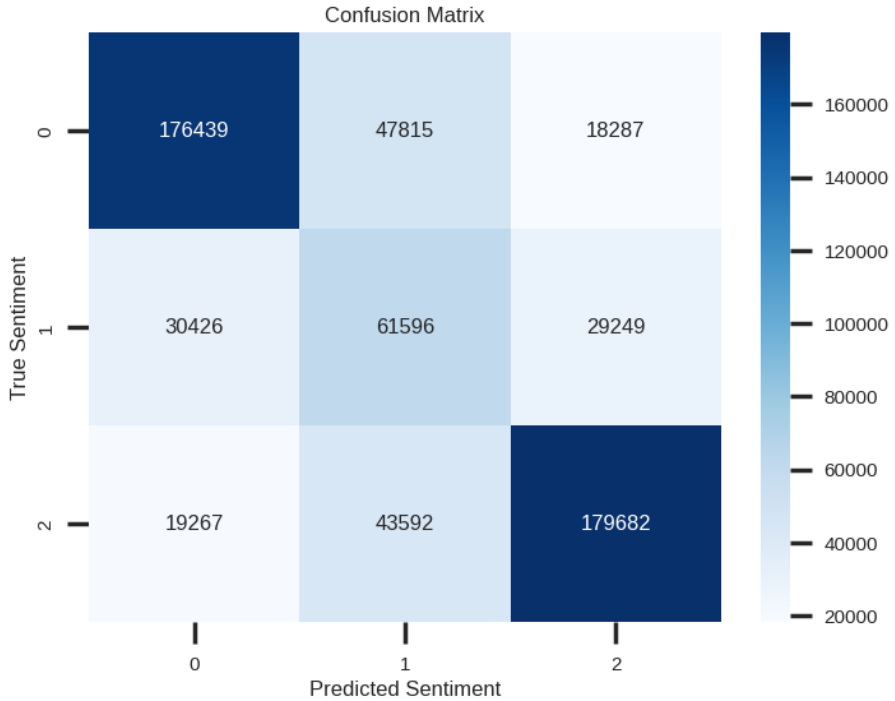
| | 0 | 1 | 2 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|---|
| **precision** | 0.781062 | 0.402663 | 0.794673 | 0.690256 | 0.659466 | 0.710937 |
| **recall** | 0.728457 | 0.510918 | 0.741562 | 0.690256 | 0.660312 | 0.690256 |
| **f1-score** | 0.753843 | 0.450376 | 0.767200 | 0.690256 | 0.657140 | 0.698582 |
| **support** | 169796.000000 | 84769.000000 | 169882.000000 | 0.690256 | 424447.000000 | 424447.000000 |

```
Test Set Metrics:
Accuracy: 0.6889
Matthews Correlation Coefficient (MCC): 0.5234
Classification Report:
                     0              1              2  accuracy
precision     0.780248       0.402580       0.790791  0.688901
recall        0.727461       0.507920       0.740831  0.688901
f1-score      0.752930       0.449157       0.764997  0.688901
support     242541.000000  121271.000000  242541.000000  0.688901

                macro avg   weighted avg
precision        0.657873       0.708931
recall           0.658737       0.688901
f1-score         0.655694       0.697002
support     606353.000000  606353.000000
```

| | 0 | 1 | 2 | accuracy | macro avg |
|---|---|---|---|---|---|
| **precision** | 0.780248 | 0.402580 | 0.790791 | 0.688901 | 0.657873 |
| **recall** | 0.727461 | 0.507920 | 0.740831 | 0.688901 | 0.658737 |
| **f1-score** | 0.752930 | 0.449157 | 0.764997 | 0.688901 | 0.655694 |

```
Confusion Matrix:
[[176439  47815  18287]
 [ 30426  61596  29249]
 [ 19267  43592 179682]]
```




- **Train Accuracy**: 72.21%

- **Validation Accuracy**: 69.03%
- **Test Accuracy**: 68.89%
- **MCC (Validation)**: 0.526
- **MCC (Test)**: 0.523
- **Classification Report (Test)**:
    - Positive (Class 2): Precision 79.08%, Recall 74.08%, F1-score 76.50%
    - Neutral (Class 1): Precision 40.26%, Recall 50.79%, F1-score 44.92%
    - Negative (Class 0): Precision 78.12%, Recall 72.75%, F1-score 75.29%

Random Forest performs better than Decision Tree but worse than Logistic Regression. Moderate overfitting suggests it is less robust for unseen data compared to simpler models. Neutral sentiment detection remains challenging, with significant misclassifications.

Between the 4 models implemented, logistic regression appears to have the best performance. With high accuracy (75.46%) and MCC (0.6268) indicating it effectively balances precision and recall across all classes, and strong performance for polar sentiments. All models struggle with neutral sentiments, with low precision and recall across the board, this can be highlighted as a key challenge. A possible improvement would be to use deep learning models like BERT for better context understanding, which I had planned to implement in this project. However, BERT requires more computational power than is available currently.