Sentiment Analysis of Twitter Data: A Comparative Study of Traditional and Deep Learning Approaches

[Veerendragouda.T.Patil]

December 9, 2024

1 Executive Summary

This study implements and compares three distinct approaches to sentiment analysis on Twitter data: Support Vector Machine (SVM), Bidirectional LSTM (BiLSTM), and DistilBERT. Each model demonstrates different strengths in classifying tweets into three sentiment categories: negative (-1), neutral (0), and positive (1).

2 Dataset Analysis and Preprocessing

2.1 Dataset Characteristics

The analysis was performed on a substantial dataset comprising 154,278 tweets, distributed across three sentiment categories:

- Positive (1.0): 68,228 tweets (44.2%)
- Neutral (0.0): 52,318 tweets (33.9%)
- Negative (-1.0): 33,732 tweets (21.9%)

2.2 Text Analysis Insights

Our preliminary analysis revealed several important patterns in the data:

- Average text length: 124 characters
- Average word count: 20 words
- Text length variation by sentiment:
  - Negative tweets: 148 characters (average)
  - Neutral tweets: 88 characters (average)
  - Positive tweets: 140 characters (average)

A notable observation is that users tend to write more extensively when expressing strong opinions, particularly negative ones.

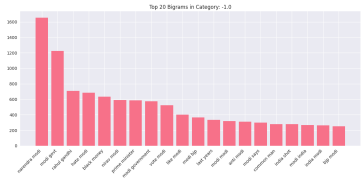

Figure 1: bigram frequency of category -1

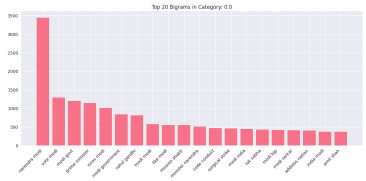

Figure 2: bigram frequency of category 0

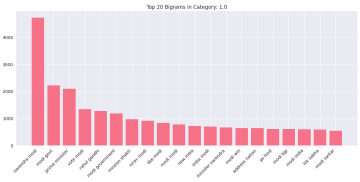

Figure 3: bigram frequency of category 1

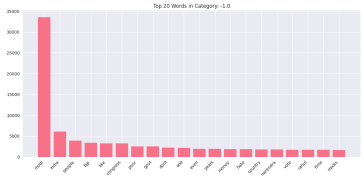

Figure 4: word frequency of category -1

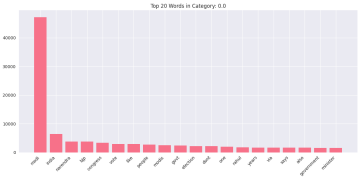

Figure 5: word frequency of category -1

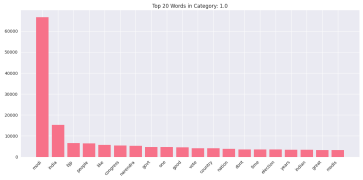

Figure 6: word frequency of category -1

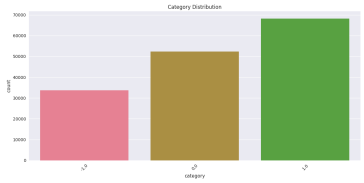

Figure 7: Category Distribution

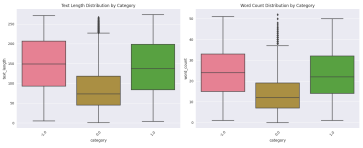

Figure 8: Lengh Distribution

2.3 Data Preprocessing Pipeline

We implemented a comprehensive preprocessing pipeline that included:

1. Text normalization (lowercase conversion)
2. URL and special character removal
3. Hashtag and mention cleaning
4. Contraction handling
5. Stop word removal
6. Lemmatization
7. Token standardization

3 Model Implementation and Results

3.1 Support Vector Machine (SVM)

The traditional machine learning approach using SVM demonstrated strong baseline performance:

3.1.1 Performance Metrics

- Accuracy: 88%
- Macro Average F1-score: 0.87
- Class-wise Performance:
  - Negative: F1-score 0.81
  - Neutral: F1-score 0.90
  - Positive: F1-score 0.90

3.1.2 Features

- TF-IDF vectorization with 5000 features
- Unigram and bigram consideration
- Linear kernel implementation

3.2 Bidirectional LSTM (BiLSTM)

The BiLSTM model showed significant improvement over the traditional approach:

3.2.1 Performance Metrics

- Final Test Accuracy: 96.93%
- Training Accuracy: 98.91%
- Validation Accuracy: 97.01%

3.2.2 Training Characteristics

- Rapid convergence (6 epochs)
- Consistent performance improvement
- Good generalization with minimal overfitting

3.3 DistilBERT

The transformer-based approach achieved the highest performance:

3.3.1 Performance Metrics

- Final Accuracy: 97%
- Macro Average F1-score: 0.97
- Class-wise Performance:
  - Negative: F1-score 0.95
  - Neutral: F1-score 0.99
  - Positive: F1-score 0.98

4 Comparative Analysis

4.1 Performance Comparison

Metric	SVM	BiLSTM	DistilBERT
Accuracy	88%	96.93%	97%
Training Time	Slow	Moderate	Slow
Resource Usage	Low	Moderate	High

Table 1: Model Performance Comparison

4.2 Trade-offs Analysis

4.2.1 Resource Requirements vs Performance

- SVM offers good performance with minimal resources
- BiLSTM provides excellent performance with moderate resources
- DistilBERT achieves best performance but requires significant computational power

4.2.2 Training Time vs Accuracy

- SVM: Slow training and lower accuracy
- BiLSTM: Moderate training time with high accuracy
- DistilBERT: Longest training time but highest accuracy

5 Conclusions and Recommendations

5.1 Choice of Model

Based on our analysis, we recommend:

- For quick deployment with limited resources: SVM
- For balanced performance and resources: BiLSTM
- For highest accuracy regardless of resources: DistilBERT

5.2 Best Practices

Our study highlighted several important practices:

- Thorough data preprocessing is crucial for all approaches
- Class imbalance handling improves performance
- Proper validation strategies are essential

5.3 Future Improvements

Future work could focus on:

- Ensemble methods combining multiple approaches
- Domain-specific fine-tuning for DistilBERT
- Advanced data augmentation techniques

6 Conclusion

This project demonstrates the evolution of sentiment analysis techniques from traditional machine learning to modern deep learning approaches. Each method offers unique advantages depending on the specific requirements of the application, with DistilBERT showing the highest accuracy but requiring the most resources, while SVM provides a solid baseline with minimal computational requirements.