Capstone Project Documentation

Sentiment Analysis of Movie Reviews Using Support Vector Machine

1. Problem Statement

Online movie review platforms contain a large volume of user-generated text expressing opinions about movies. These reviews significantly influence audience decisions and movie ratings. However, manually analyzing thousands of reviews to understand public sentiment is time-consuming and impractical.

Therefore, there is a need for an automated system that can classify movie reviews as **positive** or **negative** based on their textual content. This project addresses this problem by applying machine learning techniques to perform sentiment analysis on movie reviews.

2. Project Objective

The main objectives of this project are:

- To build a machine learning model that automatically classifies movie reviews into positive and negative sentiments
- To preprocess and convert textual data into numerical features using TF-IDF
- To train and evaluate a Support Vector Machine (SVM) classifier
- To compare the performance of different SVM kernels (Linear, Polynomial, RBF)
- To analyze results using standard evaluation metrics and visualization techniques

3. Dataset Description

The project uses the **Stanford Large Movie Review Dataset (IMDB)**, a widely used benchmark dataset for sentiment analysis tasks.

Dataset characteristics:

- Total reviews: 50,000
- Training set: 25,000 reviews
- Testing set: 25,000 reviews

- Balanced classes:
    - 12,500 positive reviews
    - 12,500 negative reviews

Each review is stored as a plain text file, and sentiment labels are inferred from the folder structure (`pos` for positive, `neg` for negative).

Dataset Source:
http://ai.stanford.edu/~amaas/data/sentiment/

4. Tools and Environment

- **Platform:** Google Colab
- **Programming Language:** Python
- **Libraries Used:**
    - NumPy
    - Pandas
    - Scikit-learn
    - Matplotlib
    - Seaborn

All experiments were conducted in a Google Colab notebook, ensuring reproducibility and ease of execution.

5. Approach and Methodology

The project follows a supervised machine learning pipeline consisting of the following steps:

5.1 Data Loading

Movie reviews were loaded from structured directories containing positive and negative reviews for both training and testing datasets. Each text file represents one review.

5.2 Text Preprocessing

Basic preprocessing steps were applied during vectorization:

- Conversion to lowercase
- Removal of stopwords
- Tokenization

These steps help reduce noise and improve model performance.

5.3 Feature Extraction (TF-IDF)

Text data was converted into numerical form using **Term Frequency–Inverse Document Frequency (TF-IDF)** vectorization. TF-IDF assigns higher importance to informative words while reducing the influence of frequently occurring but less meaningful words.

5.4 Model Training

Support Vector Machine (SVM) classifiers were trained using three different kernels:

- **Linear Kernel**
- **Polynomial Kernel**
- **Radial Basis Function (RBF) Kernel**

This allowed a comparative analysis of linear and non-linear decision boundaries for text classification.

5.5 Model Evaluation

Models were evaluated on the test dataset using:

- Accuracy
- Precision
- Recall
- F1-score

A confusion matrix was also used to visually analyze classification performance.

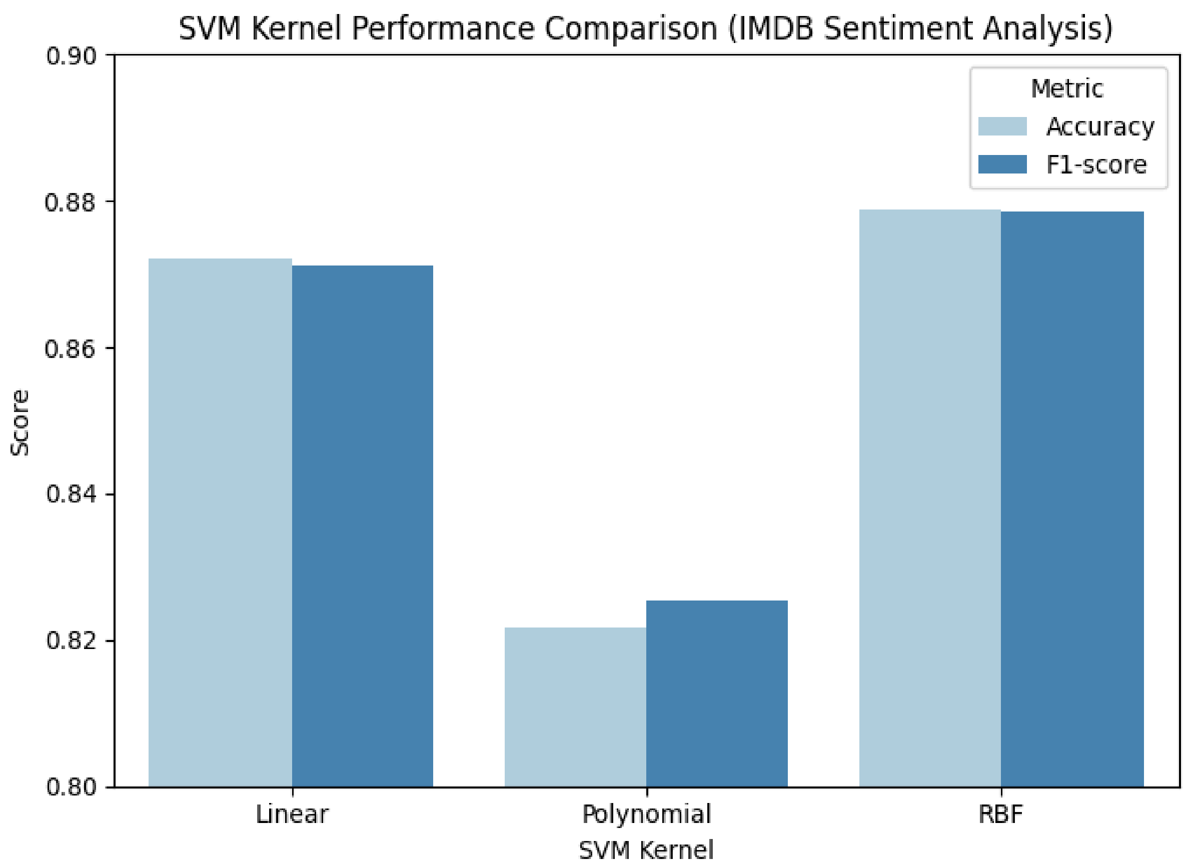
6. Results and Analysis

6.1 Kernel Performance Comparison

Kernel	Accuracy	F1-score
Linear	~87%	~0.8711
Polynomial	~82%	~0.8254
RBF	~88%	~0.8785

The **Radial Basis Function (RBF)** achieved the best overall performance.

- **Observation:**
  The RBF kernel achieved the highest accuracy and F1-score among the evaluated models, indicating its ability to capture non-linear sentiment patterns in the data. The improvement over the linear kernel is modest, suggesting that both kernels are effective, with RBF offering slightly better performance at higher computational cost.

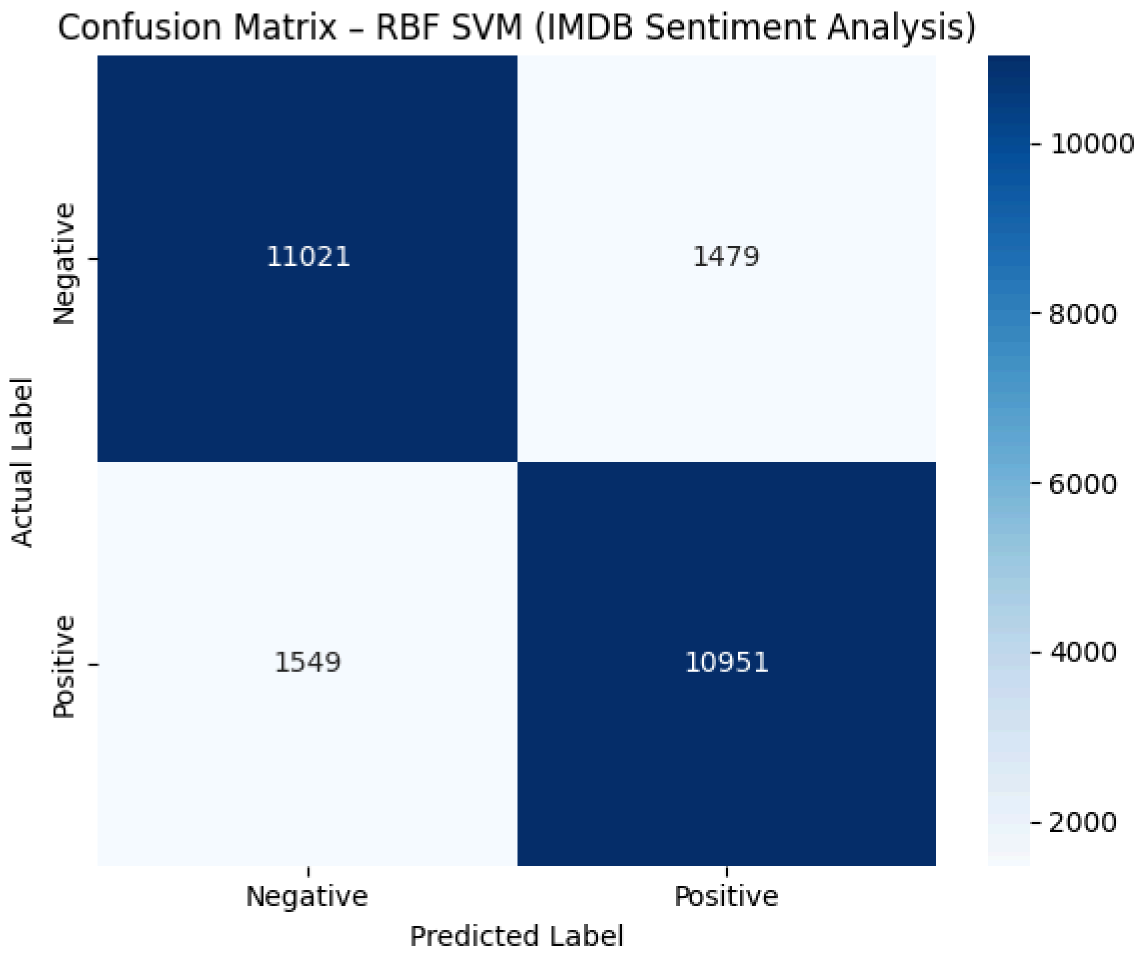


6.2 Confusion Matrix Analysis

The confusion matrix for the **RBF SVM**, which achieved the highest accuracy, shows improved performance across both positive and negative classes. Most predictions lie along the diagonal, indicating correct classification.

Remaining misclassifications mainly occur in reviews containing:

- Mixed sentiments
- Implicit opinions
- Sarcasm



7. Key Insights

- The RBF kernel provided the best overall performance for sentiment classification.
- Confusion matrix analysis confirms balanced and accurate predictions for both classes.
- Non-linear modeling helps capture complex sentiment patterns more effectively.

8. Limitations

- The model cannot effectively handle sarcasm or implicit sentiment.
- Contextual understanding is limited due to bag-of-words based feature representation.
- Performance may vary when applied to domains other than movie reviews.

9. Future Scope

- Use word embeddings or transformer-based models for better context understanding.
- Extend the model to multi-class or aspect-based sentiment analysis.
- Apply the approach to real-time review or social media data.

10. Conclusion

This project demonstrates the application of Support Vector Machines for sentiment analysis of movie reviews using TF-IDF features. Among the evaluated kernels, the RBF kernel achieved the best performance by effectively modeling non-linear sentiment patterns. The results show that classical machine learning techniques can deliver strong performance for text classification tasks when combined with appropriate feature engineering.