Sentiment Analysis of Movie Reviews

Team number 43

2022170410	مريم تامر محمد محمد محمد القندقلي
2022170136	حنين هانى عبد المنعم
2022170149	رحمه محمد خطاب حامد
2022170446	منه محمد مصطفي احمد
2022170147	رحاب خالد عبدالحميد البهنسى
2022170503	يوستينا متشو سمير شكرى

Overview

This project applies Natural Language Processing (NLP) and Machine Learning techniques to classify movie reviews as **positive** or **negative**. The workflow includes cleaning and preprocessing text, transforming it into numerical features using TF-IDF, training various models, and evaluating their performance.

The goal is to build an accurate sentiment classifier and visualize the model performance using suitable charts.

Project Requirements

This project meets the five essential requirements:

1. **Data Preprocessing**
    - Clean text (remove HTML tags, punctuation, lowercase conversion).
    - Tokenization using spaCy.
    - Stopwords removal using nltk.
    - Data augmentation and lemmatization.
2. **Feature Extraction**
    - Applied **TF-IDF vectorization** using sklearn.
    - Vocabulary size: 41,214 features.
3. **Model Training and Testing**
    - Trained multiple models using an 80-20 train-test split (3200 train, 800 test).
    - Used 8 Machine Learning models and 1 Deep Learning model.
    - Evaluated using accuracy & confusion matrix.
4. **Visualizing Results**
    - Confusion matrices, bar chart, PCA, ROC, TSNE, heat map and word cloud
5. **Saving the Model**
    - All ML models were saved using joblib including our best performing models.

Dataset Description

After preprocessing and augmentation, the dataset contains:

- df: 2,000 reviews (original, 2 columns)
- augmented_df: 4,000 reviews (augmented, 3 columns: label, text, tokens)

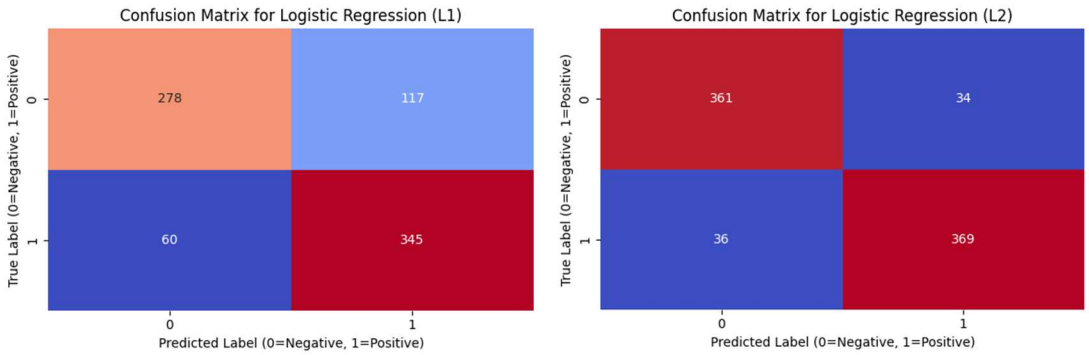
Data Summary

- **Original Dataset Dimensions**: (2000, 2)
- **Augmented Dataset Dimensions**: (4000, 3)
- **TF-IDF Matrix Shapes**:
  - Train: (3200, 41214)
  - Test: (800, 41214)
  - Full: (4000, 41214)

Models and Results

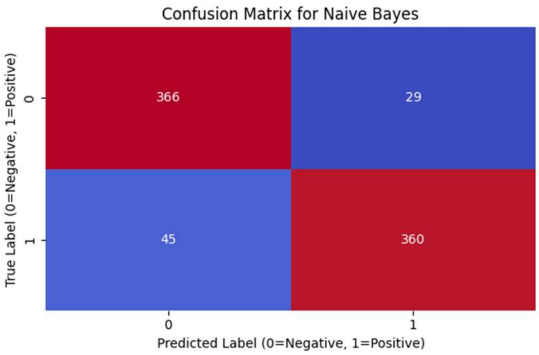
1. Logistic Regression

Model Version	Train Accuracy	Test Accuracy
L1 Penalty	79.59%	76.50%
L2 Penalty	97.59%	91.75%



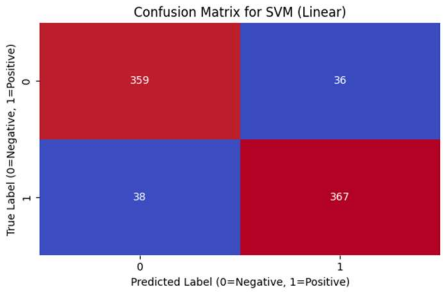
2. Naive Bayes

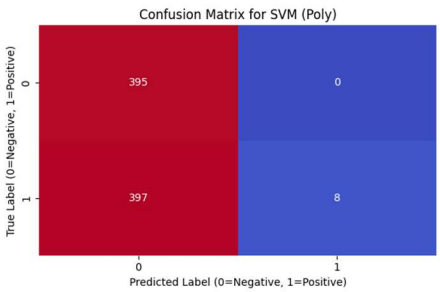
Train Accuracy	Test Accuracy
97.03%	91.75%

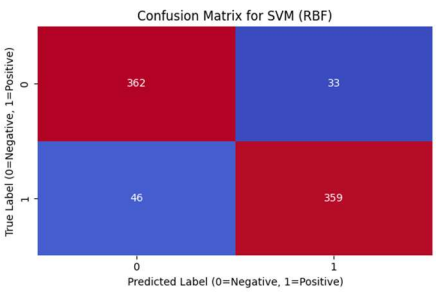


3. Support Vector Machine (SVM)

Kernel	Train Accuracy	Test Accuracy
Linear	96.38%	90.25%
RBF	98.41%	92.50%
Polynomial (deg=4)	83.84%	50.38%

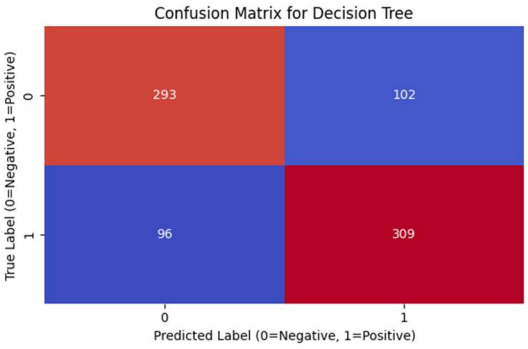






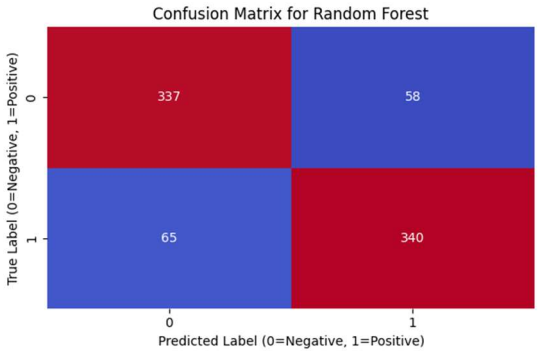
4. Decision Tree

Train Accuracy	Test Accuracy
88.38%	71.25%



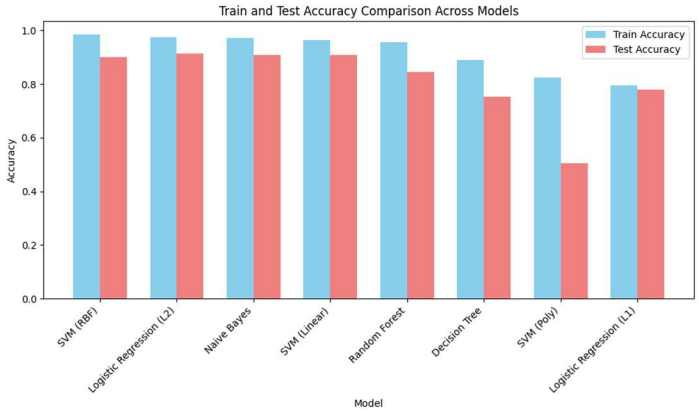
5. Random Forest

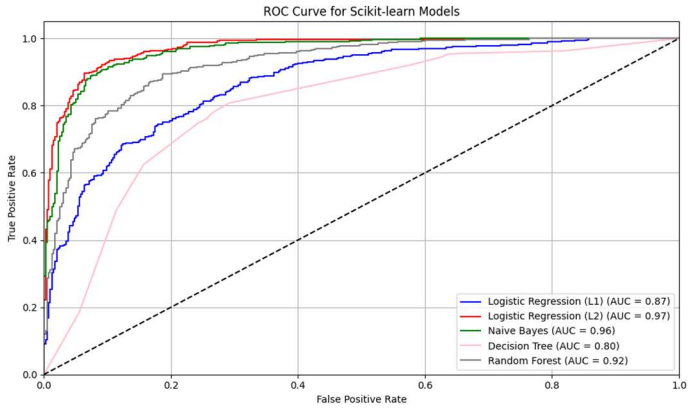
Train Accuracy	Test Accuracy
95.19%	85.88%

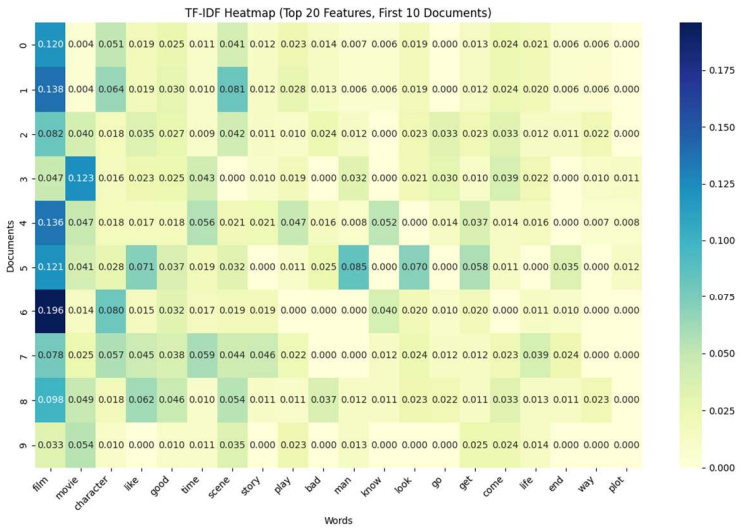


Best Model

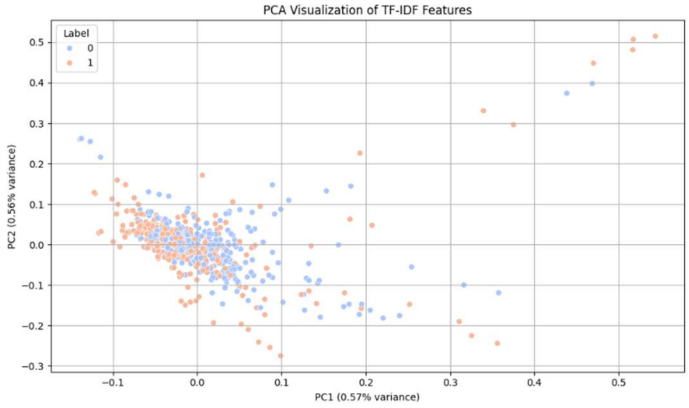
- **SVM with RBF Kernel** performed best overall:
  - **Train Accuracy**: 98.41%
  - **Test Accuracy**: 92.50%
- The model generalizes well and avoids overfitting, unlike Polynomial SVM which suffered from significant overfitting.

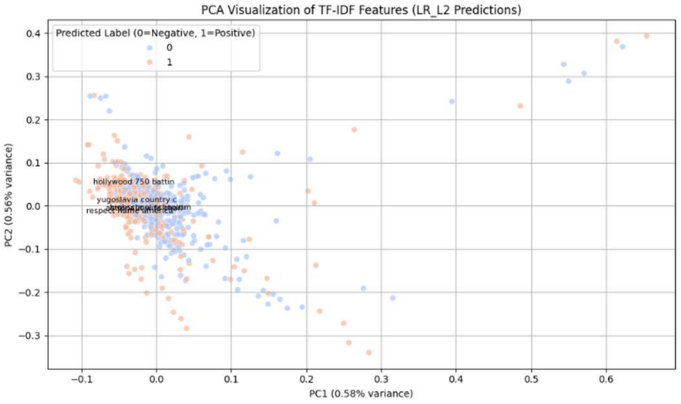


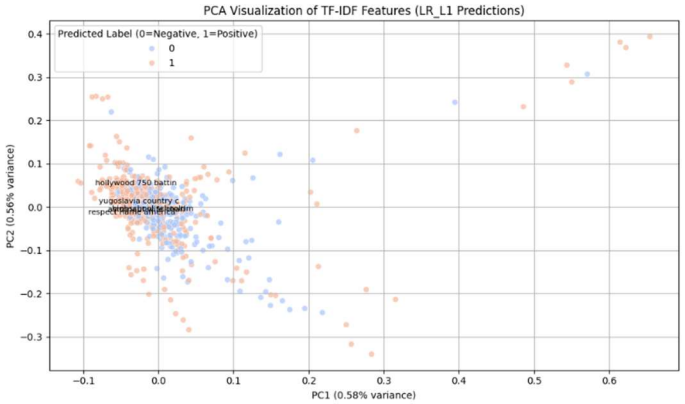


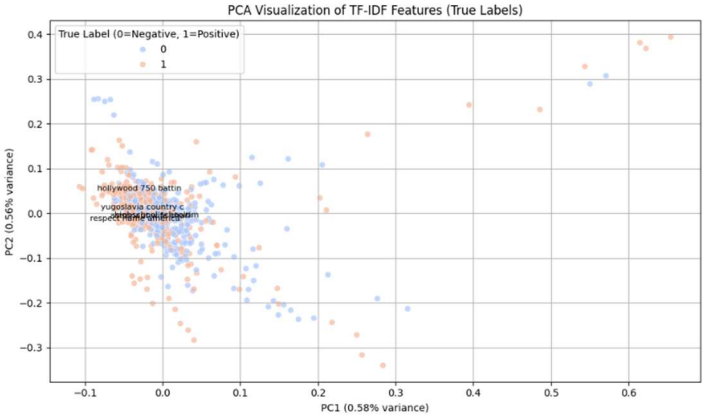


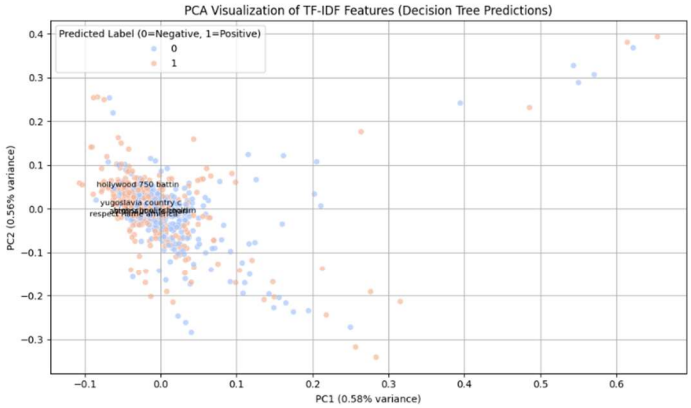
Visualizing Results

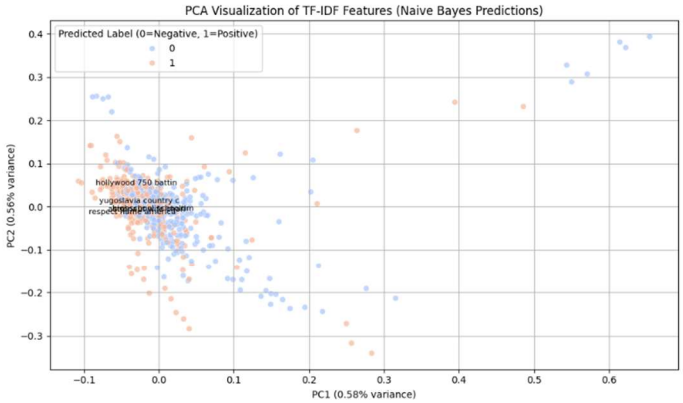


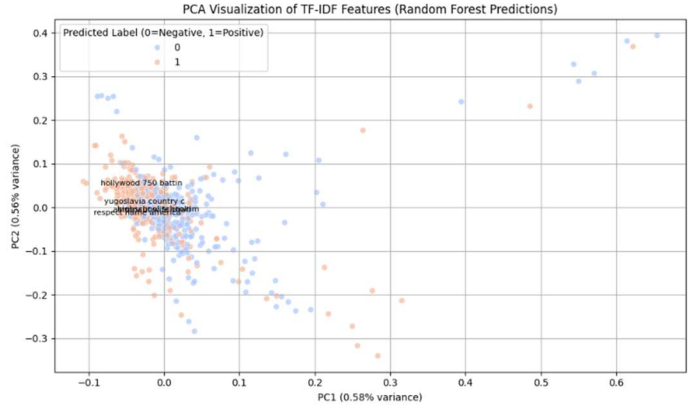


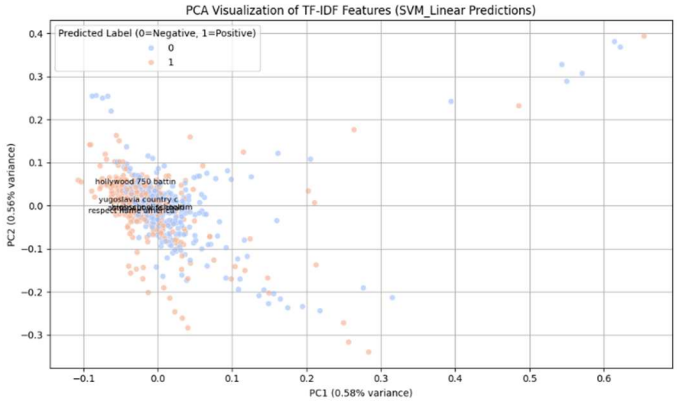


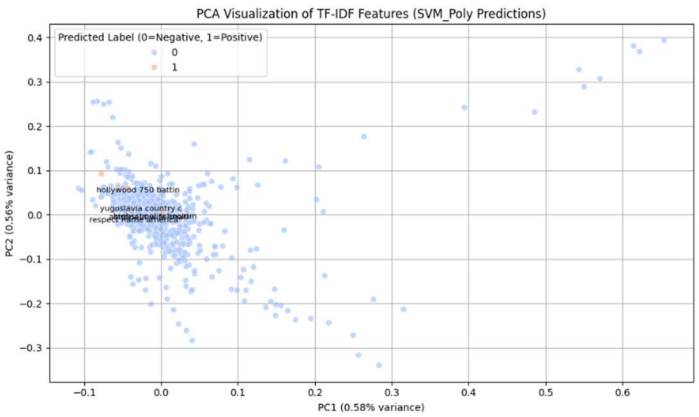


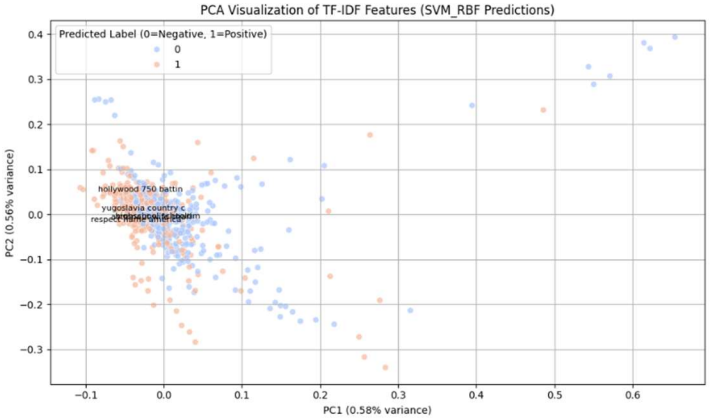


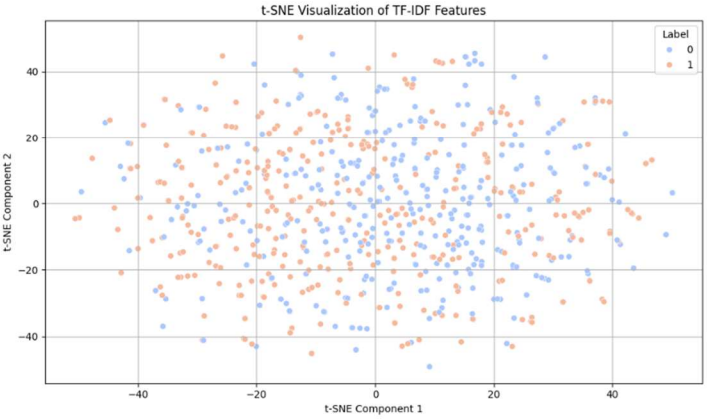


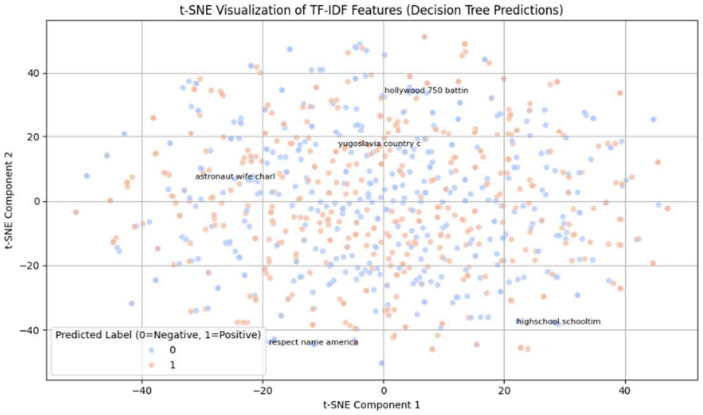


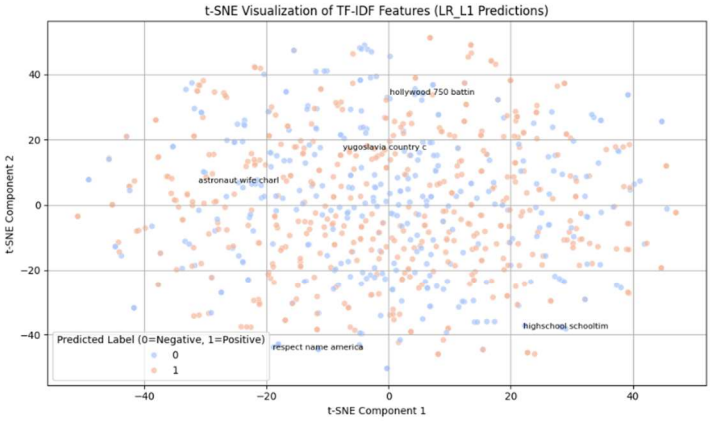


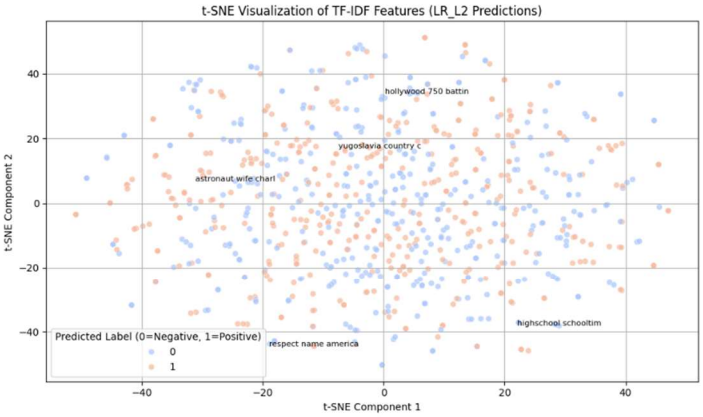

PCA Visualization of TF-IDF Features (SVM_Poly Predictions)
Predicted Label (0=Negative, 1=Positive)
0
1
hollywood 750 battin
yugoslavia country c
PC1 (0.58% variance)
PC2 (0.56% variance)

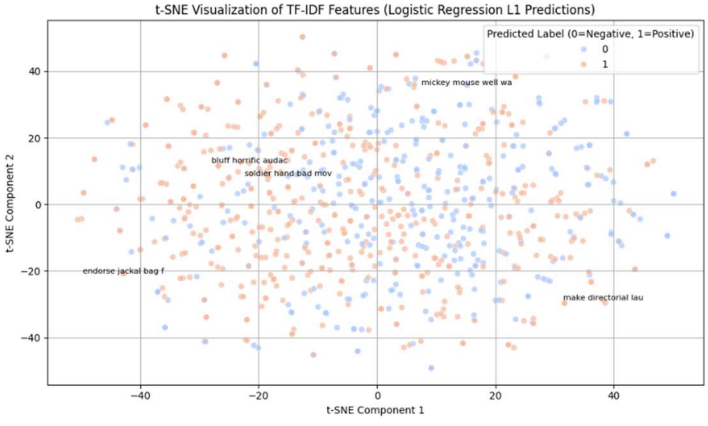

PCA Visualization of TF-IDF Features (SVM_RBF Predictions)
Predicted Label (0=Negative, 1=Positive)
0
1
hollywood 750 battin
yugoslavia country c
PC1 (0.58% variance)
PC2 (0.56% variance)

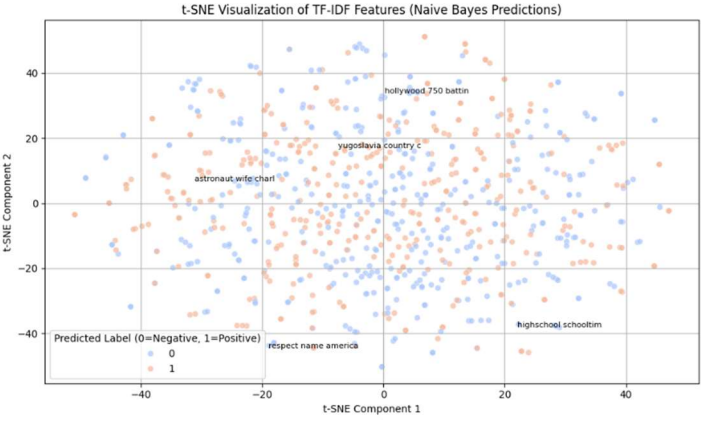

t-SNE Visualization of TF-IDF Features
Label
0
1
t-SNE Component 1
t-SNE Component 2

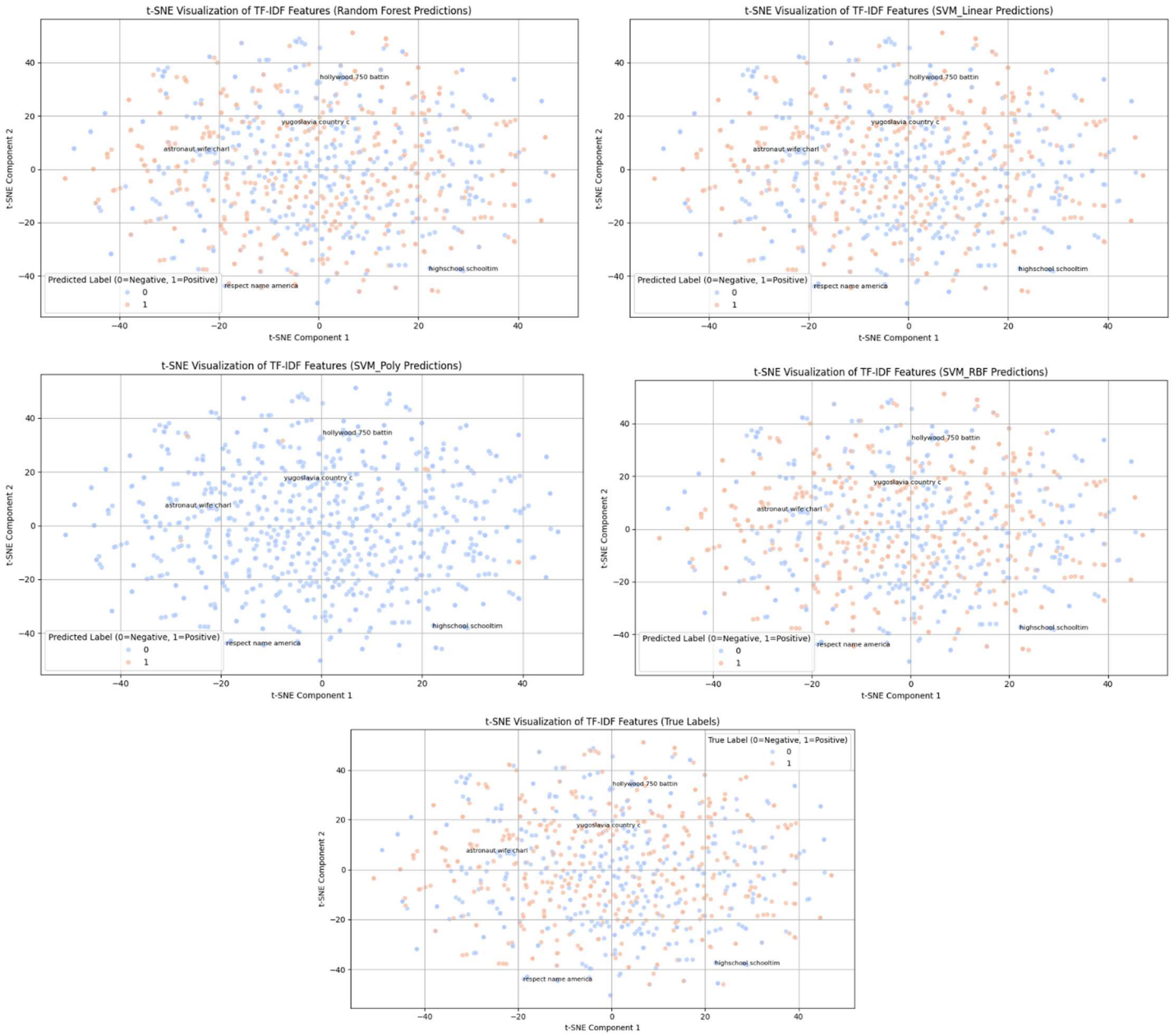

t-SNE Visualization of TF-IDF Features (Decision Tree Predictions)
hollywood 750 battin
yugoslavia country c
astronaut wife charl
highschool schooltim
respect name america
Predicted Label (0=Negative, 1=Positive)
0
1
t-SNE Component 1
t-SNE Component 2


t-SNE Visualization of TF-IDF Features (LR_L1 Predictions)
hollywood 750 battin
yugoslavia country c
astronaut wife charl
highschool schooltim
respect name america
Predicted Label (0=Negative, 1=Positive)
0
1
t-SNE Component 1
t-SNE Component 2


t-SNE Visualization of TF-IDF Features (LR_L2 Predictions)
hollywood 750 battin
yugoslavia country c
astronaut wife charl
highschool schooltim
respect name america
Predicted Label (0=Negative, 1=Positive)
0
1
t-SNE Component 1
t-SNE Component 2


t-SNE Visualization of TF-IDF Features (Logistic Regression L1 Predictions)
Predicted Label (0=Negative, 1=Positive)
0
1
mickey mouse well wa
bluff horrific audac
soldier hand bad mov
endorse jackal bag f
make directorial lau
t-SNE Component 1
t-SNE Component 2


t-SNE Visualization of TF-IDF Features (Naive Bayes Predictions)
hollywood 750 battin
yugoslavia country c
astronaut wife charl
highschool schooltim
respect name america
Predicted Label (0=Negative, 1=Positive)
0
1
t-SNE Component 1
t-SNE Component 2



Conclusion

This project successfully implemented a sentiment analysis pipeline that processes raw movie reviews and classifies them with high accuracy. Among all tested models, **SVM with RBF kernel** showed the best balance of performance and generalization.