CSP571 - Data Preparation and Analysis

Instructor - Oleksandr Narykov

TWITTER SENTIMENT ANALYSIS

Group No - 5

Pushkar Visave - A20582714

Saurabh Dighe - A20568301

Vibhav Rane - A20548428

TABLE OF CONTENTS

ABSTRACT

This report presents a comprehensive analysis of Twitter sentiment classification using machine learning approaches. The study implements and compares multiple classification models, including Naive Bayes and Logistic Regression, on a large-scale Twitter dataset containing 1.6 million tweets. Through extensive feature engineering, dimensionality reduction, and model optimization, we achieved significant classification accuracy in distinguishing between positive and negative sentiments. The research demonstrates the effectiveness of TF-IDF vectorization combined with feature selection in improving model performance.

INTRODUCTION

Social media sentiment analysis has become increasingly important for understanding public opinion, customer feedback, and market trends. Twitter, with its vast user base and real-time nature, provides a rich source of data for sentiment analysis. This study focuses on developing and evaluating machine learning models for accurate sentiment classification of tweets.

PROBLEM STATEMENT

The primary challenge addressed in this research is the accurate classification of tweet sentiments in a large-scale dataset. Specific challenges include:

- Handling and processing a large dataset of 1.6 million tweets
- Dealing with high-dimensional text data
- Balancing model complexity with performance
- Selecting optimal features for sentiment classification
- Achieving high classification accuracy while maintaining computational efficiency

PROPOSED METHODOLOGY

Dataset Overview

The study utilizes the Sentiment140 dataset from Kaggle, This dataset contains 1.6 million tweets and has the following characteristics:

- Binary classification (0 for negative, 4 for positive)
- Features include: target (sentiment), id, date, flag, user, and text
- Balanced class distribution between positive and negative sentiments

The dataset was constructed using distant supervision, where tweets were automatically labeled based on the presence of emoticons - tweets with positive emoticons were labeled as 4 (positive) and those with negative emoticons as 0 (negative), providing a large-scale collection for sentiment analysis research.

Data Preprocessing

The preprocessing pipeline includes:

1. Text Cleaning and normalization
    a. Conversion to lowercase
    b. Removal of special characters and numbers
    c. Handling of contractions
    d. Removal of URLs and emails
2. Advanced preprocessing techniques
    a. Custom feature creation

    b. Text length calculation
    c. Unique words ratio computation

Feature Extraction

The feature extraction process employs:

1. TF-IDF Vectorization
    a. Maximum 5000 features
    b. English stop words removal
    c. Custom parameters for document frequency
2. Additional Features
    a. Text length metrics
    b. Unique word ratios
    c. Custom engineered features

Dimensionality Reduction

Multiple dimensionality reduction techniques were implemented:

1. Principal Component Analysis (PCA)
2. Truncated SVD
3. t-SNE
4. UMAP

The analysis showed that:

- SVD achieved optimal performance for initial dimensionality reduction
- t-SNE and UMAP provided better visualization of the feature space
- Optimal number of components was determined through variance analysis

Model Training

The following models were implemented and trained:

1. Baseline Model: Multinomial Naive Bayes
2. Advanced Model: Logistic Regression with regularization
3. Optimized Model: Feature-selected Logistic Regression

Feature Selection

Feature selection was performed using:

1. Chi-squared test
2. SelectKBest method with k=3000 features
3. Correlation analysis for feature importance

ANALYSIS AND RESULT

Performance Metrics

The models were evaluated using:

1. Classification Accuracy
    a. Naive Bayes: 75.0%

Classification Report:				
	precision	recall	f1-score	support
0	0.75	0.76	0.75	159494
1	0.76	0.74	0.75	160506
accuracy			0.75	320000
macro avg	0.75	0.75	0.75	320000
weighted avg	0.75	0.75	0.75	320000

b. Logistic Regression: 77.0%

Classification Report:				
	precision	recall	f1-score	support
0	0.78	0.74	0.76	159494
1	0.75	0.79	0.77	160506
accuracy			0.77	320000
macro avg	0.77	0.77	0.77	320000
weighted avg	0.77	0.77	0.77	320000

c. Feature-selected LR: 76.0%

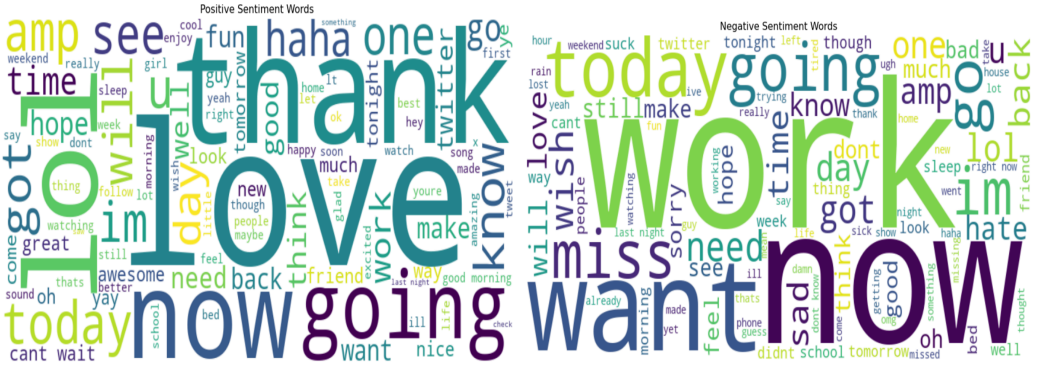
Classification Report:				
	precision	recall	f1-score	support
0	0.78	0.73	0.76	159494
1	0.75	0.80	0.77	160506
accuracy			0.76	320000
macro avg	0.77	0.76	0.76	320000
weighted avg	0.77	0.76	0.76	320000

2. ROC - AUC Scores
    a. Naive Bayes: 0.84
    b. Logistic Regression: 0.85
    c. Feature-selected LR: 0.85
3. Cross - Validation Results:
    a. Mean CV scores showed consistent performance across folds
    b. Standard deviation within acceptable range

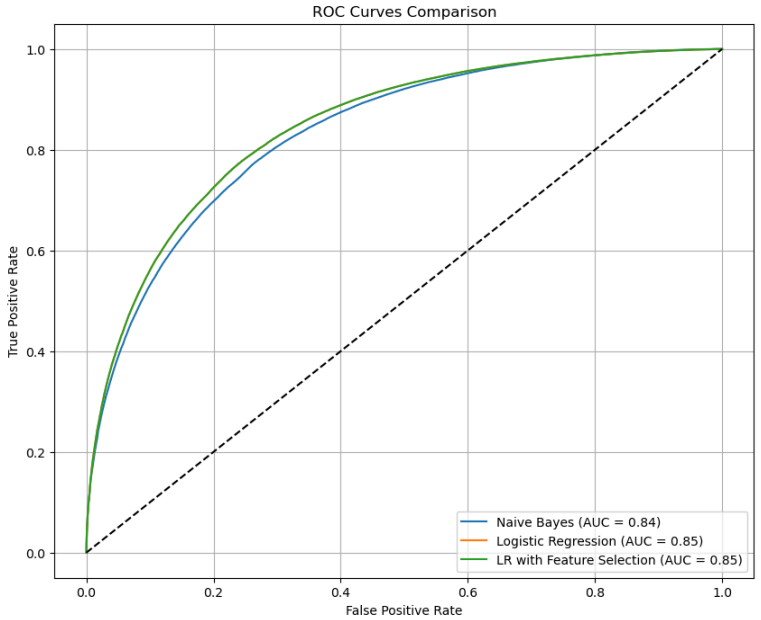
Visualization

Key Visualization included:

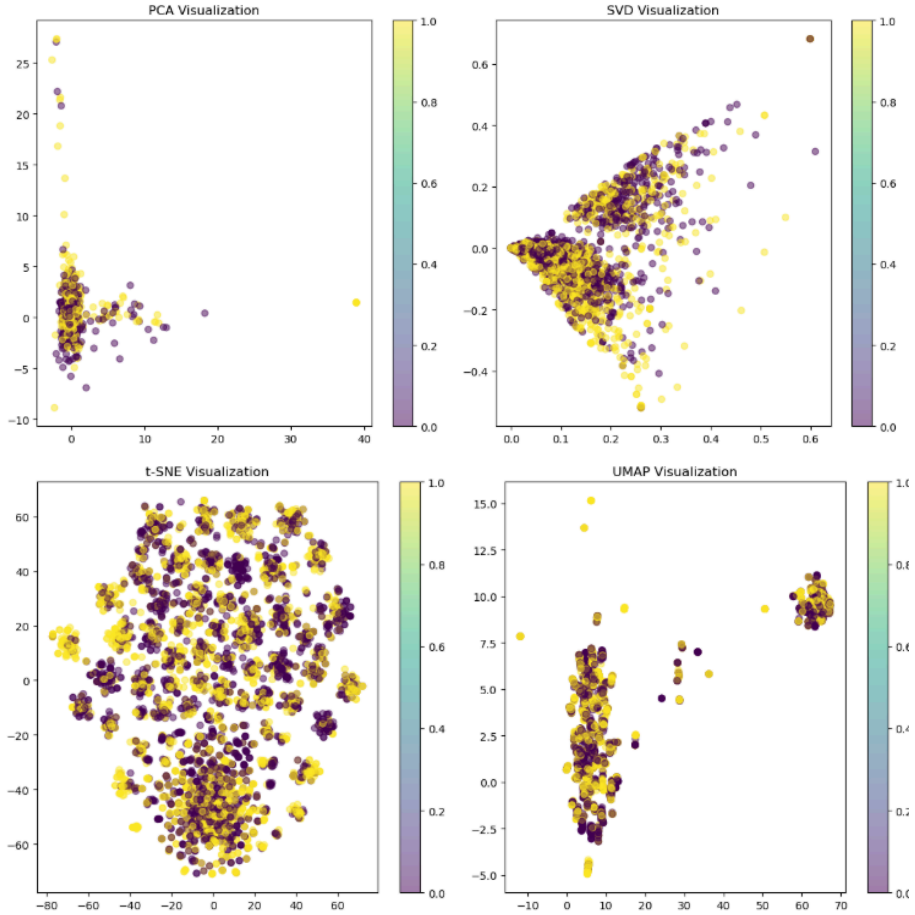
1. Word clouds for positive and negative sentiments



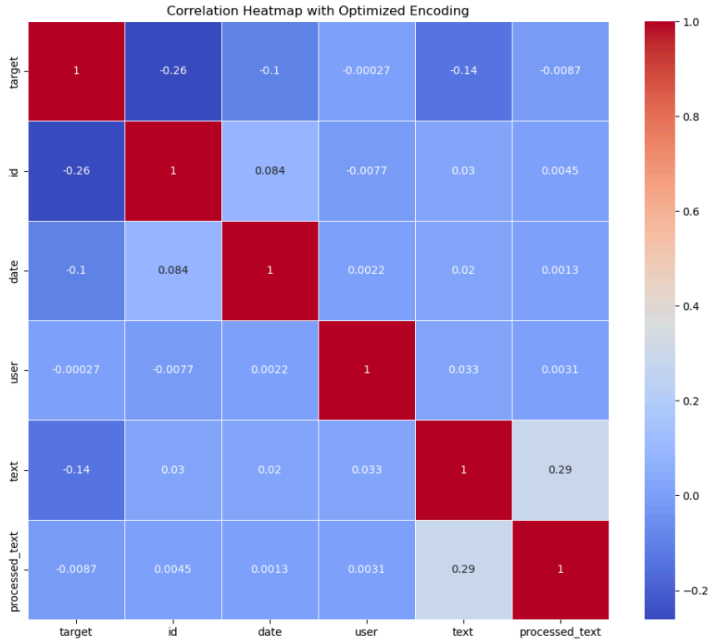
2. ROC curves comparison



3. Clustering visualizations using dimensionality reduction



4. Correlation heatmaps

5. 

CONCLUSION

The study successfully demonstrated the effectiveness of machine learning approaches in Twitter sentiment analysis. Key findings include:

- Feature selection significantly improved model performance
- Logistic Regression outperformed Naive Bayes
- Dimensionality reduction maintained performance while improving efficiency
- The importance of proper text preprocessing and feature engineering

FUTURE WORK

Potential areas for future research include:

- Implementation of deep learning models
- Exploration of multi-class sentiment classification
- Real-time sentiment analysis capabilities
- Integration of emotion detection
- Investigation of multilingual sentiment analysis

SOURCE CODE

The complete source code for this project is available on GitHub:

https://github.com/vibhav22022000/CSP571FinalProject2024.git