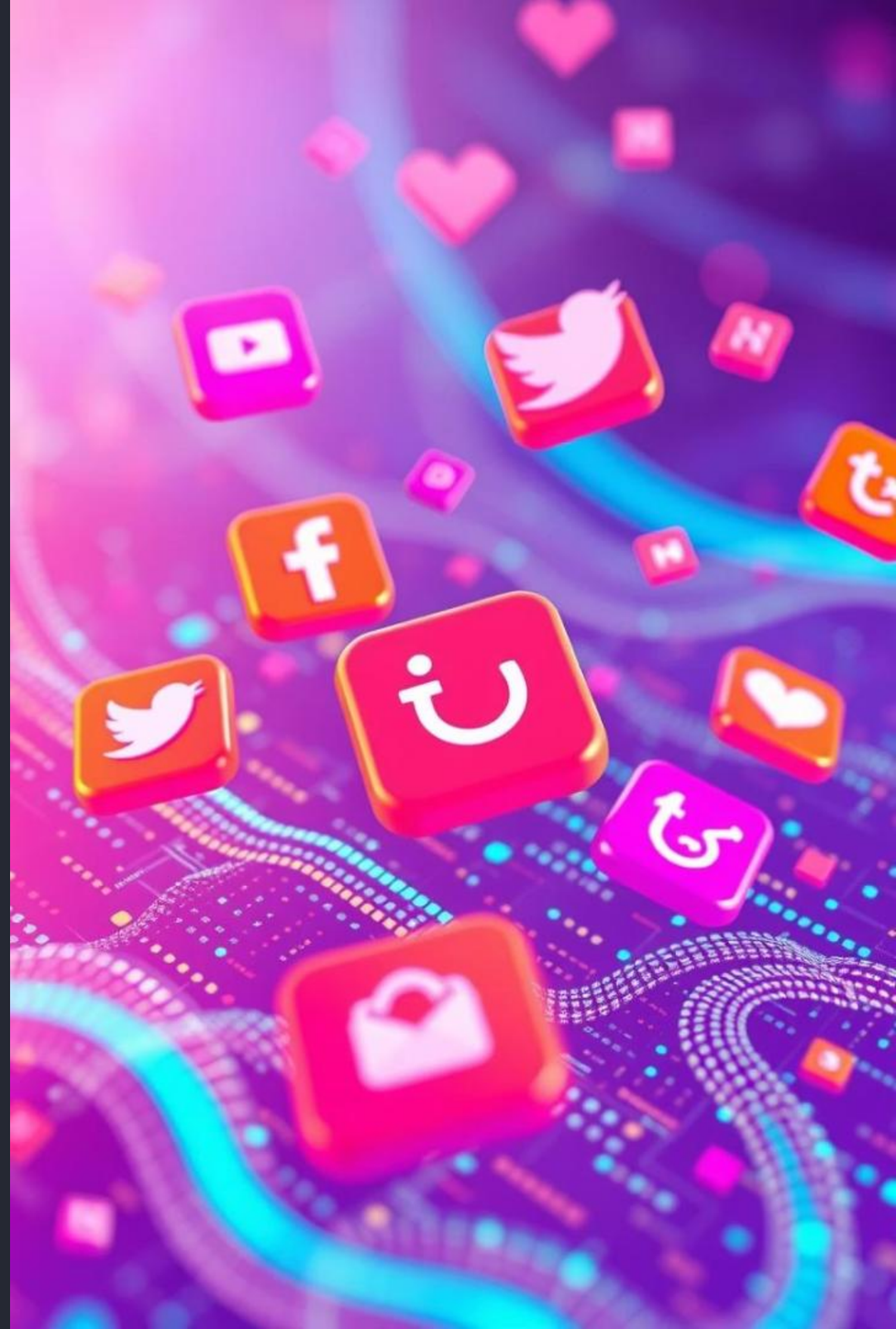
Social Media Sentiment Analysis System Using Ensemble Learning

This project develops a robust system to analyze sentiments in social media posts.



Project Overview

Objective

Analyze sentiments from Facebook and Twitter posts.

Approach

Use ensemble learning combining multiple ML and DL models.

Scope & Tools

Classify sentiments into four categories using Python and TensorFlow.

Dataset Description

Source & Structure

Facebook and Twitter datasets from Kaggle with train, validation, test sets.

Columns: text and sentiment labels.

Preprocessing & Size

Cleaned text by removing URLs, mentions, hashtags, and lowercased.

Test set has about 11,353 balanced samples.

Methodology

Data Preprocessing

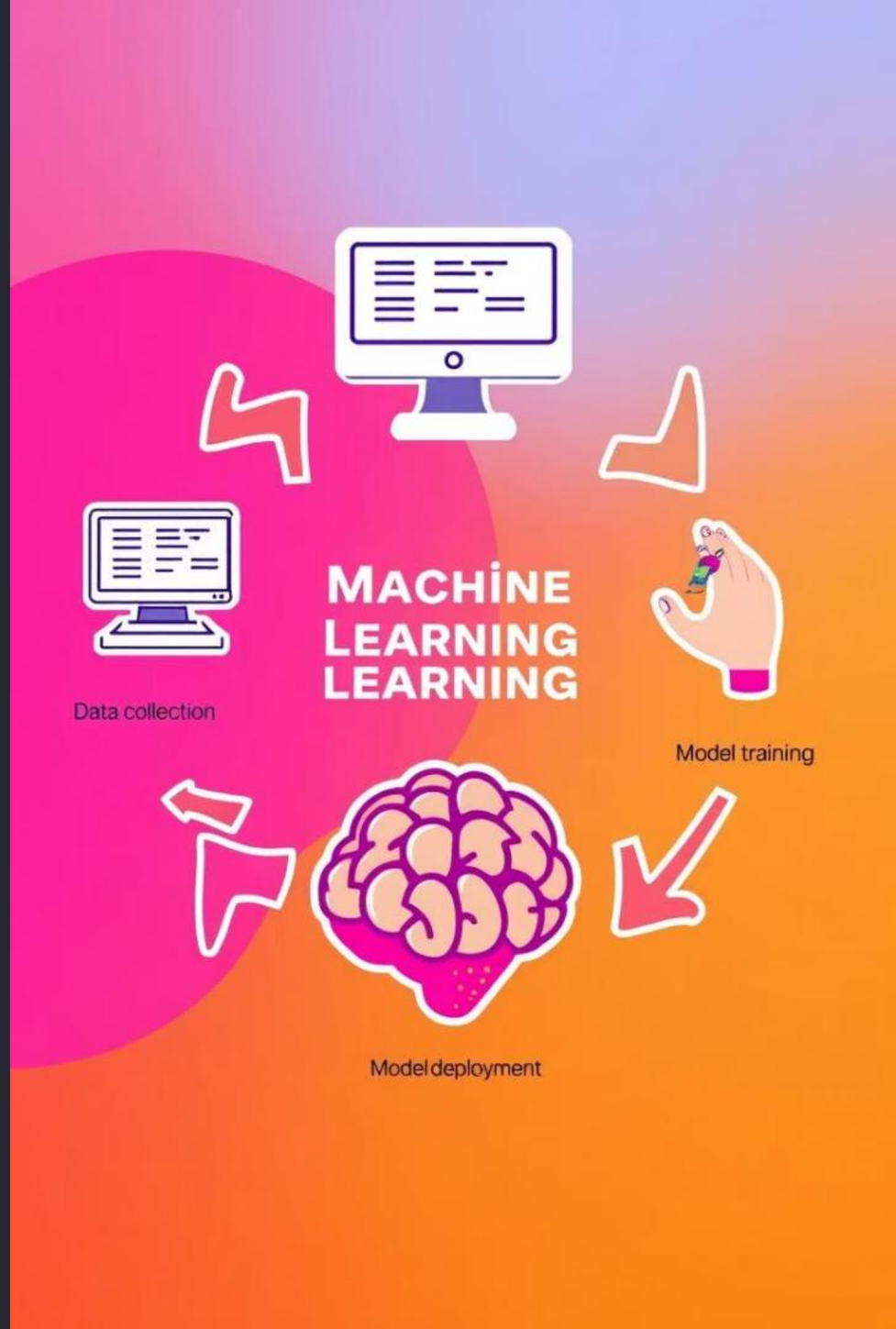
Regex cleaning, TF-IDF for ML, tokenization and padding for LSTM.

Models

- Naive Bayes, Logistic Regression, SVM, KNN, XGBoost
- Bidirectional LSTM
- Ensemble: Hard Voting, Soft Voting, Stacking

Evaluation

Measured by Accuracy and Macro F1 Score.



Model Architecture

Traditional ML

- Naive Bayes, Logistic Regression, SVM, KNN, XGBoost
- Balanced weights and tuned parameters

LSTM

Embedding layer, Bidirectional LSTM with dropout, dense layers.

Ensemble

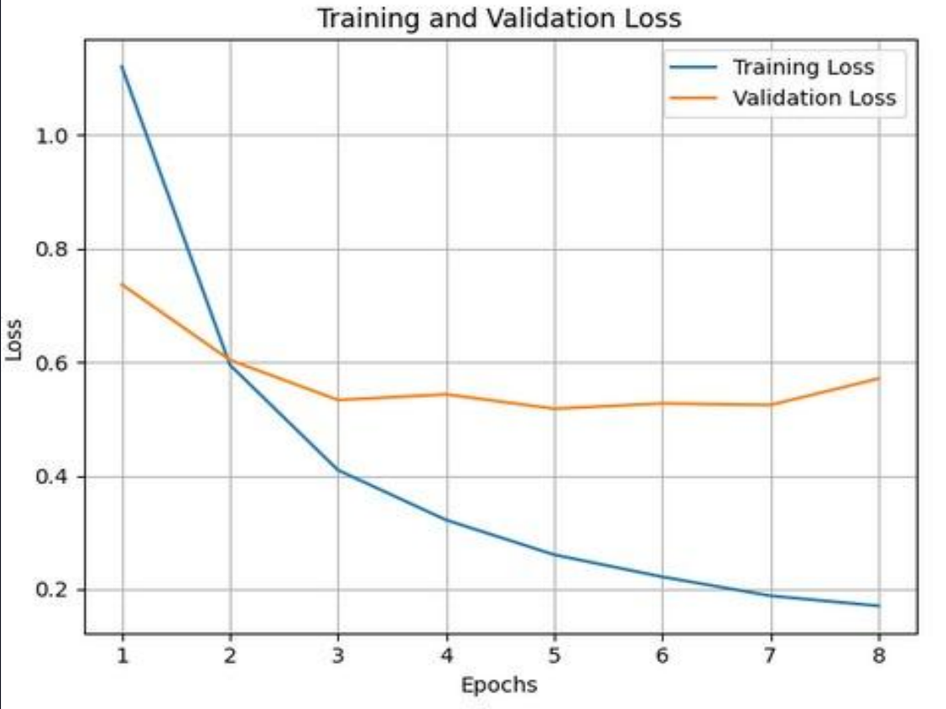
- Hard Voting: majority vote
- Soft Voting: weighted probabilities
- Stacking: meta-learner Logistic Regression

Results

KNN	Accuracy: 0.8641	Macro F1: 0.8611
Soft Voting Ensemble	Accuracy: 0.8779	Macro F1: 0.8745 (Best)
Stacking Ensemble	Accuracy: 0.8673	Macro F1: 0.8647
LSTM	Accuracy: 0.8287	Macro F1: 0.8239
Basic	Naive Bayes: 0.6278 Logistic Regression: 0.6770, SVM: 0.6996, XGBoost: 0.6503, Hard Voting: 0.7351	0.5957 0.6697 0.6906 0.6225 0.7261

Soft Voting Ensemble outperforms individual models by leveraging strengths.

LSTM Training and Validation Accuracy



Challenges

Data Quality

Noisy text with slang, emojis; class imbalance handled by weights.

Model Integration

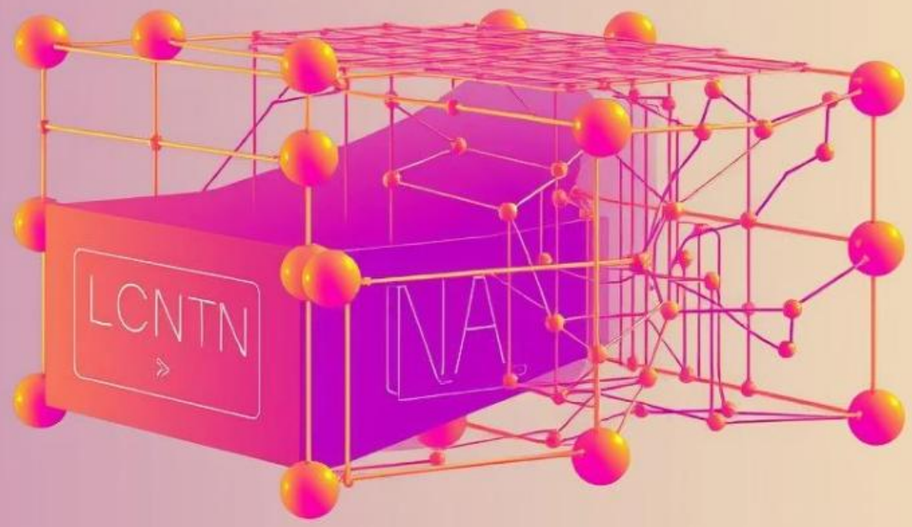
Stacking combined LSTM and ML models with different input formats.

Computational Cost

LSTM training was resource-heavy; tuning took significant time.

Overfitting Risk

Mitigated using early stopping and dropout in LSTM.



Future Work

1

CNN Implementation

Capture local text patterns with 1D convolutions and pooling.

2

Hybrid Ensemble

Combine CNN, LSTM, and ML models with advanced stacking.

3

Additional Improvements

- Use pretrained embeddings like GloVe and BERT
- Expand dataset with more social media platforms
- Optimize hyperparameters via grid or Bayesian search

Conclusion

Achievements

Built a high-performing sentiment system using ensemble learning.

Soft Voting Ensemble achieved best accuracy and F1 scores.

Integrated diverse ML and DL models for robust predictions.

Impact & Next Steps

Supports social media monitoring, brand management, market research.

Plan to implement CNN, enhance ensembles, and use pretrained embeddings.