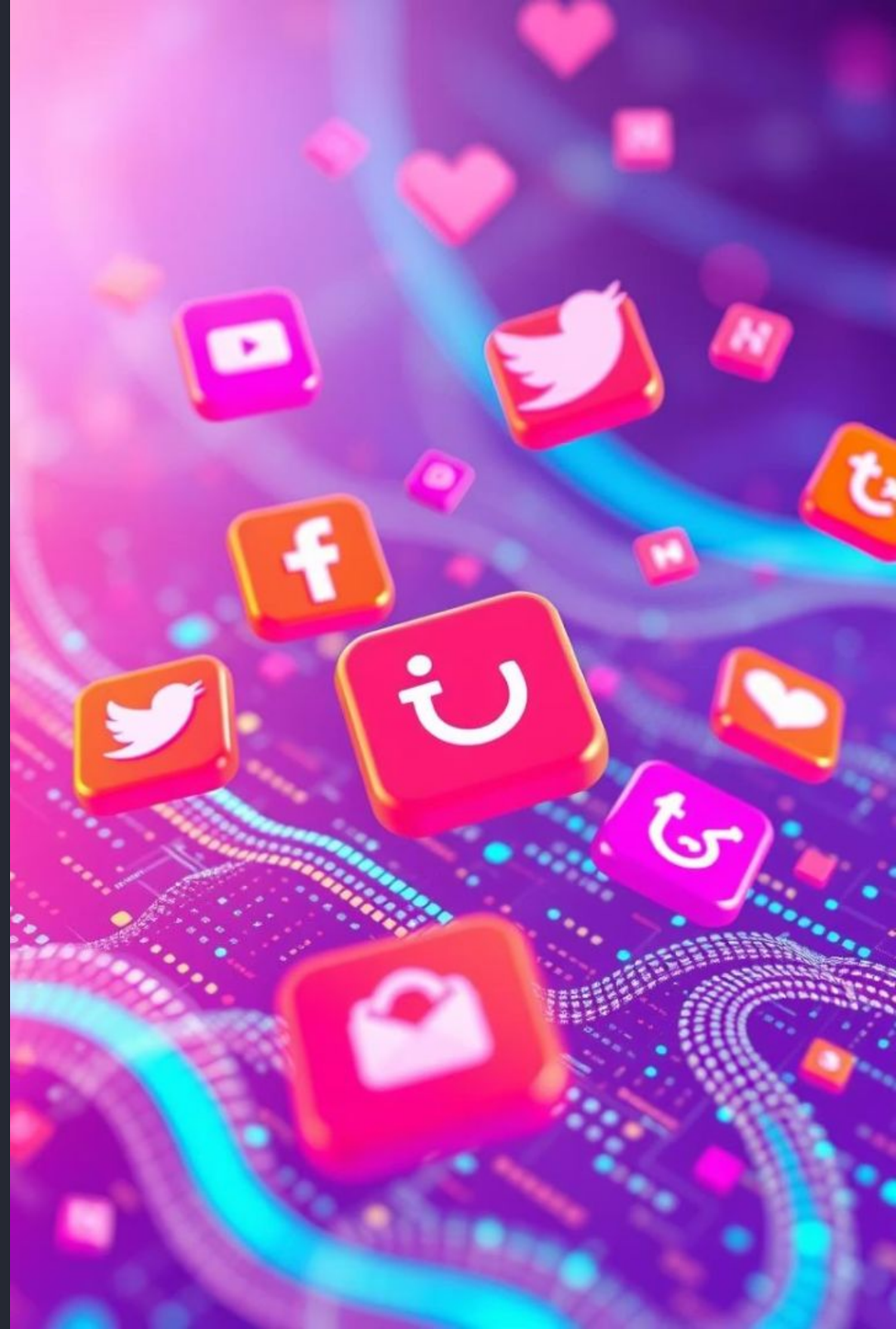
Social Media Sentiment Analysis System Using Ensemble Learning

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.

Methodology

Data Preprocessing

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

Models

- Naive Bayes, Logistic Regression, SVM, KNN, XGBoost
- CNN, 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

DL

CNN, Bidirectional LSTM

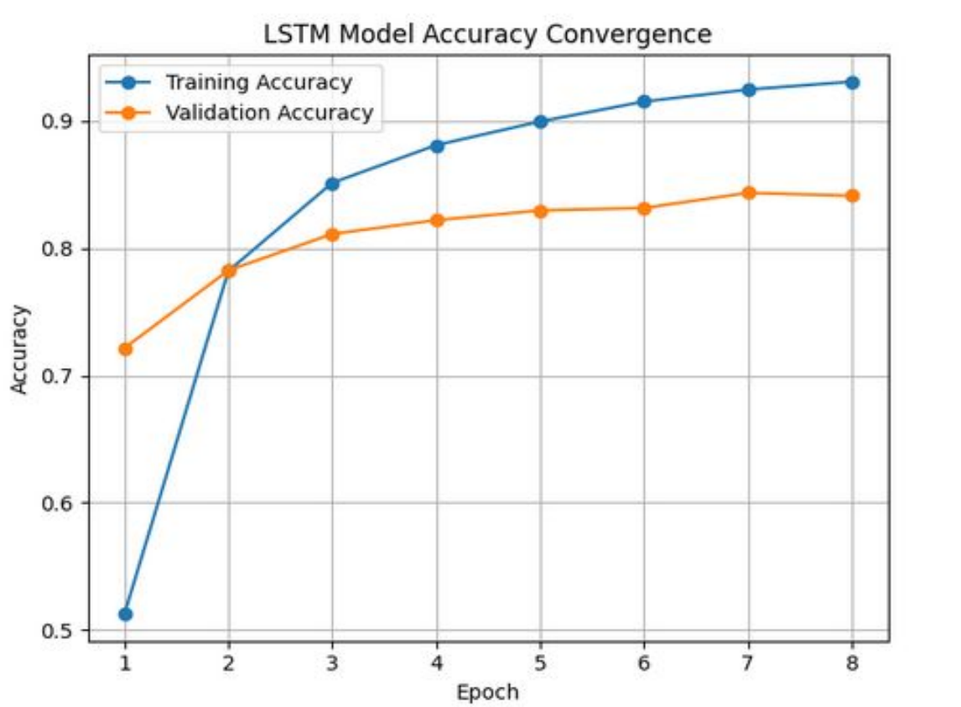
Ensemble

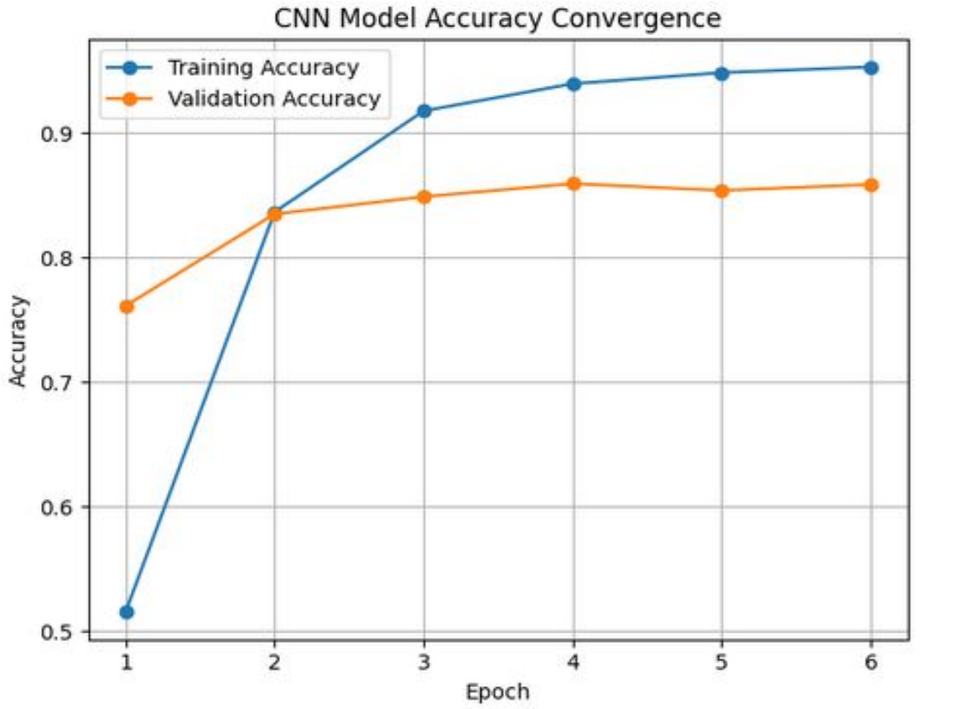
- Hard Voting: majority vote
- Soft Voting: weighted probabilities
- Stacking: meta-learner Logistic Regression
- Soft voting (LSTM + CNN)

Results

Model	Accuracy	Macro F1
Multinomial Naive Bayes	0.6367	0.6025
Logistic Regression	0.6766	0.6693
Support Vector Machine	0.7005	0.6920
K-Nearest Neighbors	0.8597	0.8568
XGBoost	0.6537	0.6267
Ensemble (Hard Voting)	0.7365	0.7272
Ensemble (Soft Voting)	0.8771	0.8738
Stacking Ensemble	0.8588	0.8565
LSTM	0.8230	0.8186
CNN	0.8459	0.8422
CNN + LSTM Ensemble	0.8613	0.8581

Soft Voting Ensemble outperforms individual models by leveraging strengths.





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.

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.