Group Project 499:

Twitter Sentiment Classification with NLP

Team Members

- Costas Frantzides - Stefanos Englezou - Vasilis Kynigaris

Project Overview

This project implements a sophisticated sentiment analysis system for Twitter data that classifies tweets as either positive or negative sentiment. The system combines modern NLP techniques (GloVe word embeddings) with extensive feature engineering to achieve robust sentiment classification performance.

Feature Types Used

1. Text Representation Features

- **GloVe Twitter Embeddings (200d)**: Pre-trained word vectors from GloVe Twitter corpus providing semantic representation

- **TF-IDF Features** : N-gram features (unigrams and bigrams) with 3100 maximum features

(TF-IDF was eventually not used because of GloVe's superiority)

2. Lexical & Syntactic Features

- **POS Tagging**: Counts and ratios of nouns, verbs, adjectives, adverbs, pronouns, interjections, auxiliary verbs, determiners, and particles

- **Negation Detection**: Count of negation words and particles

- **Word Statistics** : Word count, average word length, uppercase ratio

- **Elongated Words :** Detection of words with repeated characters

3. Sentiment & Emotion Features

- **TextBlob Sentiment** : Polarity and subjectivity scores

- **Opinion Lexicon** : Counts of positive and negative words from NLTK opinion lexicon

- **Emoji/Emoticon Analysis** : Custom sentiment scoring for emojis and emoticons

- **Sentiment Contrast** : Detection of polarity flips within text indicating sarcasm

4. Social Media Specific Features

- **Profanity Detection** : Count of profane words

- **Slang Usage** : Count of slang terms

- **Capitalization Patterns** : All-caps token count and uppercase ratio

- **Punctuation Analysis** : Special punctuation counts (!, #, @, ?) and repeated sequences

- **Hashtag Analysis** : Hashtag count and sentiment polarity

- **Named Entity Recognition** : Counts of persons, organizations, locations, dates, etc.

- **Mention & Hashtag Counts** : Social media specific markers

5. Advanced Feature

- **Sarcasm Detection** : Integration of a separate sarcasm classifier model

- **Emoticon Sentiment** : Separate counts for happy and sad emoticons

- **Contrast Analysis** : Detection of sentiment contrast within sentences

Model Implemented

Primary Classifier

- **SGDClassifier with Log Loss** : Optimized using Optuna hyperparameter tuning
- Optimized parameters: alpha, eta0, learning_rate
- 100 trial optimization with 3-fold cross validation
- F1-macro score optimization

```
[I 2025-11-26 15:16:15,306] Trial 95 finished with value: 0.8457568109979406 and parameters: {'alpha': 0.00040776016602363165, 'eta0': 0.04777317431707379, 'learning_rate': 'adaptive'}. Best is tria
l 62 with value: 0.8468842783849757.
[I 2025-11-26 15:16:16,327] Trial 96 finished with value: 0.834668165360207 and parameters: {'alpha': 0.0011911909795896804, 'eta0': 0.034308517188156244, 'learning_rate': 'optimal'}. Best is trial
62 with value: 0.8468842783849757.
[I 2025-11-26 15:16:16,719] Trial 97 finished with value: 0.5624272201428199 and parameters: {'alpha': 0.0020934410461260207, 'eta0': 0.025284417438184693, 'learning_rate': 'constant'}. Best is tria
l 62 with value: 0.8468842783849757.
[I 2025-11-26 15:16:17,748] Trial 98 finished with value: 0.8459676268630788 and parameters: {'alpha': 0.0008817841408442878, 'eta0': 0.020432316170001533, 'learning_rate': 'adaptive'}. Best is tria
l 62 with value: 0.8468842783849757.
[I 2025-11-26 15:16:18,817] Trial 99 finished with value: 0.846317854453818 and parameters: {'alpha': 0.001426984682629484 5, 'eta0': 0.057411150705927956, 'learning_rate': 'adaptive'}. Best is trial
 62 with value: 0.8468842783849757.
Best trial:
F1_macro: 0.8468842783849757
Params: {'alpha': 0.0007854783455671682, 'eta0': 0.02476139931496488, 'learning_rate': 'adaptive'}
```

Baseline/Comparison Models

- Random Forest Classifier : Used for feature importance analysis

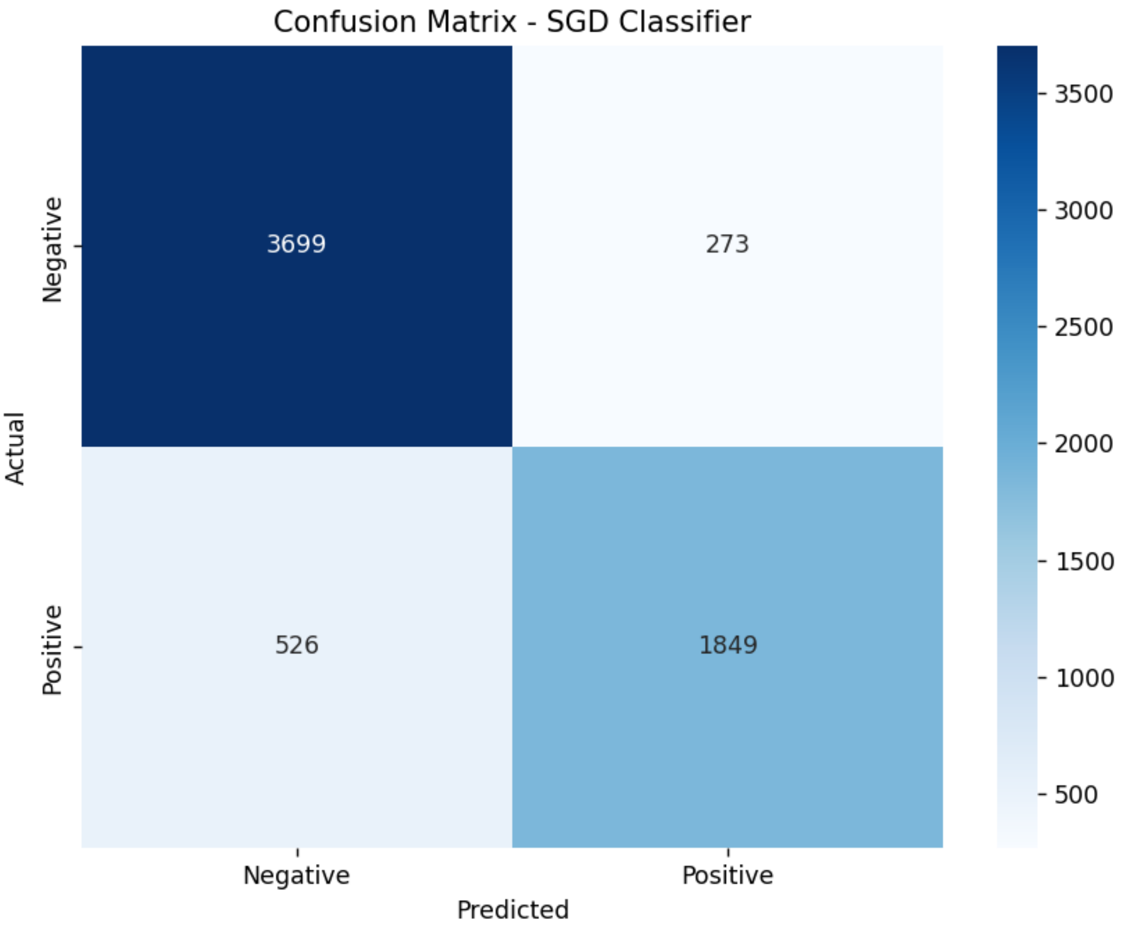
Preprocessing Pipeline

- Text normalization and Unicode cleaning- URL/email replacement with tokens
- Contraction expansion
- Tokenization, lowercasing, stopword removal
- Lemmatization using spaCy

Key Results and Findings

Performance Metrics

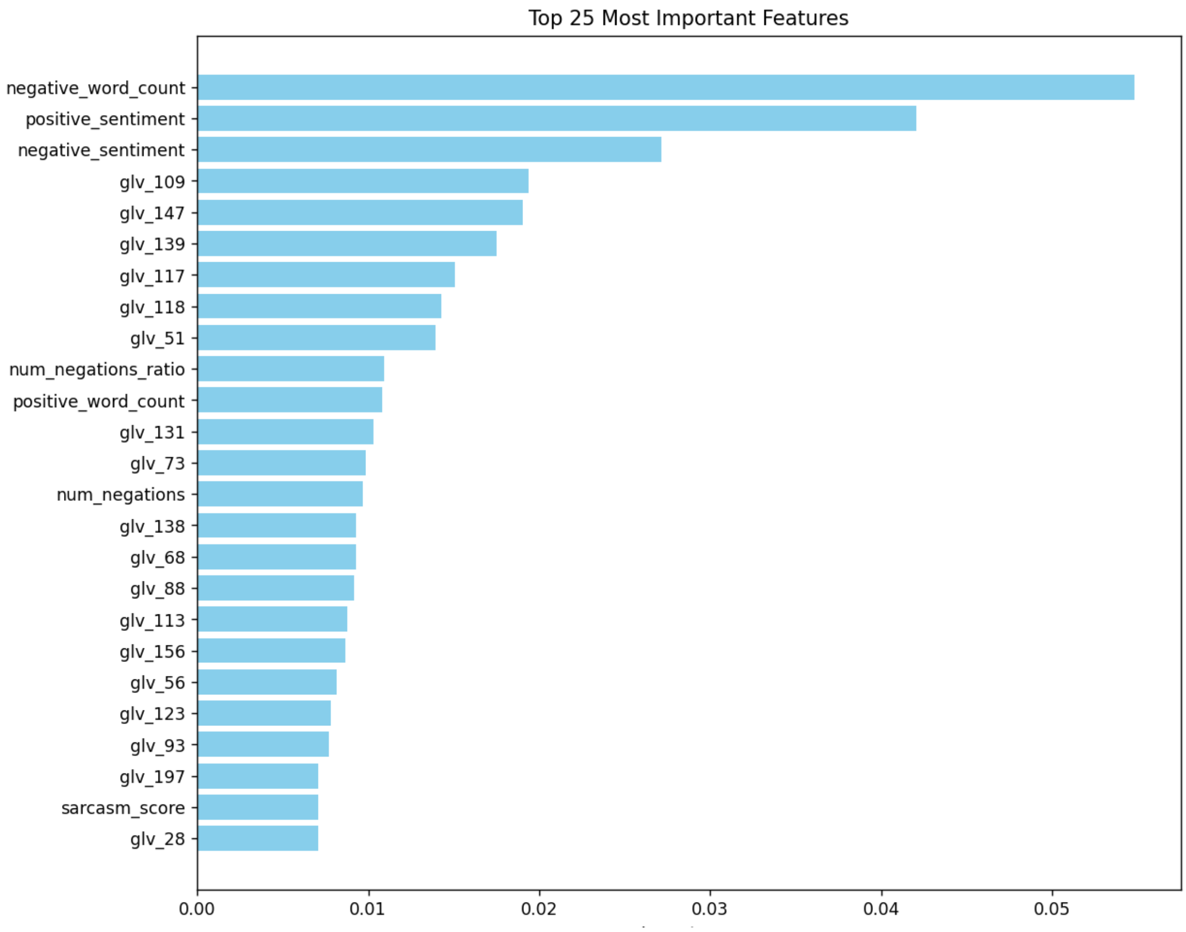
- Optimized SGDClassifier achieved strong classification performance
- Comprehensive evaluation using accuracy, F1-score, precision, and recall
- Confusion matrix analysis for error pattern identification




```
=== Final SGDClassifier (Optuna tuned) ===
Accuracy: 0.8741137545296991
              precision    recall  f1-score   support

    negative       0.88      0.93      0.90      3972
    positive       0.87      0.78      0.82      2375

    accuracy                           0.87      6347
   macro avg       0.87      0.85      0.86      6347
weighted avg       0.87      0.87      0.87      6347


================================================================================
Total Misclassified: 799 out of 6347 (12.59%)
================================================================================
```

Feature Importance Insights

- Random Forest feature importance revealed the most impactful features
- Top 25 features visualization provided interpretability
- GloVe embeddings combined with engineered features showed complementary strengths



Rationale for Feature Choices

1. Linguistic Foundation: Capturing Language Structure

POS Tagging Features

- ****Rationale****: Different parts of speech carry varying sentiment weight. Adjectives and adverbs are strong sentiment indicators, while nouns provide context. Verbs can indicate actions with emotional connotations.

- ****Expected Impact****: Help model understand grammatical structure and identify sentiment-heavy sentence components

Negation Detection

- ****Rationale****: Negation words fundamentally flip sentiment polarity (e.g., "good" vs "not good"). Critical for accurate sentiment interpretation.

- ****Expected Impact****: Prevent misclassification of negated positive/negative statements

2. **Social Media Specificity: Platform-Aware Features**

Emoji/Emoticon Analysis

- ****Rationale****: Emojis and emoticons are direct emotional indicators in social media. They often convey sentiment more reliably than text.

- ****Expected Impact****: Capture non-textual emotional expressions common in Twitter communication

Capitalization and Punctuation Patterns

- ****Rationale****: ALL CAPS indicates emphasis or strong emotion. Repeated punctuation (!!!, ???) shows emotional intensity.

- ****Expected Impact****: Detect emotional intensity and emphasis patterns unique to informal digital communication

Hashtag and Mention Analysis

- **Rationale**: Hashtags often contain sentiment-bearing words. Mentions indicate social context.

- **Expected Impact**: Leverage Twitter-specific metadata for sentiment cues

3. **Lexical Semantics: Word-Level Sentiment Indicators**

GloVe Twitter Embeddings

- **Rationale**: Pre-trained on Twitter data, capturing semantic relationships in social media language. Provides dense vector representations that understand context and word similarity.

- **Expected Impact**: Semantic understanding beyond bag-of-words, handling of vocabulary not seen in training

Opinion Lexicon Features

- **Rationale**: Established sentiment lexicons provide reliable baseline for word-level polarity

- **Expected Impact**: Strong baseline sentiment indicators from validated lexical resources

Profanity and Slang Detection

- **Rationale**: Profanity often indicates strong negative sentiment. Slang terms common in Twitter may carry sentiment not in formal dictionaries.

- **Expected Impact**: Capture informal language and strong emotional expressions

4. **Psychological and Pragmatic Features**

Sentiment Contrast and Polarity Flips

- **Rationale**: Sarcasm often manifests as sentiment contrast within a sentence. Polarity flips can indicate ironic or sarcastic content.

- **Expected Impact**: Improve performance on sarcastic or ironic tweets that would otherwise be misclassified

Elongated Word Detection

- **Rationale**: Word elongation ("soooo goooood") indicates emphasis and emotional intensity in informal communication

- **Expected Impact**: Capture emphasis patterns common in expressive social media language

5. **Statistical and Structural Features**

Text Statistics

- **Rationale**: Tweet length and word complexity can correlate with sentiment expression style

- **Expected Impact**: Capture writing style differences in emotional expression

Named Entity Recognition

- **Rationale**: Mentions of specific entities (people, organizations) can provide context for sentiment interpretation

- **Expected Impact**: Contextual understanding of what the sentiment is directed toward

6. **Advanced Composite Features**

Sarcasm Detection Integration

- **Rationale**: Sarcasm is a major challenge in sentiment analysis. Dedicated sarcasm detection can flag tweets where surface sentiment is misleading.

- **Expected Impact**: Significant improvement in handling ironic and sarcastic content

Weighted Sentiment Fusion

- **Rationale**: Different sentiment indicators (text, emojis, emoticons) may have varying reliability. Weighted combination leverages strengths of each source.

- **Expected Impact**: More robust sentiment scoring by combining multiple signal types

Feature Selection Philosophy

Diversity Overlap

Each feature category was chosen to capture different aspects of sentiment expression:

- **Semantic**: GloVe embeddings

- **Lexical**: Opinion words, profanity

- **Syntactic**: POS tags, negation

- **Pragmatic**: Capitalization, punctuation

- **Contextual**: Named entities, hashtags

- **Psychological**: Sentiment contrast, sarcasm

Redundancy as Robustness

Deliberate feature overlap (e.g., multiple ways to detect emphasis) to provide robustness - if one feature type fails on certain data, others may succeed.

Twitter Platform Optimization

All features were selected with Twitter's unique communication style in mind: short, informal, emotive, and rich with non-standard language features.

This comprehensive feature set was designed to give the model multiple pathways to detect sentiment, making it robust to the varied ways people express emotions on social media.