Twitter Sentiment Analysis Using Machine Learning

Name- Krishna Kumar

College Name- NIT Raipur

Semester- 6th

Abstract

This project performs sentiment analysis on Twitter data using machine learning to classify tweets as positive or negative. The dataset containing 1.6 million labeled tweets was utilized. The tweets were pre-processed through cleaning, stop word removal, and stemming, followed by vectorization using TF-IDF. Logistic Regression was employed for classification, achieving a test accuracy of approximately 77.6%. The model effectively distinguishes public sentiment, demonstrating the applicability of machine learning for social media analytics. Future work could explore more advanced models and deeper text representations.

---

1. Introduction

Problem Statement

Understanding public sentiment on social media platforms like Twitter is valuable for marketing, politics, and social research. The challenge is to automatically classify tweets into positive or negative sentiments based on their textual content.

Motivation

Manual sentiment analysis is not scalable given the vast number of tweets generated daily. Automating this process helps gain timely insights into public opinion and trends.

Goals

The main goal is to develop a machine learning model that accurately classifies tweet sentiment using natural language processing (NLP) techniques and to evaluate its performance on a large-scale dataset.

Dataset

The Sentiment140 dataset was used, which contains 1.6 million tweets labeled as positive (1) or negative (0). This dataset serves as a standard benchmark for sentiment classification tasks.

.

2. Data Description

Dataset Source

- Kaggle: Sentiment140 Dataset
- Size: 1.6 million tweets

Features

- **`text`:** The tweet content
- **`target`:** Sentiment label (0 = negative, 1 = positive)
- Other columns like, **`id`, `date`, `flag`, `user`** are present but not used for modeling.

Pre-processing Steps

- Removed special characters and numbers from tweets
- Converted all text to lowercase
- Removed stop words using NLTK's English stop word list
- Applied Porter stemming to reduce words to their root form
- Re-joined processed words into cleaned tweets

Handling Missing Data

No significant missing values were found in the text or target columns.

Descriptive Statistics

- Class distribution is roughly balanced after mapping label 4 to 1 (positive)
- Number of tweets: 1,600,000 (approximate from dataset size)

---

3. Exploratory Data Analysis (EDA)

Sentiment Distribution

The dataset contains nearly equal proportions of positive and negative tweets, supporting balanced classification.

Text Length

The average tweet length is short (due to Twitter's character limits), influencing feature extraction.

Word Frequency

Common words vary across sentiments; stemming reduces feature complexity by grouping similar words.

4. Modeling

Model Selection

Logistic Regression was chosen for its simplicity, interpretability, and effectiveness on text classification tasks.

Feature Extraction

TF-IDF Vectorizer converts cleaned text into numerical feature vectors, highlighting important words relative to their frequency.

Train-Test Split

The data was split into 80% training and 20% testing sets with stratification to preserve label distribution.

Hyperparameters

- Logistic Regression was configured with a maximum of 1000 iterations for convergence.

Evaluation Metrics

- Accuracy was used to measure performance on both training and testing data.

---

5. Results

Performance

- Training Accuracy: 79.8%
- Testing Accuracy: 77.6%

Discussion

The model shows reasonable accuracy given the complexity of natural language and noisy Twitter data. However, it may struggle with sarcasm, slang, or ambiguous tweets.

Limitations

- Logistic Regression might not capture complex semantic relationships
- Basic pre-processing may miss nuances in language

---

6. Conclusion

This project demonstrates a practical approach to sentiment classification of Twitter data using logistic regression. The pre-processing pipeline and model training process provide a solid foundation for social media sentiment analysis. Future improvements may include using deep learning models (e.g., LSTM, BERT) and enhanced feature engineering for better accuracy and generalization.

7. References

- Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision.
- Kaggle. Sentiment140 Dataset: https://www.kaggle.com/datasets/kazanova/sentiment140
- Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
- Scikit-learn Documentation: https://scikit-learn.org/
- NLTK Documentation: https://www.nltk.org/

Thank You