

Twitter Sentiment Analysis - Using Machine Learning

1. Project Overview

This project focuses on analyzing the sentiment of tweets using machine learning techniques. It classifies tweets as either positive or negative based on their textual content. The analysis provides valuable insights into public opinion on various topics using natural language processing (NLP).

2. Dataset

Name: Sentiment140

Size: 1.6 million labeled tweets

Source: <https://www.kaggle.com/datasets/kazanova/sentiment140>

Labels: 0 (negative), 4 (positive)

3. Technologies Used

Python

Jupyter Notebook

Pandas

NLTK (Natural Language Toolkit)

Scikit-learn

4. Data Preprocessing

The tweet text is cleaned and transformed to prepare it for machine learning:

- Removal of special characters, numbers, and punctuation
- Conversion to lowercase
- Tokenization (splitting into words)
- Stop word removal
- Stemming using Porter Stemmer
- Rejoining stemmed words into clean tweet strings

5. Model Training

Algorithm: Logistic Regression

Input Features: Processed tweet text converted to numeric format using TF-IDF or CountVectorizer

Target: Tweet sentiment (positive or negative)

Split: Training and test sets (typically 80:20)

6. Model Evaluation

Accuracy measured on test data

Other metrics (optional): Precision, Recall, F1 Score

The model shows good generalization in identifying sentiment trends in unseen tweets

7. Most Useful Concept: Stemming

Stemming is crucial for reducing word complexity and improving model performance by treating related word forms as one. It simplifies text data, enabling better pattern recognition by the machine learning model.

- Implemented using NLTK's PorterStemmer
- Reduces redundant features and improves model generalization

8. Conclusion

This project effectively demonstrates how machine learning can be used to classify tweet sentiment with high accuracy. The pipeline-from preprocessing to model training-is scalable and adaptable to real-world applications such as social listening, product feedback, or political sentiment tracking.