Project Proposal: Analyzing Public Sentiment on Social Media Text Using Machine Learning

Jason Park

June 28, 2025

1. Problem Statement

Text sentiment analysis on social networks (Twitter) is the process of automatically identifying and categorizing the sentiment expressed in text, positive, negative, or neutral, using natural language processing (NLP) and machine learning techniques. Given the real-time and large-scale nature of social media data, this form of analysis plays a crucial role in understanding public emotions and opinions related to products, events, or social issues.

This project aims to analyze and classify the sentiment of texts using the Sentiment140 dataset, which consists of 1.6 million labeled texts. Each text is annotated with its sentiment polarity (positive or negative), and the dataset includes various metadata such as tweet ID, timestamp, user information, and tweet text. The goal is to build a classification pipeline that can accurately distinguish between positive and negative sentiments in text content.

2. Motivation and Applications

Twitter is a popular microblogging platform where users share thoughts and opinions in real-time. Analyzing tweet sentiment can provide actionable insights across domains:

- **Business Insight:** Understand customer satisfaction and detect complaints.
- **Reputation Management:** Monitor public perception of brands or organizations.
- **Political Sentiment:** Gauge voter mood and reactions to political events.
- **Event Response:** Track public response during natural disasters, pandemics, or crises.

3. Proposed Methodology

The analysis will follow a structured NLP pipeline that includes the following stages:

1. **Dataset Loading:** Load the Sentiment140 dataset using `pandas`. Each tweet includes sentiment labels (0 = negative, 4 = positive), tweet ID, date, user ID, and text.
2. **Data Cleaning and Preprocessing:**
    - Remove HTML tags, special characters, and URLs using regular expressions.
    - Normalize casing, strip whitespace, and remove accented characters.

- Perform tokenization, remove stopwords, and apply stemming or lemmatization.
- Use the NLTK library to apply standard NLP preprocessing routines.

3. **Label Transformation:**
   - Convert sentiment labels to binary: 0 (negative), 1 (positive).
   - Remove neutral tweets to ensure binary classification.
4. **Feature Extraction:**
   - Apply **TF-IDF vectorization** to convert tweets into numerical feature representations.
   - Extract both unigrams and bigrams to improve feature richness.
5. **Model Training:**
   - Train a **Multinomial Naive Bayes** model using `scikit-learn`'s `MultinomialNB`.
   - Use `fit()` to train the model on the TF-IDF transformed tweets.
   - Apply the trained model to the test set and generate predictions.
6. **Model Evaluation:**
   - Evaluate the model using `classification_report()` which includes accuracy, precision, recall, and F1-score.
   - Visualize performance using a confusion matrix.

4. Planned Evaluation Strategy

The proposed model will be evaluated through:

- **Classification Report:** Evaluate accuracy, precision, recall, and F1-score for positive and negative classes.
- **Confusion Matrix:** Understand true positive, true negative, false positive, and false negative distributions.
- **Cross-Validation (optional):** If computationally feasible, use 5-fold cross-validation to ensure generalizability.

5. Dataset Description

The Sentiment140 dataset includes 1.6 million labeled tweets collected using the Twitter API. Each record includes:

- **target:** Sentiment label (0 = negative, 4 = positive)
- **id:** Unique identifier for the tweet
- **date:** Timestamp of tweet posting
- **query:** Search query used (may be "NO QUERY")
- **user:** Username who posted the tweet
- **text:** Content of the tweet

Only tweets with positive or negative sentiment will be retained.

6. Novelty and Contribution

This project highlights how a simple yet effective pipeline combining text preprocessing, TF-IDF vectorization, and a Naive Bayes classifier can yield strong baseline results on a challenging real-world dataset. Contributions include:

- Reproducible ML pipeline for binary sentiment classification
- Evaluation of Multinomial Naive Bayes on large-scale social media text
- Foundation for future extensions using deep learning or transformer-based models

7. Team Member Responsibilities

This is an individual project conducted by Jason Park. All tasks including data handling, model development, evaluation, and report writing will be completed independently.

References

- Go, Alec, Richa Bhayani, and Lei Huang. "Twitter Sentiment Classification using Distant Supervision." Stanford University (2009).
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 2011.
- `https://www.kaggle.com/datasets/kazanova/sentiment140`