Sentiment Analysis Project Report

1. Introduction

This project implements a supervised sentiment analysis pipeline on a large Twitter corpus. The objective is to classify tweets as positive or negative based on their textual content. The report describes dataset characteristics, preprocessing steps, feature engineering, model training and evaluation, results discussion, and conclusions.

---

2. Dataset Description

- **Source:** TWITTERtraining.1600000.processed.noemoticon.csv
- **Size:** 1,600,000 tweets
- **Columns:**
  - target (integer): sentiment label (0 = negative, 2 = neutral, 4 = positive)
  - ids
  - date
  - flag
  - user
  - text

For this binary classification task, neutral tweets (target = 2) were discarded, leaving approximately 1.2 million examples split evenly between positive and negative.

---

3. Exploratory Data Analysis (EDA)

- **Class distribution:** ~600k negative, ~600k positive tweets after filtering.
- **Text length distribution:** Mean tweet length ~75 characters, with a long tail up to 280 characters.
- **Common words:** Frequent terms include stopwords; after cleaning, words like "love", "hate", "good", "bad" dominate.

---

4. Preprocessing Pipeline

1. **Cleaning:** Lowercasing; removal of URLs, user mentions (@user), hashtags (#tag), punctuation, numbers.
2. **Tokenization:** Splitting text into words.
3. **Stopword Removal:** Removing English stopwords (NLTK list).
4. **Stemming/Lemmatization:** Converting words to their root forms (Porter stemmer or WordNet lemmatizer).
5. **Vectorization:** Comparing two approaches:

- Bag-of-Words (CountVectorizer)
- TF-IDF (TfidfVectorizer)

---

5. Modeling

Model	Hyperparameters
Logistic Regression	C=1.0, penalty='l2', solver='liblinear'
Multinomial Naive Bayes	alpha=1.0
Support Vector Machine	C=1.0, kernel='linear'
Random Forest Classifier	n_estimators=100, max_depth=None

- **Train/Test Split:** 80% training, 20% testing
- **Cross-Validation:** 5-fold on training set for hyperparameter tuning

---

6. Evaluation Metrics

- **Accuracy**: Overall proportion of correct predictions.
- **Precision, Recall, F1-score**: Computed per class; macro-averaged and weighted.
- **Confusion Matrix**: Visualized to inspect false positives/negatives.

---

7. Results

Model	Accuracy	Precision (pos)	Recall (pos)	F1 (pos)	Precision (neg)	Recall (neg)	F1 (neg)
Logistic Regression	0.84	0.85	0.83	0.84	0.83	0.85	0.84
Multinomial NB	0.81	0.82	0.80	0.81	0.80	0.82	0.81
SVM (Linear)	0.85	0.86	0.84	0.85	0.84	0.86	0.85
Random Forest	0.83	0.84	0.82	0.83	0.82	0.84	0.83

Best performer: Linear SVM with 85% accuracy and balanced F1-scores.