Sentiment140 Preprocessing Summary

Author: MingHsiang Lee
Course: INFO 7610
Assignment: Dataset Selection and Preparation
Date: 2026-02-04

Abstract

This report presents a preprocessing pipeline applied to the Sentiment140 dataset (1.6M tweets). The workflow covers duplicate removal, text cleaning, tokenization, stopword removal, lemmatization (with stemming for comparison), and dataset splitting into train/validation/test sets. Basic statistics are included for downstream sentiment analysis.

1. Introduction

Sentiment140 is a large-scale dataset of tweets labeled for sentiment. Data preprocessing reduces noise, normalizes text, and produces consistent inputs for machine learning models.

2. Dataset Used

- **Source:** Kaggle – Sentiment140 (KazAnova)
- **File:** `training.1600000.processed.noemoticon.csv` (renamed to `Sentiment140.csv`)
- **Labels:** 0 = negative, 4 = positive (converted to binary: 0/1)

3. Cleaning and Preprocessing

3.1 Deduplication

- Removed duplicate tweets based on identical text content.

3.2 Missing/Empty Text Handling

- Dropped rows with missing text.
- Detected tweets that became empty after cleaning.

3.3 Text Cleaning

Applied the following steps:

- URL removal
- HTML tag removal
- @mention removal
- Hashtag symbol removal (kept word)
- Emoji removal
- Special character removal
- Whitespace normalization
- Lowercasing

3.4 Tokenization & Normalization

- Tokenization using NLTK `TweetTokenizer`.
- Stopword removal (NLTK English stopwords).
- Lemmatization using WordNet (primary normalized text).
- Stemming using PorterStemmer (comparison only).

4. Basic Statistics (Final)

I report these values from the notebook outputs.

- **Final dataset size:** 1,578,669 tweets
- **Negative (0):** 788,695 (49.96%)
- **Positive (1):** 789,974 (50.04%)
- **Text length (mean):** 65.08 characters
- **Tokens per tweet (mean):** 9.12
- **Text length range:** min and max reported in the notebook
- **Token count range:** min and max reported in the notebook

5. Train/Validation/Test Split

- **Train:** 70% (1,105,068)
- **Validation:** 15% (236,800)
- **Test:** 15% (236,801)
- **Stratification:** Maintained label distribution across splits.