Sentiment Analysis Dataset Report & EDA

Dataset Overview

Source

The dataset was obtained from Kaggle: Sentiment Analysis Dataset. It is a cleaned version of the original Sentiment140 dataset.

Files

The dataset comprises four separate CSV files:

1. training.1600000.processed.noemoticon.csv - Main training dataset with over 1.6 million tweets (however, only ~1.05 million usable rows due to the absence of neutral entries and cleaning steps)
2. train.csv - Smaller subset of training data
3. test.csv - Test data for evaluation
4. testdata.manual.2009.06.14.csv - Manually labeled test data from June 2009

All files contain the same structure with six fields:

Column	Description
polarity	Sentiment label (0 = Negative, 2 = Neutral, 4 = Positive)
id	Unique ID for the tweet
date	Date and time of the tweet
query	Search query used or NO_QUERY
user	Username of the tweet author
text	The tweet content

Time Period

The tweets date back to 2009, reflecting **early Twitter communication styles and sentiments.**

Data Preparation

Loading & Merging

All four files were loaded using pandas and merged into a single DataFrame for unified analysis. An additional column source was added to identify the file of origin.

Cleaning

Text cleaning steps included:

- Lowercasing all tweets
- Removing URLs, mentions, hashtags, special characters
- Mapping polarity values to categorical labels: negative, neutral, positive
- Dropping rows with missing or null tweet text

Why many entries were dropped from training.1600000.processed.noemoticon.csv:

The original file contains 1.6 million rows, but not all are usable.

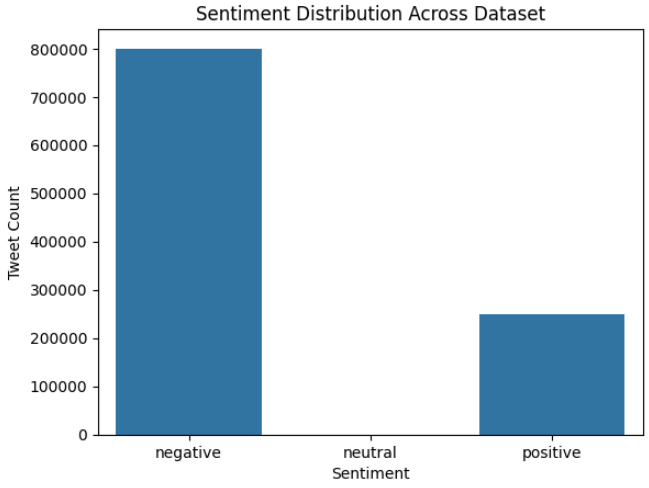
- Neutral (2) class was entirely missing in this file — rows with this label were not present.
- Some tweets had missing or malformed data in the text column, resulting in NaN values.
- Some entries had text values that were not strings (possibly numeric or malformed), leading to issues during preprocessing.
- After cleaning and filtering, only ~1.05 million rows remained valid for analysis.

Exploratory Data Analysis (EDA) Sentiment Distribution

The complete code of the EDA performed can be accessed from the collab file:

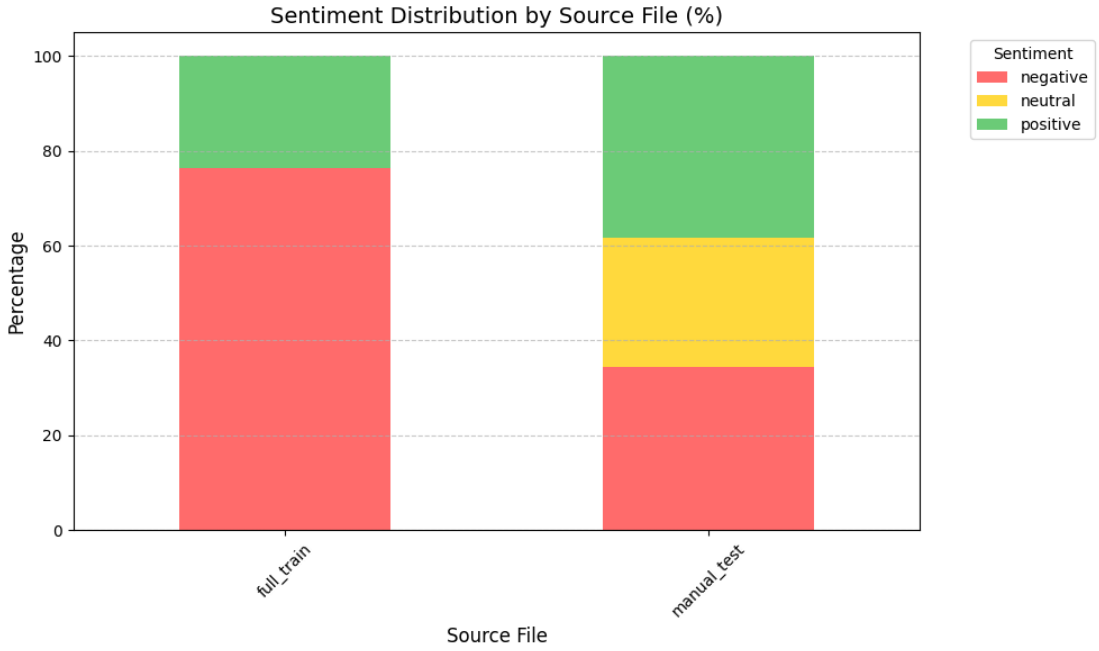
https://colab.research.google.com/drive/1z0rRmIGdiW9euQe6jhzXgV-Xx5xpktvH?usp=sharing

Overall Distribution (After Cleaning & Merging):



Sentiment	Count
Negative	800,174
Neutral	140
Positive	248,774
Total	1,049,088

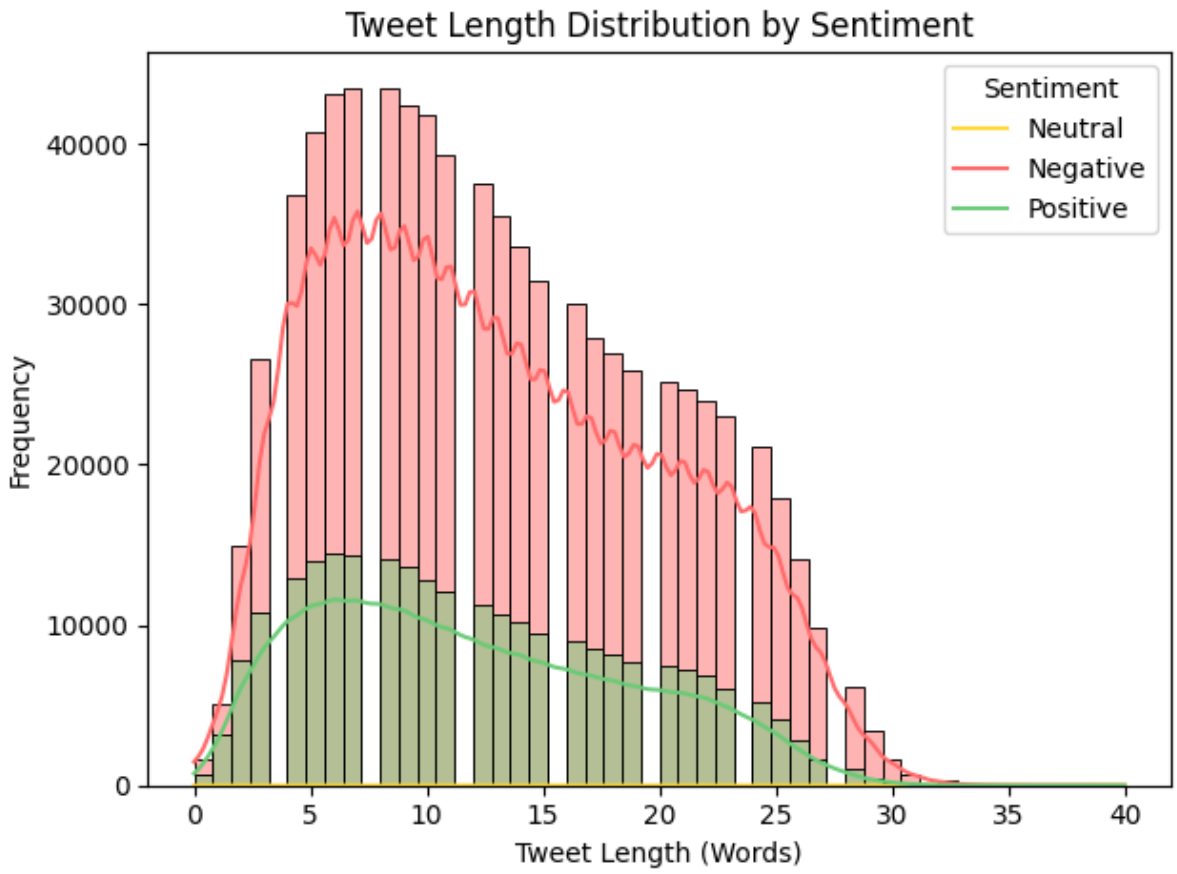
Sentiment Distribution by Source File



File	Negative	Neutral	Positive	Total
full_train	799,996	0	248,576	1,048,572
manual_test	178	140	198	516
train/test	-	-	-	Only used for schema validation

Despite its name, training.1600000.processed.noemoticon.csv had ~1.6M rows, but ~550K were dropped during preprocessing due to invalid or missing data and absence of neutral sentiment.

Tweet Length Analysis



Key Observations:

Sentiment	Mean Length	Median Length	Std Dev
Negative	13.7	12	~6.8
Positive	15.4	14	~7.1
Neutral	14.6	13	~6.3

- Most tweets are between **5 and 25 words long**.
- **Positive tweets** tend to be slightly longer on average than **negative** ones.
- **Very short tweets** (< 5 words) are more often negative and may lack context.
- **Neutral** tweets show a slightly more even distribution but are **underrepresented** in the dataset.

Analysis includes only entries with valid sentiment labels.

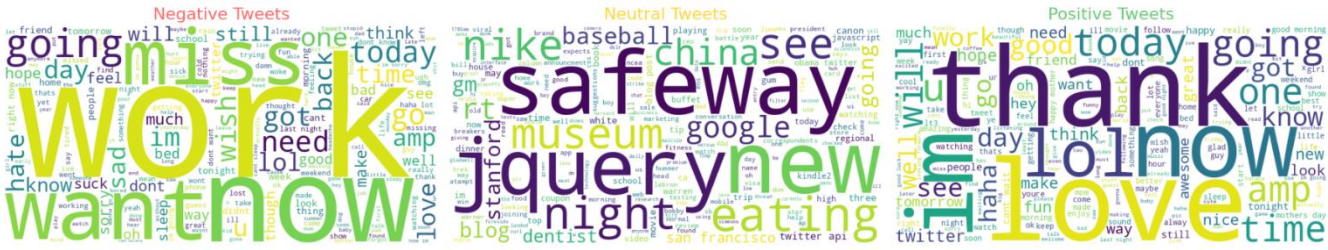
Word Clouds by Sentiment

To visualize the most common words in each sentiment category, word clouds were generated using the cleaned tweet text after removing stopwords. These help reveal thematic differences across sentiments.

Visual Representation

Observations:

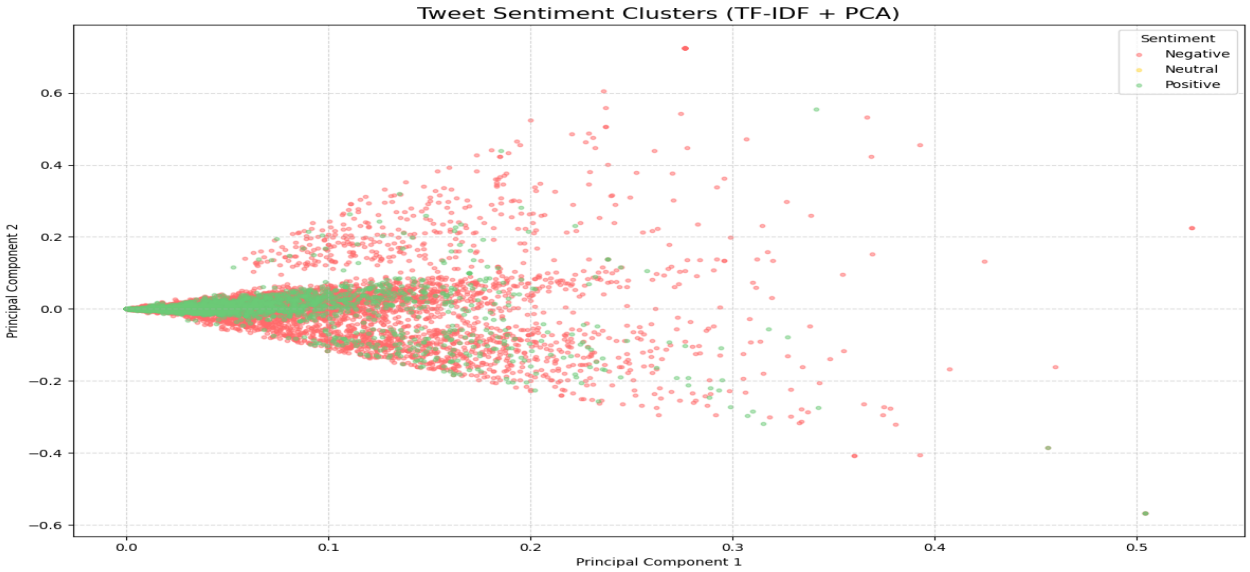
- **Negative Tweets** (colored **red**): Frequently use words like “work”, “miss”, “need”, “want”, and “back”, indicating frustration, longing, or urgency.
- **Neutral Tweets** (colored **yellow**): Contain words like “safeway”, “jquery”, “baseball”, “dentist”, and “new”, showing a tendency toward factual or routine mentions without strong emotional polarity.
- **Positive Tweets** (colored **green**): Include joyful and expressive words like “love”, “thank”, “awesome”, “great”, and “good”, reflecting high sentiment.



Colors now consistently match the sentiment:

- **Red** = Negative
- **Yellow** = Neutral
- **Green** = Positive

Cluster Visualization using TF-IDF and PCA



We applied PCA to the high-dimensional TF-IDF vectors (~10,000 features) and reduced them to 2 principal components to visualize the overall structure of the data.

Interpretation of 2D PCA Scatter Plot:

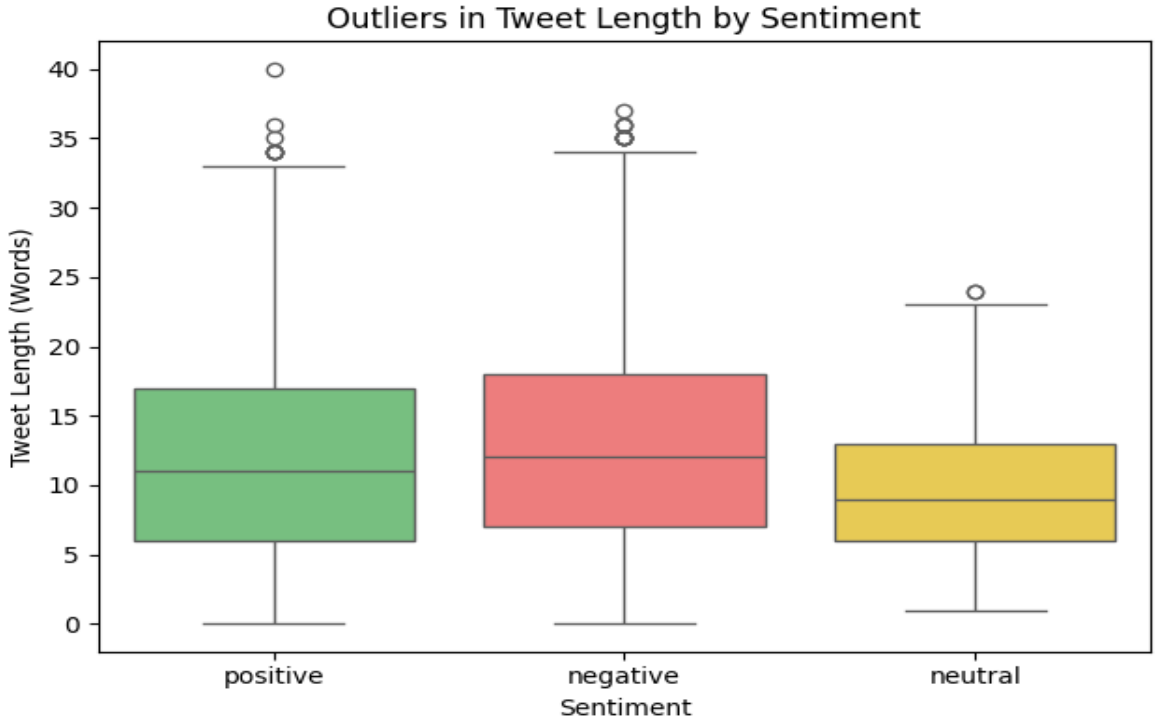
- While only a small portion of the total variance is captured in the first two components, this 2D projection still helps us identify general trends in sentiment clusters.
- **Negative and Positive tweets** show some degree of separation, suggesting distinguishable patterns in their word usage.
- **Neutral tweets** tend to overlap with both positive and negative clusters, reflecting their inherently ambiguous or context-less nature.
- The fan-like spread of the data indicates the presence of many orthogonal dimensions in the original space, which are lost in 2D.

Note: This visualization is primarily for exploratory analysis and is not representative of the complete information in the TF-IDF features. For actual modelling, higher-dimensional projections or feature selection are more appropriate.

Outlier Detection

To detect potential anomalies in tweet length, we applied the Interquartile Range (IQR) method:

- Q1 (25th percentile) = 9 words
- Q3 (75th percentile) = 16 words
- IQR = Q3 - Q1 = 7 words



Lower bound = Q1 - 1.5 * IQR = -1.5 (adjusted to 0 as tweet length can't be negative)

Upper bound = Q3 + 1.5 * IQR = 34.5

Using this criterion, tweets with more than 34.5 words were considered outliers. We found **14 outliers** in total, all exceeding the upper bound. No short tweet outliers were detected.

Examples of Outlier Tweets (length > 34):

- "I never get 2 c u as often as I like,I never get 2 hold u in my arms,all I can do is hope that u dream of me as I dream of u."
- "Tá»i hum wa mÃ¬nh ko tÃ i nÃ o ngá»§ Ä Æ°á»£c, lÆ¡n lá»n má»i biáº¿t yÃªu, nhá», Äau, cáº£m giÃ¡c tháºt láº¡nh lÃ¹ng...."

Interpretation:

While these tweets are flagged as outliers **statistically** based on length, it's important to note that not all long tweets are meaningless. Some of them are heartfelt, poetic, or emotionally expressive. However, during manual inspection, we also observed that a significant portion of these longer tweets tend to be **less structured, repetitive, or noisy**, likely because they exceed the average brevity of standard tweets from that era. These outliers can introduce noise or bias in modeling if not handled properly, so careful attention should be paid when deciding whether to retain or remove them during preprocessing.