

DATA APPENDIX

Project: Sentiment Shifts in Vaccine-Related Tweets by Theme

Group: Model Citizens

Members: Shaina Banduri, Neil Parikh, Nishana Dahal

Course: DS 4002

Date: Feb. 2026

This appendix documents every variable in the datasets used in this project. Three CSV files are described below:

1. covid-19_vaccine_tweets_with_sentiment.csv (raw data)
 2. covid19_vaccine_tweets_cleaned.csv (after preprocessing)
 3. covid19_vaccine_tweets_analyzed.csv (after analysis)
-

DATASET 1: Raw Data

File: covid-19_vaccine_tweets_with_sentiment.csv

Rows: 14,151 Columns: 3

Unit of observation: One tweet from Twitter related to COVID-19 vaccines, with a human-annotated sentiment label.

Variables:

tweet_id (float64)
Unique numerical ID for each tweet.
Example: 1.360342e+18

label (int64)
Human-annotated sentiment label.
Values: 1 = Negative, 2 = Neutral, 3 = Positive

tweet_text (string)
Full text content of the tweet, including hashtags, URLs, and @mentions.

DATASET 2: Cleaned Data

File: covid19_vaccine_tweets_cleaned.csv

Rows: 6,000 Columns: 3

Unit of observation: One cleaned tweet. Rows with missing tweet text were removed. Text was lowercased, URLs removed, @mentions removed, # symbols stripped, whitespace normalized.

Variables:

tweet_id (float64)
Same as raw data. Unique tweet identifier.

label (int64)
Same as raw data.
Distribution in cleaned set:
1 (Negative): 420 (7.0%)
2 (Neutral): 3680 (61.3%)
3 (Positive): 1900 (31.7%)

tweet_text (string)
Cleaned tweet text. All lowercase, no URLs, no @mentions, no # symbols (hashtag text preserved), single-spaced.

DATASET 3: Analyzed Data
File: covid19_vaccine_tweets_analyzed.csv
Rows: 6000 Columns: 13

Unit of observation: One cleaned tweet enriched with VADER sentiment scores and theme/brand indicator flags.

Original variables (same as cleaned data):
tweet_id, label, tweet_text

--- Added Variables ---

vader_compound (float64)
VADER normalised compound sentiment score [-1, 1].
-1 = most negative, +1 = most positive.
Mean: 0.1313
Median: 0.0000
Std: 0.4457
Min: -0.9816
Max: 0.9718

vader_pos (float64)
Proportion of text with positive sentiment [0, 1].
Mean: 0.0941

vader_neg (float64)
Proportion of text with negative sentiment [0, 1].
Mean: 0.0494

vader_neu (float64)
Proportion of text with neutral sentiment [0, 1].
Mean: 0.8565

tweet_length (int64)
Character count of the cleaned tweet text.
Mean: 161.8
Median: 164.0
Min: 12
Max: 293

DATASET 3 (continued): Theme & Brand Indicator Flags

theme_safety (bool)
True if the tweet contains at least one keyword from the safety/side-effects dictionary.
True: 716 (11.9%)
False: 5284 (88.1%)

theme_access (bool)
True if the tweet contains at least one keyword from the access/appointments dictionary.
True: 1671 (27.9%)
False: 4329 (72.2%)

theme_eligibility (bool)
True if the tweet contains at least one keyword from the eligibility dictionary.
True: 0 (0.0%)
False: 6000 (100.0%)

theme_general (bool)
True if the tweet contains at least one keyword from the general-information dictionary.
True: 4179 (69.7%)
False: 1821 (30.3%)

brand_mention (bool)
True if the tweet mentions at least one vaccine brand (Pfizer, Moderna, AstraZeneca, Covaxin, etc.).
True: 5640 (94.0%)
False: 360 (6.0%)

=====

END OF DATA APPENDIX