Distinguishing AI-Generated Tweets from Human Language

COURSE NAME

CSDS 413 Introduction to Data Analysis

Authors:

Wiam Skakri, Jacob Anderson, and Aishani Patil

October 28, 2025

Research Question

Can linguistic and stylistic features (e.g., vocabulary density, sentence length, word length ...) statistically distinguish AI-generated tweets from human-written tweets?

Hypotheses

H_0: There is no significant difference between human-written and AI-generated tweets in their textual feature distributions.
H_1: There is a significant difference between human-written and AI-generated tweets in their textual feature distributions.

Dataset(s)

Dataset: TweepFake (Fagni et al., 2020).
Source: Kaggle.
Represents: Real tweets posted on Twitter, balanced between human-written and AI-generated content.
Variables: Tweet text, binary label (human/bot), generation method (Markov, RNN, LSTM, GPT-2).
Size: 25,572 tweets.
Samples: 12,786 human, 12,786 bot-generated from 23 bot accounts imitating 17 humans.

Execution

Below are the five features we intend to extract from each tweet for the purpose of our analysis. The dataset prior to extracting these features is cleaned of any URLs, mentions, and hashtags, as from initial inspection, bot tweets don't appear to include these and inclusion may also throw off extraction:

Vocabulary Richness (V)

V measures the ratio of unique words in tweet i to the total number of words in tweet i:

$$V_i = \frac{\text{total unique words in tweet i}}{\text{total words in tweet i}}$$

Sentence Length (S)

S is the average length of each sentence in a tweet, delimited by standard punctuation symbols as well as newline characters:

$$S_i = \frac{\text{total words in tweet i}}{\text{total sentences in tweet i}}$$

Word Length (W)

W is the average length of each word in a tweet:

$$W_i = \frac{1}{n_{\text{words}}} \sum_{j=1}^{n_{\text{words}}} \text{len}(\text{word}_j)$$

Function Word Frequency (F)

F captures the ratio of function words in a tweet to total words in a tweet:

$$F_i = \frac{\text{total function words in tweet i}}{\text{total words in tweet i}}$$

Capitalization Abnormality (C)

C measures the ratio of words containing abnormal capitalization patterns:

$$C_i = \frac{\text{total words with non-standard capitalization in tweet i}}{\text{total words in tweet i}}$$

Validation

We intend to pursue a permutation testing framework using Mahalanobis distance as our test statistic. Each tweet i is represented as this five-dimensional feature vector:

$$\mathbf{x}_i = [V_i, S_i, W_i, F_i, C_i]$$

For each group, mean feature vectors will be collected:

$$\bar{\mathbf{x}}_{\text{Human}} = \frac{1}{n_{\text{human}}} \sum_{i \in \text{Human}} \mathbf{x}_i$$

$$\bar{\mathbf{x}}_{\text{AI}} = \frac{1}{n_{\text{AI}}} \sum_{i \in \text{AI}} \mathbf{x}_i$$

and a pooled covariance matrix will be collected as:

$$\mathbf{S}_{\text{pooled}} = \frac{(n_H - 1)\mathbf{S}_H + (n_{AI} - 1)\mathbf{S}_{AI}}{n_H + n_{AI} - 2}$$

where $\mathbf{S}_H$ and $\mathbf{S}_{AI}$ are the sample covariance matrices for each group.

That being said, the test procedure will first compute mean feature vectors $\bar{\mathbf{x}}_{\text{Human}}$ and $\bar{\mathbf{x}}_{\text{AI}}$, compute the pooled covariance matrix $\mathbf{S}_{\text{pooled}}$, invert the covariance matrix to obtain $\mathbf{S}^{-1}_{\text{pooled}}$, and compute our test statistic for the observed data as:

$$D_{M,obs} = \sqrt{(\bar{\mathbf{x}}_{\text{Human}} - \bar{\mathbf{x}}_{AI})^T \mathbf{S}^{-1}_{\text{pooled}} (\bar{\mathbf{x}}_{\text{Human}} - \bar{\mathbf{x}}_{AI})}$$

For $b = 1, 2, \ldots, 10,000$, we randomly shuffle the label column while keeping the feature vectors $\mathbf{x}_i$ fixed, computing new mean vectors $\bar{\mathbf{x}}^{(b)}_H$ and $\bar{\mathbf{x}}^{(b)}_{AI}$ for each permutation. Using the same covariance matrix $\mathbf{S}^{-1}_{\text{pooled}}$, we calculate the Mahalanobis distance $D_{M,b}$ for each permutation, each representing a new possibility under H_0.

A p-value will then be given for this collection of test statistics as:

$$p = \frac{\#\{D_{M,b} \geq D_{M,\text{obs}}\}}{10,000}$$

Regarding visualization of the result, we will plot a histogram of the Mahalanobis distances across the permutations, with a marker at $D_{M,\text{obs}}$ overlaying the distribution, representing the p-value visually by the distribution mass beyond the observed distance.

Visualization

We will create the following visualizations to support our analysis:

1. **Permutation test histogram:** Distribution of Mahalanobis distances across 10,000 permutations with observed distance marked, visually representing the p-value.

2. **Feature comparison box plots:** Side-by-side box plots for each feature (V, S, W, F, C) comparing human vs. AI distributions to identify which features drive group separation.

3. **PCA scatter plot:** Two-dimensional projection of the five-dimensional feature space, with points colored by tweet type, showing overall group separation.

Implementation will use Python's Matplotlib and Seaborn libraries.

Interpretation

If $p < 0.05$, we reject the null hypothesis and interpret the result as evidence that we cannot claim the difference we observed between the ai-generated and human-written textual feature centroids is what we could expect if the linguistic structures of artificial and natural tweets were actually not distinguishable. The implication of this result would be that the AI tweets are not performing well to imitate natural language because differences can be meaningfully identified by a handful of simple attributes.

Otherwise, we fail to reject the null hypothesis and must concede that artificial and natural tweets are not distinguishable from these textual features and by proxy their overall linguistic structures. More generally, this would support the idea that AI tweets do well to imitate natural language.

Work Plan

Aishani Patil (Data Preparation):

- Download and preprocess TweepFake dataset; clean tweets (remove URLs, mentions, hashtags)
- Handle edge cases and perform exploratory data analysis
- Tokenization and sentence segmentation

Wiam Skakri (Statistical Analysis):

- Extract five textual features (V, S, W, F, C) from cleaned tweets
- Compute mean vectors, covariance matrix, and Mahalanobis distances
- Execute permutation testing framework (10,000 iterations) and calculate p-value

Jacob Anderson (Validation & Visualization):

- Validate feature extraction accuracy through manual inspection
- Create all visualizations (permutation histogram, box plots, PCA, heatmap)
- Interpret results and identify key discriminative features