Overview of Twitter Scraped Data

Using a combination of Selenium and Python with the Chrome Driver, we conducted web scraping to retrieve the latest 50 tweets from each Twitter account (denoted as X). To enhance the analysis, we initially vectorized the tweets, assigning a score to each. Subsequently, we employed the Latent Dirichlet Allocation (LDA) algorithm to categorize the tweets into two clusters: Bot-generated or Human-generated. To ensure optimal data preparation, we utilized the NLTK library for comprehensive pre-processing of the tweets. Throughout this process, several observations and assumptions emerged.

First, we noticed tweets with a different tone and style, which we thought might be generated by bots or fake accounts. Second, we assigned each topic a value, 0 or 1. 0 represents normal texts written by humans, and 1 represents texts likely posted by fake accounts.

A graph of a distribution of tweets

Description automatically generated

Sentiment Analysis of Tweets:

We employed the TextBlob library in Python, specifically its "Sentiment Polarity" feature, to assess the tone of the tweets. The analysis led to the following plots.

A graph of a number of blue bars

Description automatically generated with medium confidence

A bar graph showing different colors

Description automatically generated

There are some interesting notes and assumptions:

 First, the majority of tweets had a neutral tone, indicated by a sentiment value of 0

 Second, for the remaining tweets, the overall sentiment trend was positive which indicates that the text expresses favorable or optimistic feelings. Probably because the purpose of the post was advertising China and making this assumption that everything is perfect about China and none off the rumors are true.