**Term Project**

Justin Pizzoferrato

Bellevue University

DSC530-T302: Exploration and Analysis

Professor Mathew Metzger

November 16, 2024

**Term Project**

**Statistical/Hypothetical Question**

The central question guiding this analysis was: *What factors contribute to the virality of social media posts, and how do attributes such as platform, sentiment, and hashtags influence engagement metrics like likes and retweets?* This question aimed to uncover patterns and relationships between these factors and user engagement to better understand what drives popularity and visibility online.

**Outcome of Exploratory Data Analysis (EDA)**

The EDA provided key insights into the dataset’s structure and the variables’ behaviors. Histograms revealed the distribution of five variables: platform, sentiment, likes, retweets, and hashtags. Outliers were identified, such as posts with extraordinarily high engagement metrics, which were further analyzed for context and impact. A cumulative distribution function (CDF) highlighted that the majority of posts received fewer than 30 retweets, confirming a typical “long tail” distribution of engagement. Analytical distributions showed that likes followed a near-normal distribution, with slight skewness, allowing for better modeling. Hypothesis testing suggested that sentiment significantly affects likes, and a regression analysis between likes and retweets showed a near-perfect linear relationship, confirming strong positive correlation.

**Missed Opportunities in Analysis**

While the analysis comprehensively addressed the key variables, it could have been enriched by incorporating additional metadata, such as the time of posting or the type of content (image, text, video). These factors could have added more depth to the insights regarding what drives virality. Additionally, sentiment analysis could have been improved by categorizing sentiments more granularly or by incorporating additional emotional dimensions.

**Additional Variables That Could Have Helped**

Including variables like the time of day, follower count, or post type (e.g., photo, video) would have provided more explanatory power. Such data could have helped understand how timing and content format influence engagement.

**Incorrect Assumptions**

One assumption made during the analysis was that all hashtags equally contribute to engagement. In reality, hashtags’ effectiveness likely varies based on their relevance and popularity. Similarly, it was assumed that platforms contribute equally to engagement when, in reality, each platform’s user behavior differs significantly.

**Challenges and Areas of Uncertainty**

The primary challenges included handling categorical variables, such as hashtags and sentiments, in a way that provided meaningful insights without oversimplifying. Understanding the subtle differences in platform dynamics required domain expertise that wasn’t fully available. Additionally, hypothesis testing required careful consideration of statistical assumptions, which might not always hold in social media data.

**Conclusion**

Overall, this analysis provided valuable insights into factors influencing social media engagement, though gaps in data and assumptions present opportunities for further exploration. Understanding virality is a complex problem, and future studies should incorporate additional data and more refined models to capture the nuances of user behavior.