**ML-Assisted Knowledge Extraction of Events and Public Reception.**

**1. Abstract**

In the age of social media, Twitter is a medium for the everyday person to not only voice their opinion on large topics, but also connect with and influence thousands as a result. In 2020, a year of elections, lockdowns, and vast differences of opinion, Twitter was ranked 4th in the world for engagement as external influence through social media resulted in historical events that weren’t all positive. In this research we utilize Machine Learning and Data Science techniques to analyze tweets throughout the year and visualize their correlation with large events and public reception. The project emphasizes the use of algorithms for the purpose of finding what influences people; whether that is classifying harmful bots that spread misinformation or analyzing text sentiment to quantify the attitude of users. The research is currently in progress, but preliminary results suggest a strong polarization of opinions throughout the year, as well as an increase in misinformation spreading twitter bots. Our goal is to use this research to warn about the harmful influence of social media in polarizing people’s opinions on political topics.

**2. Data Collection**

**2.1 Web Scraping**

The primary initial problem to solve is how to access Twitter data from the past year. This is hindered especially by Twitter’s recent blocking of all mining ability for tweets older than a week, creating a significant initial blockage for the study. Furthermore, the changing of web scraping endpoints in November 2020 disabled a large proportion of regularly used Twitter Web Scrapers. One scraper, however Twint, was fixed using endpoint workarounds by the general Data Science community. The Twint API makes it possible to not only mine any Tweets using given code words, but also to have a continuous scrape without rate limitations. As such, over 5 million tweets regarding two separate events were downloaded to get an initial view of the Twitter data on two large events.

Events were chosen and analyzed based on expected impact on the community. For example, events include the Coronavirus Briefing by President Trump on March 13th, 2020 and the U.S. Election on November 11th, 2020.

**2.2 Events and Local Control**

The tweets regarding all events were collected from before and after the expected “climax” of tweets by as much as 24 hours in order to establish a control for the important trends. This is important because of a trend is found for a certain event, we wanted to guarantee that the correlation was between the data and the event and not the general time period. These different periods are referred to as the **Leading Local Control**, **Event of Interest**, and **Trailing Local Control**. An example of an event opposed to its controls is shown in Figure 1. Trends looked for were regarding the ability of the event to influence other uses, including mapping the most frequently used terms, calculating a sentiment analysis of tweets, bot classification, and user interaction. More information on how these methods were utilized is in the section with experimental results.

Chart, line chart

Description automatically generated

Trailing Local Control

Leading Local Control

Area of Interest

Figure 1: Illustrates the data collection of one such event like the Coronavirus Briefing on March 13, 2020. In this case, along with the area of interest, a leading and trailing set of control data was also collected in order to identify significance of event.

**2.3 Global Control**

Since often the effects of large events on the general Twitter data can affect multiple days in a ripple effect, a **Global Control** was also established in order to track trends outside of the large event data. To do this a dataset of 366,000 tweets was gathered by mining 1000 tweets at the same time every day. This is useful in order to have something to generally compare to for trends.

**3. Results**

**3.1 Word Map**

One of the most basic initial takeaways from the text data gathered is a word cloud illustrating the most frequent words. While it is expected that Donald Trump would be a frequent subject, it is interesting to notice that he and his family are visible in the cloud far more times than expected, in terms such as “Trump supporter”, “Ivanka Trump” and “Trump say”.

Another important note is the largest term “http”, parsed from any tweets that contained a link. <A further look into correlation between bots, sentiment, and http links would be a useful addition to this section>.

Text

Description automatically generated

Figure 2: Word map of tweets from the March Coronavirus briefing.

**3.2 Sentiment Analysis**

While tweets obviously increase during the events of interest, they also tend to stay

Diagram

Description automatically generated

increased afterwards in the trailing local control. This then begs the question of if we can measure the sentiments of the tweets and how they correlate with these spikes.

To do this we built a sentiment analysis model as above in order to calculate a 0 to 1 sentiment of the tweet. Each word passes through the embedding and into the LSTM layer, before being max pooled and eventually concatenating into a probability of being considered a strong opinion based on the diction used.

This sentiment value is to try and assess strong opinions on whatever the topic at hand is. This can then be compared to the influx of tweets at each time.

Chart, line chart, scatter chart

Description automatically generated

One of the first takeaways from this analysis is the amount of variance in sentiment which increases in times of less frequency of tweets. While this of course makes sense since variance would decrease in times of high frequency, it is interesting because the trends persist outside of the averaging. <I would like to further investigate this>

The briefing data shows no large trend differences overtime because it is enough of a large spike without too much lasting effect (that we have found so far). Let’s look at a larger, steadier increase such as the data for the election tweets.

Chart, scatter chart

Description automatically generated

In this case there is a steady increase over days with related election tweets. When run through standard Linear Regression analysis, the R-Value is a 0.6017, indicating a relatively strong positive correlation between time and tweets. It is also notable that this is expected to be lower than the true increasing trend is because the shape of the Tweet count over time is obviously not directly linear. The much more linear shaped sentiment data increases with an R-value of 0.7035. All of this points toward proof of the visible correlation between sentiment and tweet count. Due to the high frequency of tweets, it is clear that this trend may persist in smaller spikes such as the briefing data, but is difficult to analyze due to its briefness. <I would also like to investigate this further>.