Twitter and Mastodon Investigation

Yuting Ji
Boston College
Chestnut Hill, MA, USA
jiyz@bc.edu

Jeremy Chan
Boston College
Chestnut Hill, MA, USA
channq@bc.edu

Abstract

In this study, we analyzed and investigated the aftereffect of Twitter acquisition. We used three models: the sentiment analysis model, the SIR compartmental model, and the Long Short-Term Memory(LSTM) model to investigate why people leave Twitter, analyze Mastodon's user number increase, and predict Mastodon's future user growth using the latter two models.

1. Introduction

Given the recent controversies surrounding Twitter, we want to find out why people were considering leaving the platform. At the same time, with rapid user increases, Mastodon, a decentralized social media platform, has been ushered into the spotlight of news articles and social media. Will the user numbers keep climbing in December?

To investigate these cases, we start with the following hypothesis:

1. People decide to leave Twitter because they dislike the new Twitter after the acquisition.

2. People will move from Twitter to other platforms in a "contagious" manner.

3. Mastodon user number will keep increasing until the end of 2022, but at a slower rate than October and November.

2. Data

To verify our hypothesis, we collected two datasets from Twitter and Mastodon.

2.1. Twitter data

We took Twitter Sentiment Data from November 2022 using Twitter's API. The data was based on several selected search queries related to the Twitter Takeover. The search queries were "Twitter, Elon Musk, Mastodon, Jack Dorsey, Decentralized Social Network, Social Media, Twitter Takeover, Tesla."

2.2. Mastodon data

The Mastodon user number data was provided by a decentralized platform tracking website called "the Federation". We extracted the total user number for each day from Jan 2022 to Nov 2022. Then we calculated the user number change every two days to get the new user number/new infection.

3. Models

3.1. Sentimental Analysis Model

For the sentiment analysis of tweets, we first used nltk's tokenizer to tokenize each Tweet we obtained from Twitter's API for each search query. Then we used the **TextBlob** library to conduct sentiment analysis on each tweet. The sentiment values range from -1 to 1. Afterward, we took the average of the sentiment values and assigned the average to the general sentiment of each search query.

3.2. SIR model

For the contagion modeling, we decided to go with the SIR model (Susceptible, Infectious, Recovered), as this compartment model would be able to divide people into different categories, and we could simulate the changes of the numbers in the S, I, and R categories. The Susceptible population represents those who have yet to sign up for Mastodon. The Infected population represents those who have signed up for Mastodon and can spread the word of Mastodon's existence to others. Lastly, the Recovered population represents inactive users or users who have left the Mastodon platform.

3.3. Long Short-Term Memory(LSTM, PyTorch Framework)

For the prediction, we decided to use the LSTM which is a Recurrent Neural Network. It can recognize data patterns over a long period, making it suitable for time-series

analysis and prediction. To make a seven-day prediction, we reorganized our data by using a window of 7 days. At each stop, seven days will be extracted as the input(X), and the 8th day will be extracted as the output or the target(y). By shifting this window across our Mastodon data, we established a time-series dataset used for the LSTM model training. The last seven pieces of data were reserved for our prediction of December.

4. Results

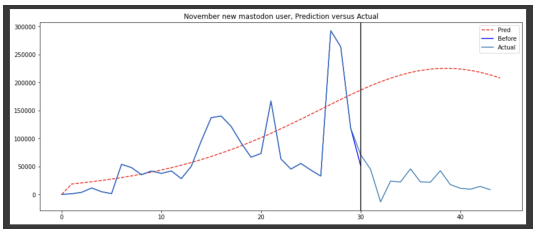
4.1. Sentimental Analysis Result

Search Query	Sentiment Value
Twitter	0.251
Elon Musk	0.035
Mastodon	0.139
Jack Dorsey	0.028
Decentralized Social Network	0.146
Social Media	0.093
Twitter Takeover	0.024
Tesla	0.186

As we can see, the sentiment on "Twitter" is higher than that of "Mastodon". The Sentiment for "Takeover" and "Elon Musk" is lower, making us believe that the general public is not as positive towards these topics. For "Takeover," the sentiment is slightly negative (as seen from the negative value). The negative sentiment is expected as the takeover was abrupt and affected many workers on Twitter. This received negative sentiment from many media outlets, so it was also surprising that the sentiment was not hostile. From an overall standpoint, the average sentiment is close to each other, and the variance of these values is also high due to the randomness of the tweets retrieved.

Based on these results, people are positive toward Twitter and relatively negative toward Elon Musk. The primary reason for leaving Twitter could be people's negative attitudes towards Elon Musk since he is the new owner of Twitter.

4.2. SIR Modeling Result

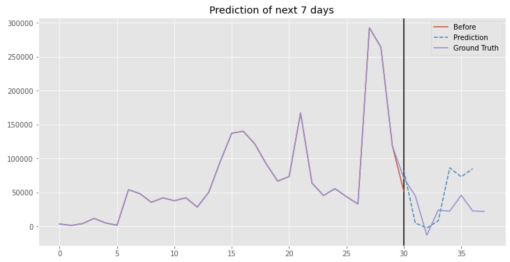


For the SIR simulation, it simulated the most user increase or the peak should appear around December 1st; after that, the number of new infections will decrease slowly. After running the model, we obtained a reproduction number of 1.1147 for new Mastodon users.

However, based on observation, the prediction doesn't match the actual result. This means that the Mastodon new users may not be a perfect contagious event, or it is a more complex situation that a simple SIR model won't be able to apply.

Although the prediction is too high, the SIR simulation generally captures the trend of decreasing the number of new infections in December. Such a trend aligns with our hypothesis.

4.3. LSTM Prediction Result



Compared to the actual results, the LSTM model predicts the values accurately. However, as observed from the graph, the number and the prediction trend are close to the actual result. This is because the LSTM model learns the trend more precisely based on historical data and can account for more drastic changes in new Mastodon users.

5. Evaluation

5.1. Twitter API limitation

Since we don't have advanced API access, we could only use a limited version with a threshold of 900 tweet requests per 15 minutes. From a limited sample size, it seems Twitter users like Twitter more than Mastodon. However, this can be attributed to the fact that we are talking about Tweets from the Twitter platform itself. Therefore, we expect that if we were to get Mastodon-related posts on Mastodon, Mastodon's sentiment values would be better than Twitter's.

If we had access to Twitter's Academic API, we would be able to get more data regarding the change in Twitter's user count and daily tweet volume to track the change of Elon Musk's takeover. We would then be able to use an SEIR model that considers the users who have left Twitter but have not signed up for Mastodon. We could also use the sentiment values of search queries to see the complete sentiment over time regarding Elon Musk's takeover of Twitter.

5.2. Prediction limitation

The SIR model is too simple to capture the daily user increases as it fluctuates too much. We believe that the reproduction number changes from range to range and varies based on many variables, such as external and internal events.

We will need to collect more Mastodon data for training LSTM and get better performance. In addition, we can also include other time-series data to improve training outcomes. Also, our current model is made for the prediction of 7 days. We should improve it to make it capable of predicting for a more extended period.

6. Source Code

Github Link

7. References

1. https://developer.twitter.com/en/docs/twitter-api
2. https://the-federation.info/
3. https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html
4. https://mc-stan.org/docs/functions-reference/index.html