COVID-19 Prediction via Vaccine Sentiment Analysis on Twitter

Jinyang Han, Jingjing Ye, Qingqing Wu

12/02/2021

Georgia Tech

Contents

- Introduction

- Related Work

- Data

- Method

- Results

- Conclusion

Jinyang Han
MS in CSE
jhan411@gatech.edu

Jingjing Ye
MS in CS
jye312@gatech.edu

Qingqing Wu
MS in CSE
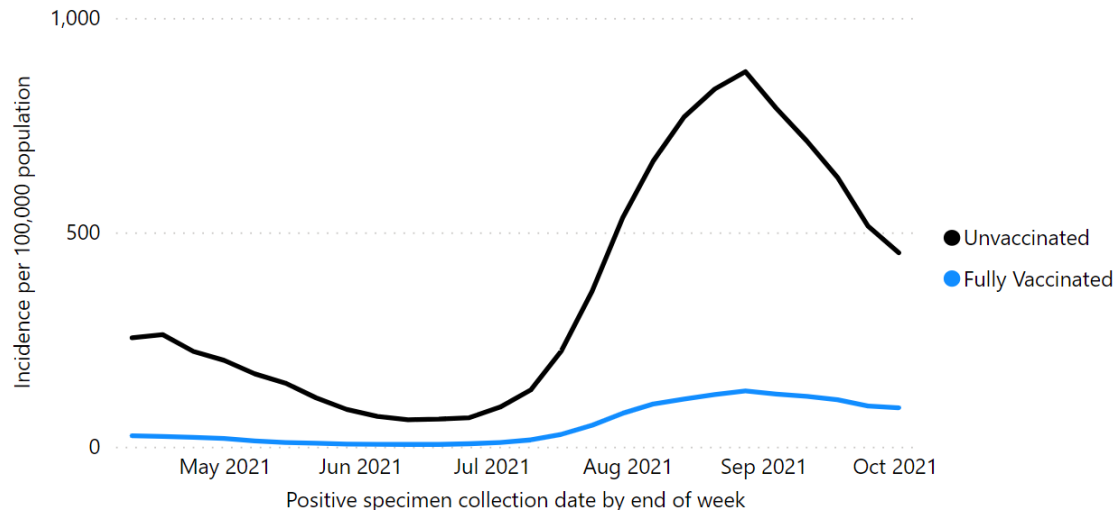qwu325@gatech.edu

Georgia Tech

Introduction

COVID-19 subverted the previous lifestyle.

Thanks to vaccine, the whole nation is on the path to being normal.

- **Lower risk of testing positive and dying**

- **Lower case/death rates in all age groups**

Rates of COVID-19 Cases by Vaccination Status

April 04 - October 02, 2021 (24 U.S. jurisdictions)

Rates of COVID-19 Cases by Vaccination Status and Age Group

April 04 - October 02, 2021 (24 U.S. jurisdictions)
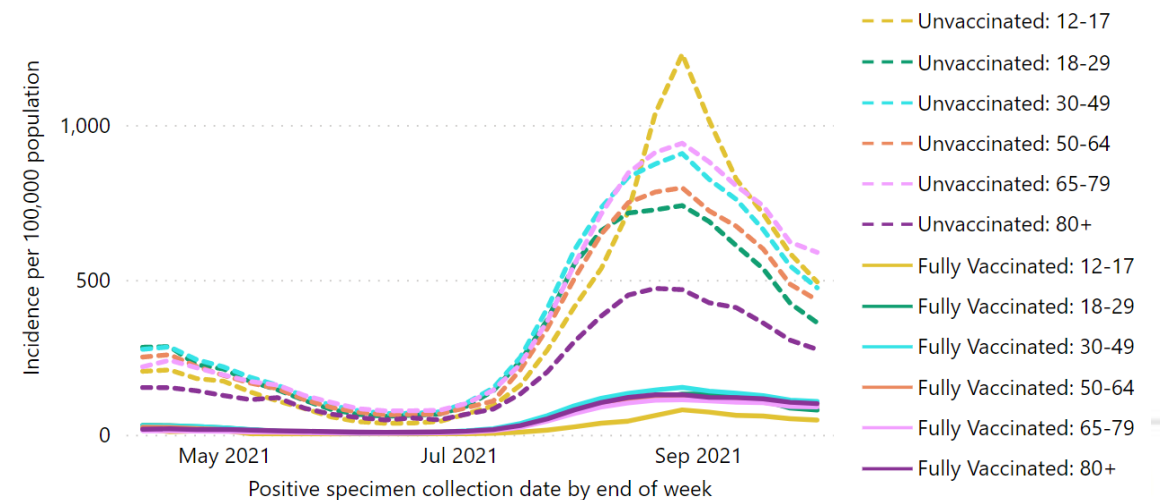
Introduction

- During the pandemic, people stayed at home are more likely to maintain connection via social media.

- Twitter has grown to be one of the most popular platform worldly. The Daily Active Users (DAU) of Twitter increased 20% and reached 199 million in the first quarter of 2021.

- Using the Twitter data , combining vaccination data the CDC to analyze the effect of public opinion towards vaccines.

- Predications of the relationship between vaccination and confirmed cases/deaths.

- Be specific, cases studies: Los Angeles, New York, Phoenix.

Georgia Tech

Related Work

Social Media and Epidemics Forecast

- Ali et al. 2017. Sentiment analysis as a service: a social media based sentiment analysis framework. In 2017 IEEE International Conference on Web Services (ICWS). IEEE, 660–667.

Methods to Analyze Sentiment in Twitter

- ZDrus & Khalid. 2019. Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science 161(2019), 707–714.

Vaccine Sentiment in Twitter

- Huet et al. 2021. Revealing public opinion towards COVID-19 vaccines with Twiter Data in the United States: aspatiotemporal perspective. medRxiv (2021).

- Satar & Arifuzzaman. 2021. COVID-19 Vaccination Awareness and Afermath: Public Sentiment Analysis on Twiter Data and Vaccinated Population Prediction in the USA. Applied Sciences 11 (6 2021). Issue 13.

- Naderi et al. 2021. COVID-19 Vaccine Hesitancy and Information Diffusion: An Agent-based Modeling Approach. (9 2021).

Machine Learning Model

- Bontempi et al. 2013. Machine Learning Strategies for Time Series Forecasting. (2013), 62–77.

- Gardner & Dorling. 1998. Artifcial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. Atmospheric Environment 32, 14 (1998), 2627–2636.

Georgia
Tech.

Data Source

Data Source

- Tweets Dataset – Panacea Lab
- National Vaccinations – CDC
- National Cases and Deaths – CDC

Case Studies

- Regional Vaccinations – CDC
- Regional Cases and Deaths – The New York Times

Data Process

1. Twitter Hydrator
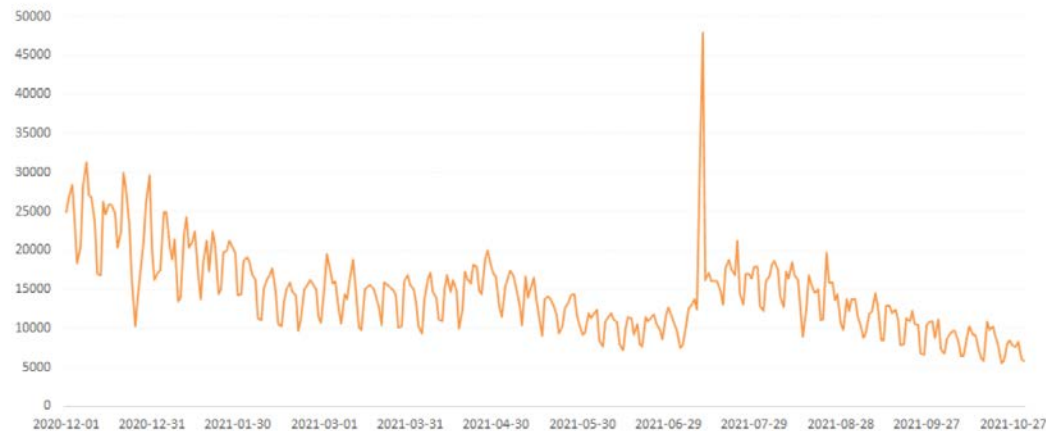The sample size fluctuates between 5625 and 48006, with an average of 14558.

Figure 1: Number of Final Selected Tweets On Each Day

2. Carmen Geolocation
50.46% of the tweets are tagged with geolocation, which is with 0.06% in the original Twitter.
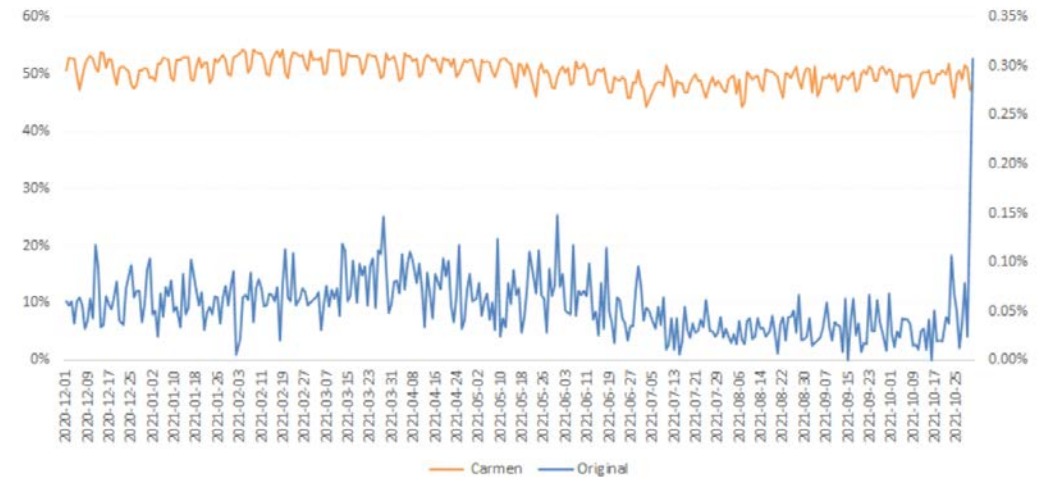

Figure 2: The Proportion of Tweets with Geolocation

3. Tweets Features
- Tweet feature: retweet counts, favorites counts of each tweet
- Users feature: followers counts, friends counts, user favorites

Method --- Sentiment Analysis

Machine Learning Method: Naive Bayes

Model Assumption

- Words position does not matter and model relies on a simple representation of document (bag of words).
- Documents are represented by features.
- The feature probabilities are independent given the class.

Training Datasets

- TextBlob
  - ➤ TextBlob's default sentiment analysis is trained on customer reviews hand-tagged with values for polarity and subjectivity.
  - ➤ Another option in TextBlob is NaiveBayesAnalyzer trained on movie reviews associated with positive or negative rating scores.
- Sentiment140
  - ➤ The dataset contains about 1.6 million tweets collected through keyword search and annotated automatically by detecting emoticons.
  - ➤ Tweets are determined to have positive, neutral, or negative sentiment.

$$c* = arg \max_{c} P_{NB}(c|d)$$

$$= arg \max_{c} \frac{P(c)(d|c)}{P(d)}$$

$$= arg \max_{c} P(d|c)P(c)$$

$$= arg \max_{c} P(x_1, \ldots, x_m|c)P(c)$$

$$= arg \max_{c} p(c) \prod_{i} P(x_i|c),$$

Georgia Tech

Method --- Multilayer Perceptron

Model Parameters

- MLP: Set three hidden layer and use Rectifed Linear Unit (ReLU) as the activation function.

- Dataloader: batchsize: 64

- Optimizer: Use Adam as the optimize and use MSELoss to calculate loss.
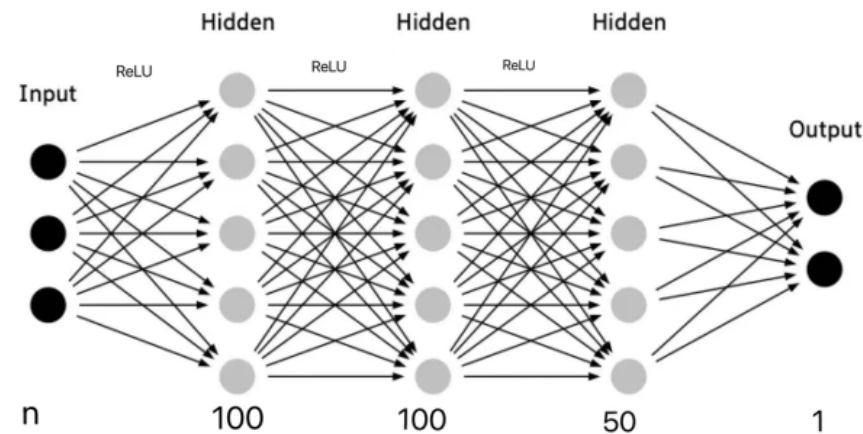
- Learning rate: 0.01.

- Epoch: 1000.

Figure 3: MLP

Results: Vaccine Sentiment

- The levels of sentiment140 and VADER show consistency with correlation of 0.4196.

- The number of vaccine tweets can also reflect vaccine sentiment.

- It is quite likely that sentiment140 and VADER are more reasonable and believable than textblob in vaccine sentiment estimate.

- To avoid bias from random samples, randomly picked 5% tweets from the total datasets and compare the sentiment estimate with that of 10% sample.
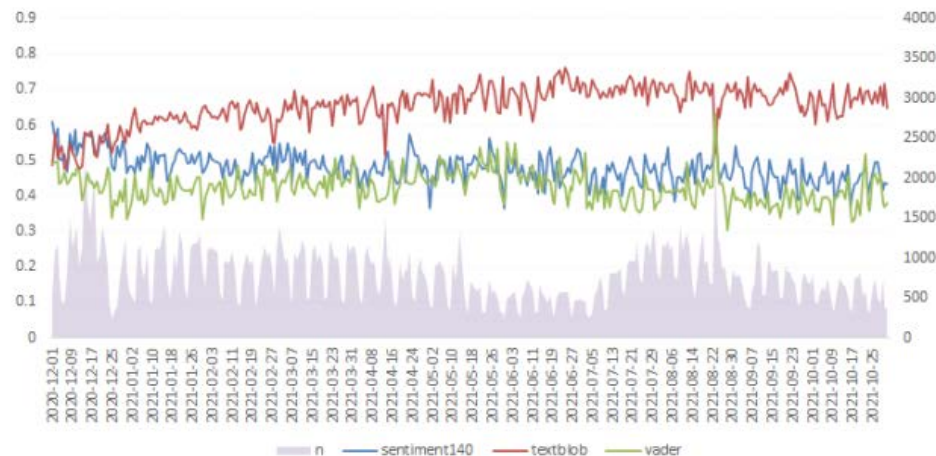

Figure 4: Daily Sentiment Index

Table 1: Correlation of Vaccine Sentiment of Three Methods

correlation	sentiment140	textblob	vader	n
sentiment140	1	-0.3873	0.4196	0.4711
textblob	-0.3873	1	-0.0785	-0.2751
vader	0.4196	-0.0785	1	0.1405
n	0.4711	-0.2751	0.1405	1

Table 2: Vaccine Sentiment Comparison of Two Samples

feature	5%	10%	Pearson	Spearman
sentiment140	0.4805	0.4798	0.7296	0.7060
textblob	0.6588	0.6553	0.8608	0.8067
vader	0.4254	0.4240	0.7329	0.7232
n	394.5493	791.0448	0.9907	0.9893

Georgia Tech

Results: Twitter Features

- Twitter features shows correlation with pandemic.

Table 3: Statistical Data of Features

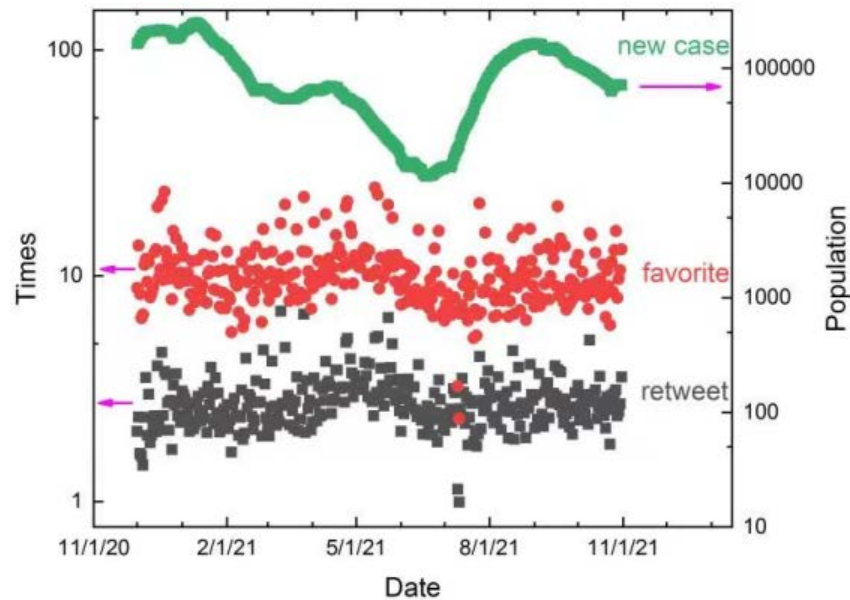| feature	retweet	favorite	followers	friends	user favorite
mean	2.82	10.37	87941.68	2007.45	24765.38
median	2.69	9.66	86321.76	2031.85	24527.52

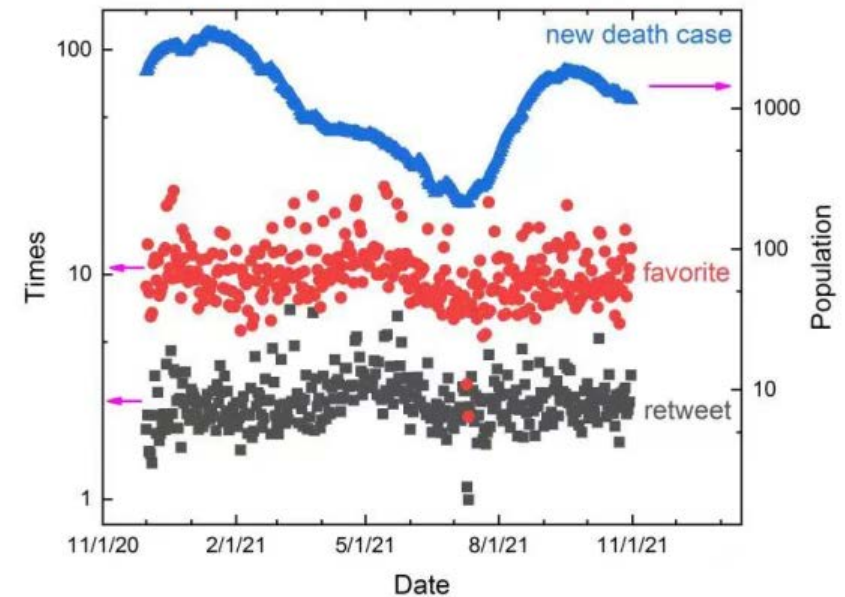**Figure 5: The Correlation among Retweet, Favorite and New Cases**

Figure 6: The Correlation among Retweet, Favorite and New Deaths

Results: Twitter VS Vaccination

- Cumulative vaccination in the U.S. is negatively related to vaccine sentiment.
- The daily vaccination is **positively** related to vaccine sentiment.
- The feature **'followers'** is also positively related to the daily vaccination, which implies that tweets posted by more influential users can cause more daily vaccination.
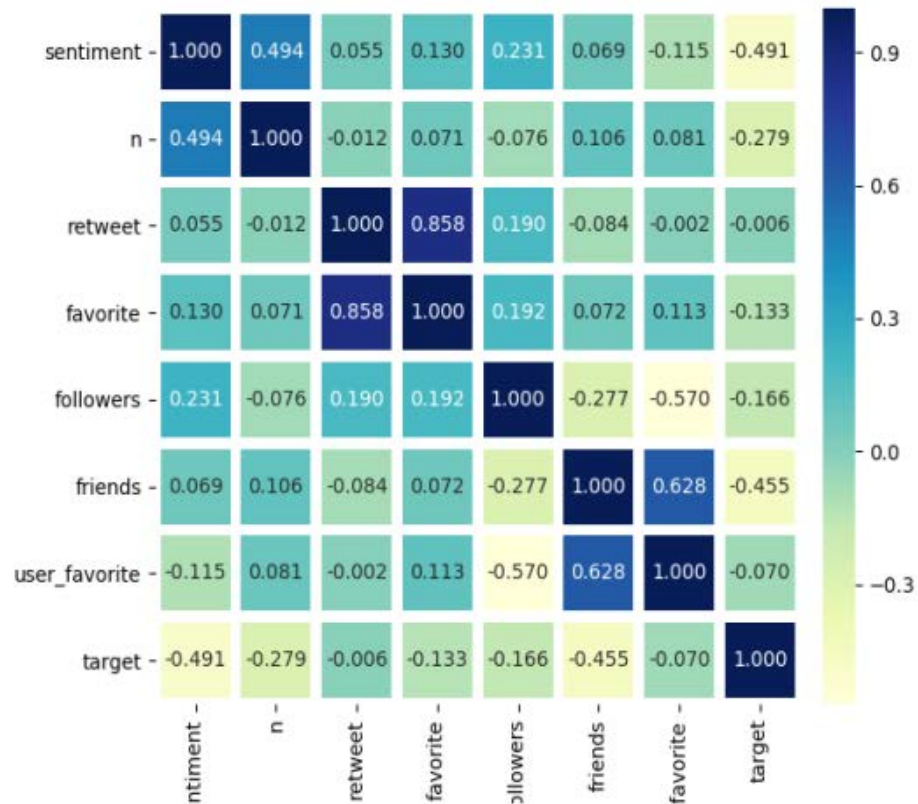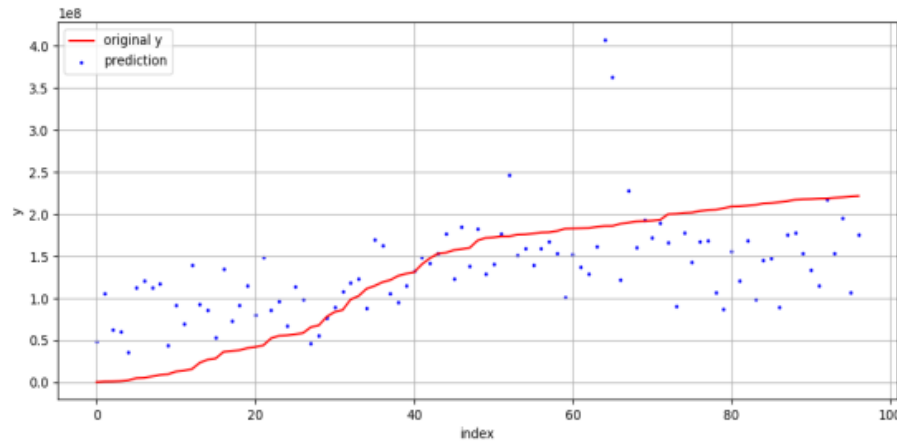

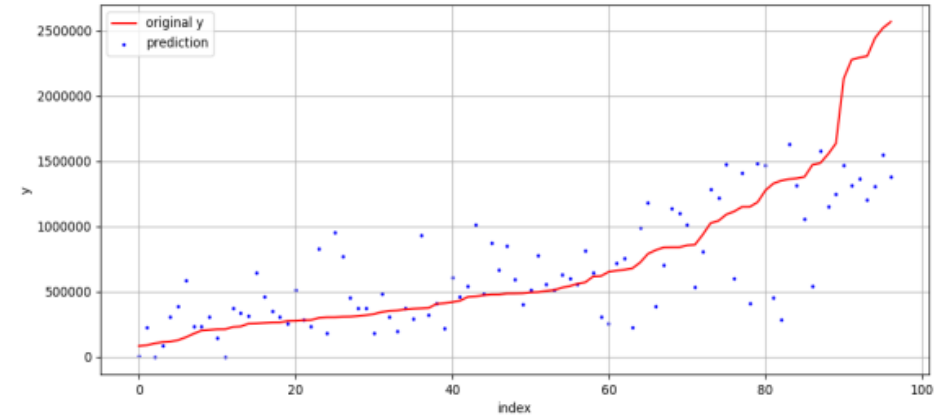(a) Heatmap of Twitter Feature and Vaccination (Cumulative)
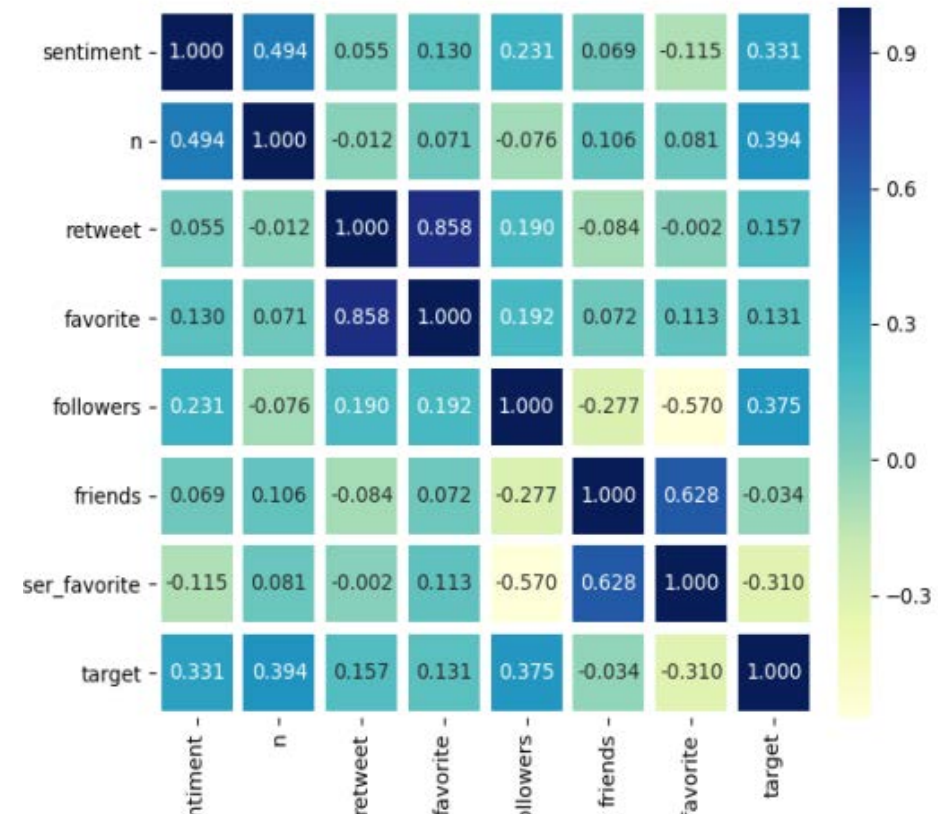
(b) Heatmap of Twitter Feature and Vaccination (Daily)

Results: Twitter VS Vaccination

- The model based on vaccine sentiment and twitter features predict the vaccination data well.

(a) Twitter and Vaccination (Cumulative)


(b) Twitter and Vaccination (Daily)

- Sentiment140 is more useful to reflect vaccine sentiment in Twiter with overall lower prediction loss, especially for daily vaccination.

**Table 4: Loss of Prediction Model for Vaccination**

Types	Percentage	Sentiment140	textblob	VADER
Cumulative	10%	$2.38*10^{15}$	$2.01*10^{15}$	$2.66*10^{15}$
	5%	$2.74*10^{15}$	$2.15*10^{15}$	$2.89*10^{15}$
Daily	10%	$1.25*10^{11}$	$1.68*10^{11}$	$1.75*10^{11}$
	5%	$1.30*10^{11}$	$2.01*10^{11}$	$8.02*10^{10}$

Georgia Tech

Results: Vaccination VS. Cases/Deaths

- Cumulative number of people receiving 1 or more doses is negatively related to cases and deaths.
- Negative correlation is more significant for deaths than cases.
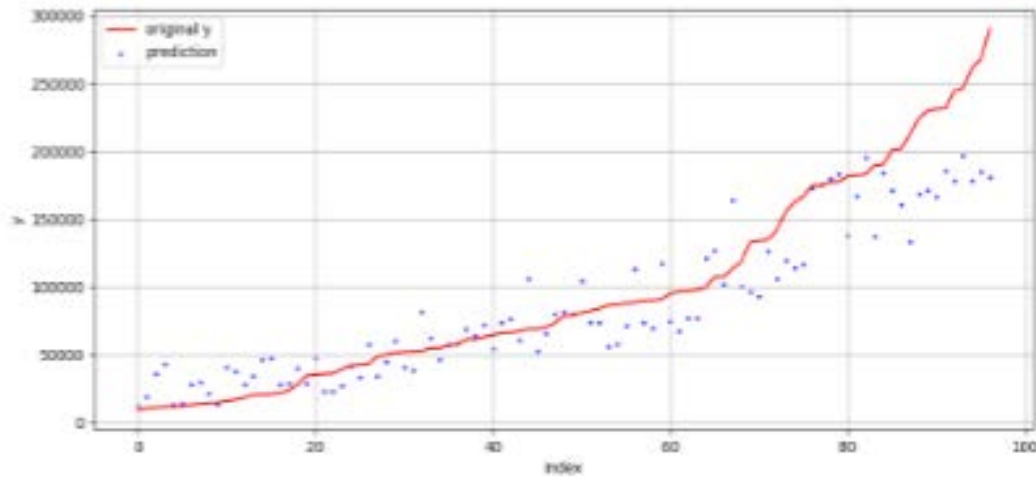- Vaccination plays more role in preventing deaths.
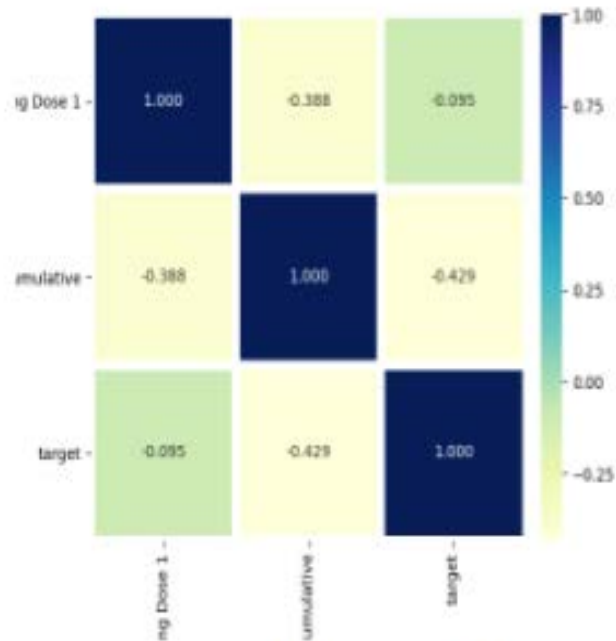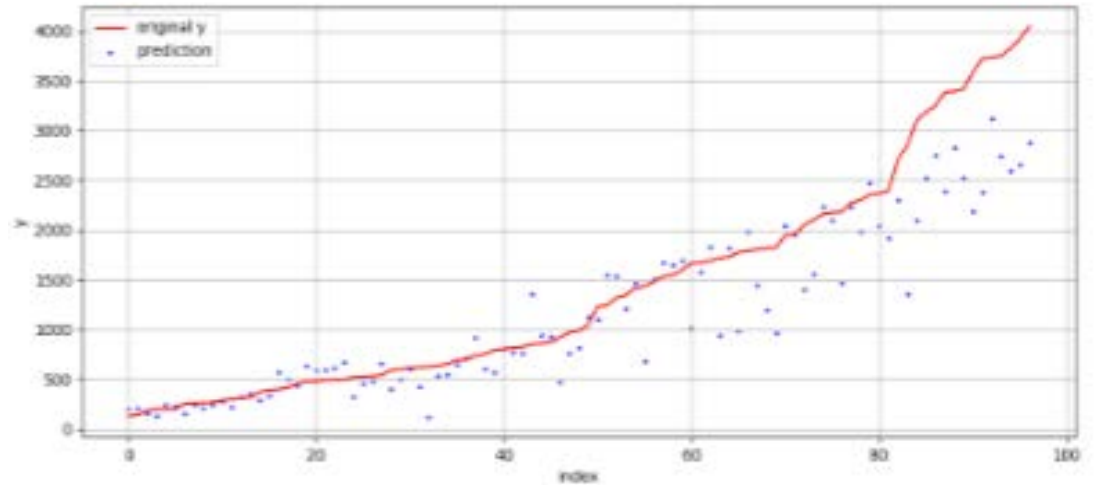

(a) Heatmap of Vaccination and Cases
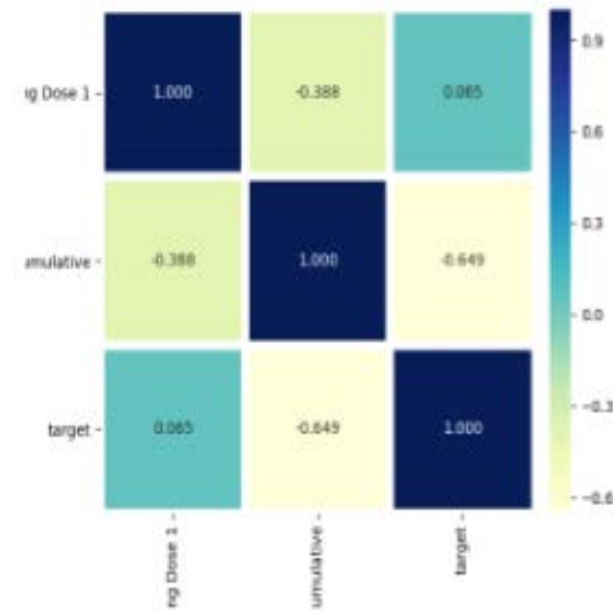
(b) Heatmap of Vaccination and Deaths

Results: Vaccination VS. Cases/Deaths

- The prediction value based on vaccination fits in well with cases and deaths.
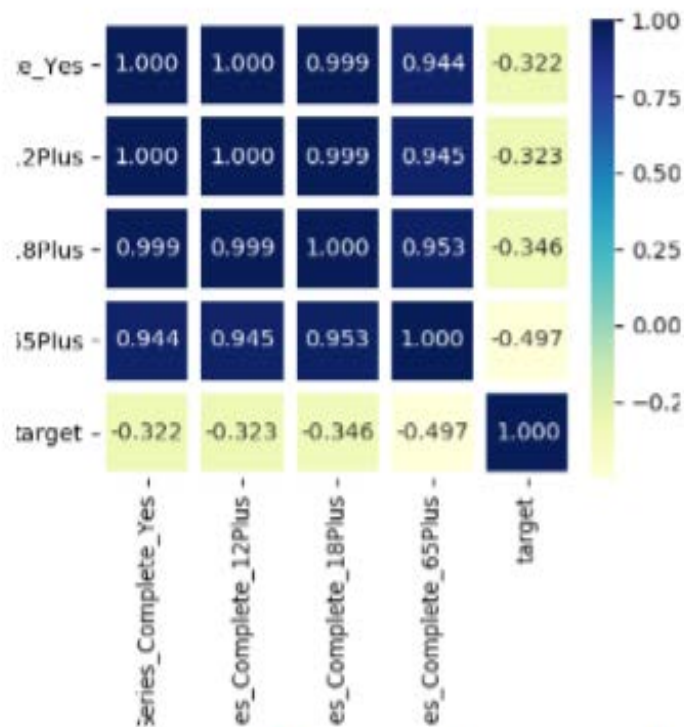

(a) Vaccination and Cases
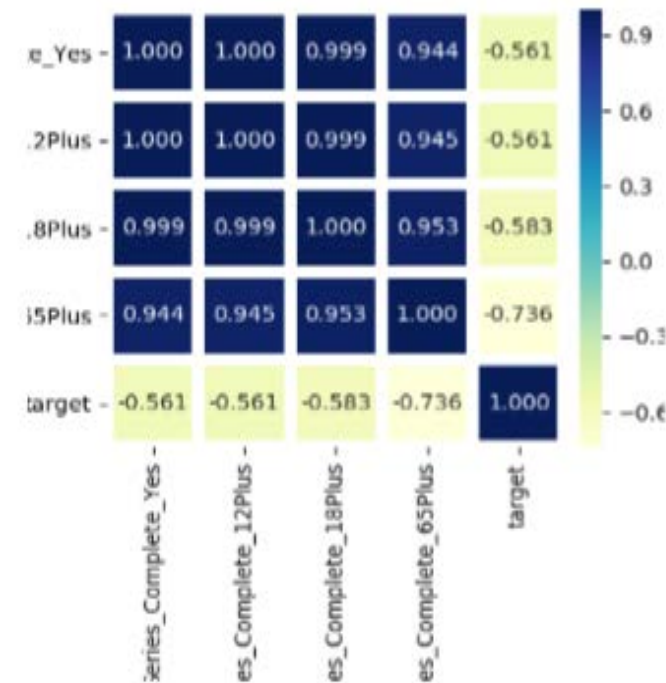
(b) Vaccination and Deaths

Result: Case Study

Los Angeles

- Vaccinations of all age-ranges people have negative relationships with cases and deaths.
- This negative correlation is increasingly significant with increasing age.
- Vaccination coverage in aged people is of more significance to control the pandemic.
- The negative correlation is more significant for deaths than cases.

(a) Heatmap of Vaccination and Cases

(b) Heatmap of Vaccination and Deaths

Result: Case Study

- Models cannot predict vaccination well, because it is only based on vaccine sentiment.

- Compared with previous results, others features of tweets improve the power of the models.

- The model based on vaccination can predict the cases and deaths in Los Angeles well.

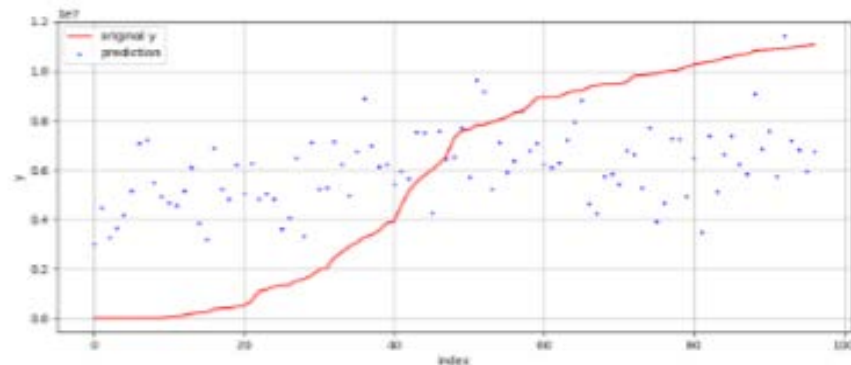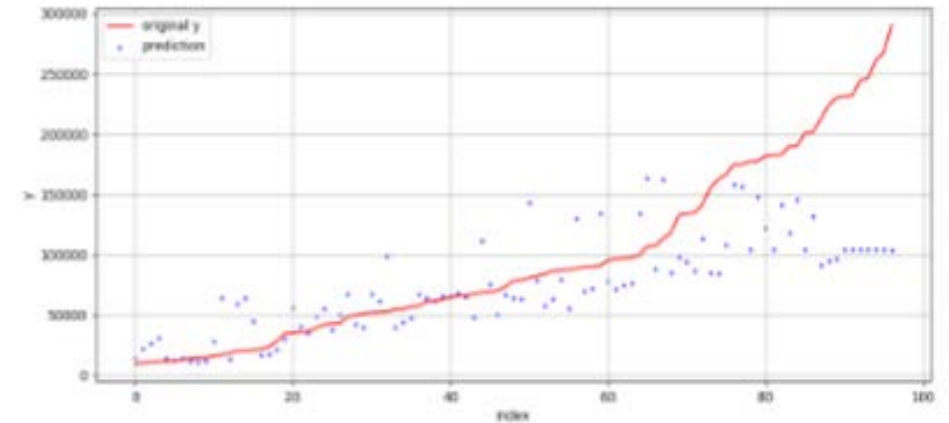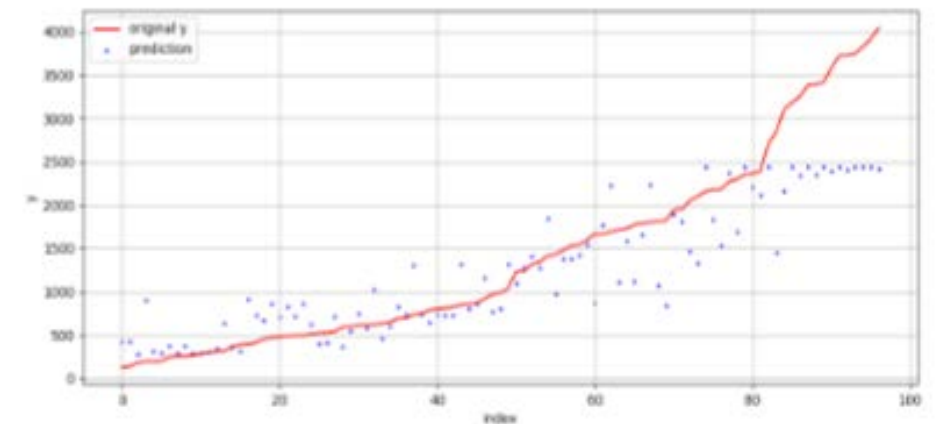- Similar results can also be obtained in New York and Phoenix.

(a) Vaccination and Cases (Los Angeles)

Figure 12: Sentiment and Vaccination

(b) Vaccination and Deaths (Los Angeles)

Conclusion

- Sentiment of vaccine extracted from Twitter can positively predict daily vaccination.

- Vaccination can significantly reduce the increase of cases and deaths, Moreover, vaccine is more effective in preventing death than reducing the risk of infection.

- Regional level: Effect of vaccine can be more obviously observed in aged group in term of cases and deaths.

- Suggestions: Government should use all feasible methods (e.g. make full use of social media advertisement) to dispel public doubts and increase the vaccination rate, since improving vaccination rate is a reasonable way to tackle COVID-19.

Georgia Tech

Thank you!