ONLINE SENTIMENT AND REACTIONS TO POSTS ON FACEBOOK: LEADING INDICATORS FOR STATE-LEVEL COVID-19 CONFIRMED CASES

Rong-Ching Chang, Chun-Ming Lai, Chu-Hsing Lin & Kai-Chih Pai
Department of Computer Science
Tunghai University
Taichung, Taiwan
{g08350003, cmlai, chlin, kcpai}@thu.edu.tw

ABSTRACT

Delays in the reporting of officially confirmed novel Coronavirus (COVID-19) cases have slowed down the process of providing essential resource allocations and policy decisions to counter the spread of the virus. Social media data have been found useful in the early detection and surveillance of infectious diseases, such as MERS-CoV, influenza, and mosquito-borne illnesses. Since very few studies examine the relationship between online sentiment and the number of local COVID-19 confirmed cases, in this study, we aim to bridge this gap. Choosing California as the location for the estimation, we collected public posts that used COVID-19 and California-related keywords from 25 January 2020 to 20 December 2020 from Facebook and evaluated the sentiments expressed in the responses. Using the Granger Causality Test, we estimated the relation of online sentiment and post reactions to the number of daily confirmed cases in California. The test has shown that online sentiment features and post reactions contain information that helps to predict the number of locally confirmed cases 24 to 70 days in advance.

1 INTRODUCTION

The novel Coronavirus (COVID-19) has caused more than 2.4 million deaths worldwide. With the lack of lab capacity to process mass testing, people have to wait several weeks to receive their COVID-19 test results in some cases. Such delays have resulted in latency in the official confirmation of cases, posing significant risks for potential spread and a threat to public health. Researches have shown that social media data may help control outbreaks and provide critical information for early disaster surveillance. A few research projects have used online sentiment analysis and social media data to detect and predict the number of future COVID-19 confirmed cases. Shen et al. (2020) sampled Weibo posts with a list of COVID-19 related symptoms and found reporting of symptoms online has a significant predictive power up to 14 days before the official statistics. Gharavi et al. (2020) observed a temporal lag between the number of reports of symptoms on Twitter and officially-reported positive cases.

Yet, we have not seen a study that navigates the relationship between public online sentiment from the social network platform, and the number of COVID-19 confirmed cases. Our research goal is to investigate whether large-scale collections of public sentiments (section 3.3) and post responses on Facebook are a suitable data source for predictive section sentiment analysis on the number of state-level Covid-19 confirmed cases (seciton 3.4). Further, we used LSTM to predict the future number of confirmed cases from the analysis of online sentiment (section 3.5).

2 RELATED WORK

Social media discussions using sentiment analysis for early discovery and alarm mechanisms related to infectious diseases such as mosquito-borne outbreaks were used in Jain & Kumar (2018). Choi

et al. (2017) used sentiment analysis to reveal timely information for MERS-CoV control. Culotta utilized tweets to identify influenza and its correlation to the number of officially confirmed cases using regression models. Szomszor et al. used Twitter to predict the course of the swine flu outbreak in 2009. Several studies have used sentiment analysis on the social media response to the COVID-19 outbreak Bhat et al. (2020), mask usage Yeung et al. (2020), and vaccine resistance (Lyu et al., 2020). The Granger causality test evaluates whether a one time series provides statistically significant information about the future value of the target time series. Using Granger causality, Smailović et al. found that sentiment polarity can indicate stock movement a few days ahead. The study conducted by Gherghina et al. (2020) used Granger-causality to measure the relationship between stock market return and Covid-19 outbreak.

3 METHODOLOGY

3.1 DATA

The data from public Facebook groups and pages were obtained using CrowdTangle (Team, 2020). We chose California, USA, as the geolocation target to study. We collected and sampled data from public Facebook groups and pages using the following keywords and conditions: Covid-19 or Covid, California or CA. The time frame ranged from 25 January 2020 to 20 December 2020. On a weekly basis, we collected data ranked by the popularity of the posts and 10,000 posts being the most collected per week. In total, we collected and sampled 19,521,124 posts. The numeric features were averaged by date. The features we included were:

- Over-performing Score: This score measures the post's interaction performance compared to similar posts on the same page or group in similar time frames.
- Message: The message is the descriptive body of the post that a user uploads or shares in a Facebook post. The polarity and subjectivity will be evaluated.
- The number of each reaction: The reactions tabulated include the number of the responses of "like", "love", "ha ha", "wow", "sad", "angry", and "care".
- Aggregation of reaction: This includes Total interaction, Likes at posting, Number of shares, the number of comments and total view.

The Covid-19 confirmed cases data was obtained from the New York Times (NYT) COVID-19 dataset[1]. We used US state-level data and filtered it for California State. The given data were in accumulated form. We calculated the daily number of newly confirmed cases within the studied time range.

3.2 PREPROCESSING

The preprocessing stages included converting the text to lower case, removing email, removing special characters, and stopping words using Spacy. Unknown or null values were filled with zeros to avoid errors during the computational process. We performed a language detection test using Spacy to filter non-English data. After the language filtering, a total of 555,078 data items were kept.

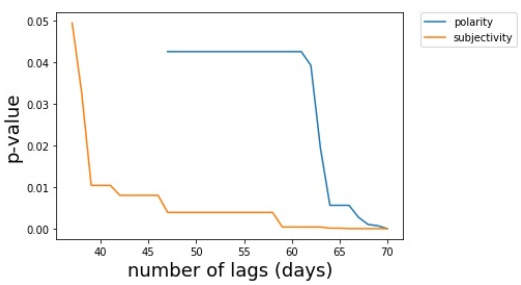
3.3 SENTIMENT ANALYSIS

With the collected data, we considered the "Message" section as the main text body. In sentiment analysis, the two major parts evaluated were polarity and subjectivity. Polarity is a score ranging from -1.0 to 1.0, with -1.0 representing negative emotion, 0.0 representing neutral emotion, and 1.0 representing positive emotion and sentiment. The Subjectivity range was from 0.0 to 1.0, with 0.0 being very objective and 1.0 being very subjective. Sentiment analysis was performed using TextBlob, where both polarity and subjectivity are measured in the given text.

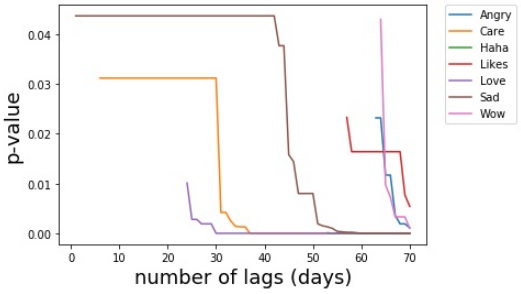
[1] https://github.com/nytimes/covid-19-datae

3.4 GRANGER CAUSALITY

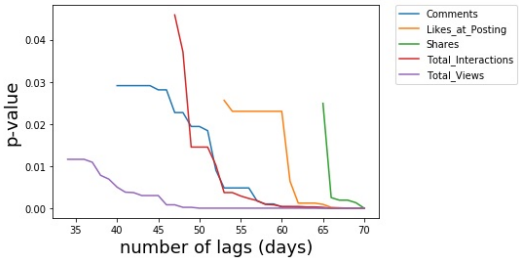
We first ran Dickey-Fuller tests to check if all of our data were stationary. The result suggested that "care" (p = .144), "wow" (p = .097), "like" at posting (p = .222), and total interactions (p = .120) were non-stationary. Having failed to reject the H0 hypothesis at the confidence level of 5%, we applied the difference scores instead of the original data to make our data stationary. We checked Granger Causality between all the possible combinations between sentiment feature time series and the target time series which is the number of local COVID-19 confirmed cases in California. As shown in figure 1 on the significant p-value of the Granger causality test, we obtained the result that, except for the over-performing score, most of our sentiment features from responses to public posts about COVID-19 in California Granger-caused the daily confirmed cases in COVID-19 in California State 24 to 70 days prior.



(a) Sentiment



(b) Reaction



(c) Aggregation of Reaction

Figure 1: Statistically significant Granger Causality results

3.5 LONG SHORT TERM MEMORY

After confirming that sentiment analysis can be a leading indicator of COVID-19 confirmed cases, we performed a multivariable time series prediction using LSTM with the sentiment features. For better performance, the data was rescaled using MinMaxScaler from sklearn. LSTM was implemented using Keras. The evaluation metric used was the Root Mean Squared Error (RMSE), which is an absolute error measure that squares the deviation. The many-to-many LSTM is implemented with two LSTM layers with dimensions 64, 32, followed by a dropout layer and a dense layer.

The comparative result is shown. After testing the different combinations of hyperparameters, we found the lowest RMSE of 1.419 when the training is conducted with 10 epochs; the number of future days we wanted to predict was 90 days on a rolling base. The n past is the number of past days we wanted to use to predict the future number.

4 DISCUSSION

In this study, we investigated the causal relationship between public sentiment on Facebook and the number of local COVID-19 confirmed cases. Based on the test results, we found that public sentiment on Facebook can be a leading indicator for the number of local COVID-19 cases for up to 59 days prior to official confirmation. The data obtained is limited to public pages and groups on Facebook, instead of private or personal opinion. The data may not and does not intend to

represent the opinions of the area's population. This study is solely conducted based on one platform, Facebook; whether the data from other platforms will yield similar results or findings is unknown. Also, the statistics on Covid-19 confirmed cases have vastly different sources with slightly different numbers, which may result in slightly different results during replication.

5 CONCLUSION

We used sentiment analysis and responses on public pages and group posts from Facebook to predict COVID-19 confirmed cases in California, USA. We have found that the sentiment and reactions to posts on public pages and groups from Facebook can be a leading factor in predicting state-level future confirmed cases 24 to 70 days in advance.

REFERENCES

Muzafar Bhat, Monisa Qadri, Majid Kundroo Noor-ul Asrar Beg, Naffi Ahanger, and Basant Agarwal. Sentiment analysis of social media response on the covid19 outbreak. *Brain, Behavior, and Immunity*, 2020.

Sungwoon Choi, Jangho Lee, Min-Gyu Kang, Hyeyoung Min, Yoon-Seok Chang, and Sungroh Yoon. Large-scale machine learning of media outlets for understanding public reactions to nationwide viral infection outbreaks. *Methods*, 129:50–59, 2017. ISSN 1046-2023.

Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pp. 115–122.

Erfaneh Gharavi, Neda Nazemi, and Faraz Dadgostari. Early outbreak detection for proactive crisis management using twitter data: Covid-19 a case study in the us. *arXiv preprint arXiv:2005.00475*, 2020.

Ștefan Cristian Gherghina, Daniel Ștefan Armeanu, and Camelia Cătălina Joldeș. Stock market reactions to covid-19 pandemic outbreak: quantitative evidence from ardl bounds tests and granger causality analysis. *International journal of environmental research and public health*, 17(18): 6729, 2020.

Vinay Kumar Jain and Shishir Kumar. Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *Journal of Computational Science*, 25:406–415, 2018. ISSN 1877-7503.

Hanjia Lyu, Wei Wu, Junda Wang, Viet Duong, Xiyang Zhang, and Jiebo Luo. Social media study of public opinions on potential covid-19 vaccines: Informing dissent, disparities, and dissemination. *arXiv preprint arXiv:2012.02165*, 2020.

Cuihua Shen, Anfan Chen, Chen Luo, Jingwen Zhang, Bo Feng, and Wang Liao. Using reports of symptoms and diagnoses on social media to predict covid-19 case counts in mainland china: Observational infoveillance study. *Journal of medical Internet research*, 22(5):e19421, 2020.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pp. 77–88. Springer.

Martin Szomszor, Patty Kostkova, and Ed De Quincey. swineflu: Twitter predicts swine flu outbreak in 2009. In *International conference on electronic healthcare*, pp. 18–26. Springer.

CrowdTangle Team. Crowdtangle. *Facebook, Menlo Park, California, United States*, 2020.

Neil Yeung, Jonathan Lai, and Jiebo Luo. Face off: Polarized public opinions on personal face mask usage during the covid-19 pandemic. *arXiv preprint arXiv:2011.00336*, 2020.