COVID-19 Sentiment Analysis Report

Zhengyang Xu[1], Yueyuan He[1], Wanyue Xiao[1], Chaoying Lyu[1]
School of Information Studies[1], Syracuse University, Syracuse, NY 13244 USA

Abstract

The outbreak of Corona Virus Disease 2019, or COVID-19, had significant consequences in terms of socio-economy, psychology, transportation, and politics, causing challenges in policy making, community mitigation, and risk management by political agencies, non-governmental organization, and world health organization. To gain more valuable understanding of public opinions and support drafting actionable policies, this study aims to find topics of English-language COVID-19 related tweets posted in April and by New York Governor during March to October respectively. This study also explores variations in how the COVID-19 related tweets, topics, and associated sentiments changed over a period of time.

1 Introduction

Since the diagnose of the 41 patients on 2 January 2020 in Wuhan, China (Huang *et al.*, 2020), the novel coronavirus has been confirmed in multiple countries in succession and infected more than 35 million patients, with above 12 million cases recorded specifically in United States by the end of October 2020 (Alwan *et al.*, 2020; CDC, 2020). To raise public awareness and reduce mortality, the U.S. government instituted a series of mandatory regulations, including lockdown, social distancing, self-quarantine, and mask wearing. Those policies, however, inevitably gave rise to other negative consequences regarding economic instability, work disruption, social isolation, and mental disorder (Fiorillo and Gorwood, 2020; Fitzpatrick *et al.*, 2020; Galea *et al.*, 2020; Gao *et al.*, 2020; Li *et al.*, 2020; Bradbury-Jones and Isham, 2020; Bonaccorsi *et al.*, 2020), which in turn cause demoralization and trust diminishment (Alwan *et al.*, 2020). Moreover, problems were compounded by religious beliefs, conspiracy beliefs, public aversion to restriction policies, medical resources shortage, and political issues (Betsch *et al.*, 2020; Weinberger-Litman *et al.*, 2020; Romer and Jamieson, 2020). Under such a situation, understanding public opinion and attempting to reach a consensus are more important than ever since the widespread disagreement on restrictions affects the effectiveness of regulation implementation significantly.

The proliferation of web-based social technologies, geo-located social network services, and the popularization of social media as a personal opinion sharing platform have enabled researchers to use social media as a data resource for public opinion mining and emotion detection with low latency. The approach using machine learning models and text mining techniques has been approved as feasible and authentic by substantial studies (Dahal *et al.*, 2019; Boon-Itt and Skunkan, 2020; Xue *et al.*, 2020; Yin *et al.*, 2020). Therefore, the combination of social media intelligence and Natural Language Processing should be harnessed to understand public responses, enhance sentiment awareness, and support decision making.

To understand public sentiments under COVID-19, this study aims to capture popular topics or opinions and their associated sentiment dynamics based on the massive social media posts on Twitter. Accordingly, three research questions had been proposed. Firstly, how could we extract and quantify the public sentiments over a period of time? Secondly, for states with high positive cases, which topics are mentioned in the highest frequency? The last one is how the sentiment varies during that period for topics obtained in question 2? This paper is organized as below. Section 2 consists of dataset acquisition, study progress illustration, model construction, and result comparison, following section 3 which interpretates the results of research question respectively. The last two sections, the limitations and conclusion will be presented.

2 Data and Method

2.1 Related Works

For sentiment analysis, either the traditional machine learning methods or the state-of-the-art deep learning methods had been applied in analyzing social media data. Although COVID-19 still remains as a relatively novel disease in human history, abundant researches resources related with text classification and sentiment categorization are available online. Wang *et al.* (2020) applied BERT on Chinese social media in for sentiment analysis and mode, achieving a better performance than classic machine learning methods. Deep learning methods also had been used to conduct cross-language sentiment analysis (Kruspe *et al.*, 2020) and cross-cultural emotion detection (Imran *et al.*, 2020).

Topic modeling had been widely used to gain a descriptive understating of unstructured data in social science research (Schwartz *et al.*, 2013). As a popular topic modeling algorithm, Latent Dirichlet Allocation (LDA) can be used to identify patterns, themes, texts structures, and interconnection among themes, enabling researchers to efficiently categorize the large amount of data. In the instances of using LDA on COVID-19 analysis, a majority of the researches obtained satisfactory results for topic mining and segmentation, proving the method validity (Paul and Dredze, 2014; Xue *et al.*, 2020).

2.2 Dataset Acquisition and Preprocessing

For the data acquisition, the tweets collection was scraped by using Python Tweepy with Twitter API 2.0. from March 29 to April 30 on a daily basis. Hashtags used to match COVID-19 related contents include but not limit to "#cornonavirus", "#coronavirusoutbreak", "#covid", "#covid19", and "#ihavecorona". 14,607,045 tweets (5.31G) had been downloaded, containing contents posted in multiple countries and in various languages. After location and language filtering, only 161,220 English written tweets originating exclusively in the United States remained. It is necessary to point out that tweets without specifying location have not been taken into account.

2.3 Study Design

A proposed framework and the utilized techniques for topic detection and sentiment dynamics identification would be introduced. After data acquisition, several data processing techniques, including emoji removing, stop-words filtering, and POS tagging matching, will be executed, following by model construction and topic modeling. For topic modeling, a LDA using Java-based Mallet software package (Song *et al.*, 2009) could be utilized to identify the most popular COVID-19 related topics (Boon-Itt and Skunkan, 2020) discussed in New York State and California. Besides, given that the existing researches proved that a LDA model is being capable of identifying semantic topic information from massive text automatically (Poria *et al.*, 2016; Han *et al.*, 2020; Liu *et al.*, 2020), the result of LDA analysis will, in turn, provides insightful topics for the sub-topics selections .

In the process of model development, for all algorithms used for training data, we mainly use transfer learning idea that used the existing annotating dataset called Sentiment Analysis 140 (Go, Richa and Lei, 2009) for supervised model tuning. Primarily, fasttext has been utilized as the baseline model due to its time-saving characteristic and accessibility. Next, two traditional machine learning models (which are Random Forest and XGBoost algorithms) and two deep learning models (which are LSTM structures and pertained BERT model) have been developed. To evaluate the model performances, three annotators had manually labeled 100 sample data extracted from the April collection, reaching an agreement Cohen's Kappa value of 0.665. Since sentiment scores could be mainly separated into three groups, which are positive, negative, and neutral, model evaluation Metrix including Accuracy, F1-score, precision, and recall, has been employed when it comes to the performance comparison. Finally, the model with best performance will be employed to predict the sentiment score on the April dataset, enabling us to solve the research questions which specified in section 3.1 and 3.2.

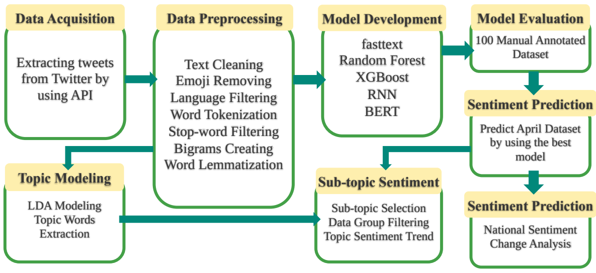

Figure 1: Overall Process of Study

2.4 Model Evaluation

For the performance evaluation of LDA model, coherence value, which measures the degree of semantic similarity between top associated words of topics, has been used to choose the best topic number. By setting multiple number of topics as input (which range from 5 to 70), the result indicates that having 35 as the topic number could yield a maximum coherence score of 0.39. According to table 1, the performance of fastText on the sample data is unsatisfactory, achieving merely 0.5677 in terms of precision and recall. XGBoost has a relatively higher accuracy (which is 0.577) and F1 score (0.575). Random Forest has slightly better performance, having 0.610 in accuracy and a F1 score that is up to 0.581. As the model with the second highest accuracy rate, LSTM architecture generate an accuracy rate of 0.645 and a F1 score of 0.6246 while BERT with 0.702 in accuracy and 0.681 in F1 score outperformed the others.

Model Name	Accuracy	Precision	Recall	F1
fastText	NA	0.5677	0.5677	NA
Random Forest	0.610	0.60	0.589	0.581
XGBoost	0.577	0.583	0.577	0.575
RNN	0.6453	0.6399	0.6275	0.6246
BERT	0.7021	0.714	0.691	0.6811

Table 1: Model Evaluation

3 Result Interpretation

As stated in table 1, BERT was perceived as the optimal model. Therefore, the fine-tuned BERT model has been employed to predict all April twitter data and generate a sentiment score for each tweet included inside.

3.1 National Sentiment Prediction

By aggregating the sentiment scores of tweets posted at the same day, we could quantify the sentiment changes through the period of March 29 to April 30. Next, combining with the timeline information of COVID-19 related events originated from NBC News (2020), a sentiment dynamics time series plot could be depicted (figure 2). To provide a better comparison result, we classified the results into positive and non-positive. Accordingly, with the evolution of COVID-19, the positive line was constantly stay below the non-positive line, indicating that titter users hold a negative attitude toward the epidemic. However, the gap between those two lines gradually diminished, suggesting that people's feelings became more positive even though more people were getting infected. In addition, the downward trend observed in two lines indicated that user prone not to discuss COVID-19 on Twitter compared the early stage.

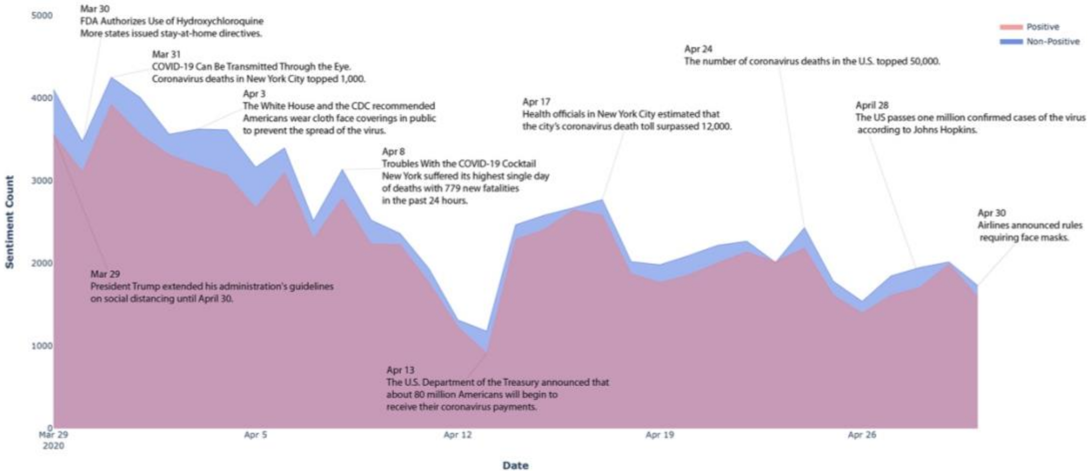

Figure 2 The Sentiment Analysis of the U.S.A. Tweets Comments During April

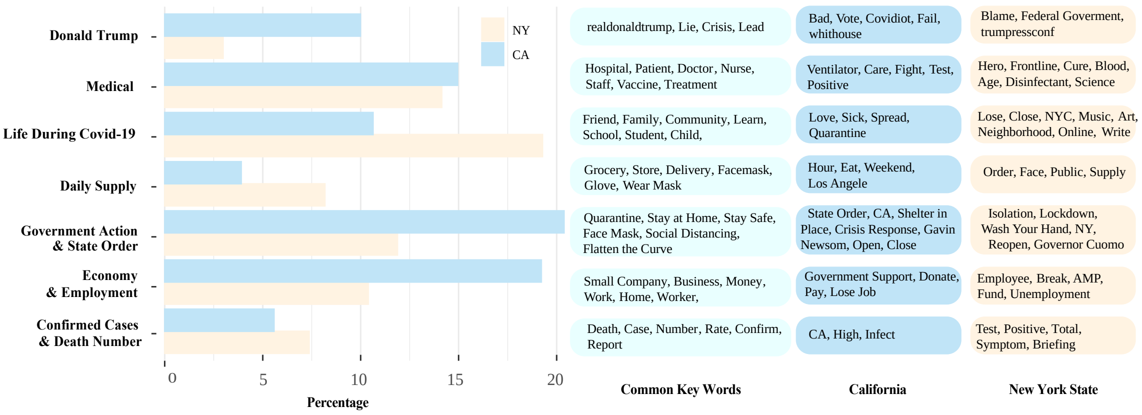

Figure 3. Themes (by %) concluded by 35 topics in California and New York, with the top associated words

3.2 Topic Extraction

Based on the probabilities included in the document list, 35 topics were generalized into seven categories after merging similar topics and eliminating irrelevant topics (Figure 3). For New York States, of the 21,164 tweets, 19.28% of the contents pertained to the theme of Life During COVID-19, followed by the approbation toward medical staff and concerns toward disease (14.14%), discussion on state policy (11.88%), the impact of COVID-19 on the economy and employment (10.39%), daily confirmed cases and death number (7.37%), daily supply (6.07%), and trump the president (2.21%). Compare with New York, people in California showed more concern on Policy related topics (which is 20.38%), followed by the impact on economy and employment (19.22%), hospital and medical issues (14.95%), daily life during the epidemic (10.63%), aversion toward Trump (9.99%), daily confirmed cases (4.14%), and daily supply (2.90%).

3.3 Sub-topic Level Sentiment Analysis

Based on the results obtained in section 3.2 and the ranking result of hashtags with a high frequency value, top seven hashtags are selected, including COVID related tages (e.g., covid, coronavirus), trump, quanrentine, job related tags (e.g., unemployement, work), vaccine, stayhome, and unemploy benefits. In the next step, we used rules and regex to find group each tweet belongs to, and then to predict sentiment score. Interestingly, scores related to Donald Trump are lower than any other frequent topics, except for "Unemploy Benefit" whose low value was caused by inadequate amount of data. Moreover, people hold a relatively positive attitude toward Stay-at-Home policy compared with Quarantine policy. One speculation is that people tend to espouse a voluntary order instead of a mandatory action.

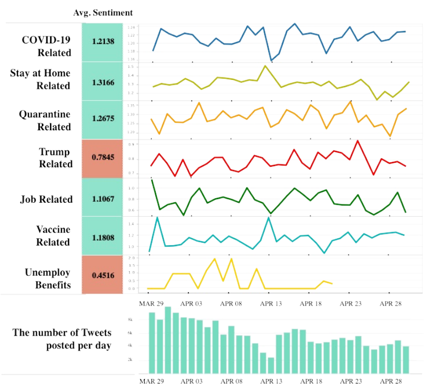

Figure 4. Topic level Sentiment Changes

4 Limitation

There are several other points should be considered in future studies. Firstly, the results may be biased to some extent due to limited model performance because of the fashion of emotion expression (E.g., using emoji) and the ambiguity of Internet words which were excluded from the proposed models. Secondly, the current method used for model training belongs to transfer method, indicating a potential issue of discrepancy between the training dataset and prediction. Additionally, our prediction data has larger range of content than training set. Hence, a majority of text, words, and language using in prediction data did not learned by models. Annotating our own training dataset in further studies can be a potential breakthrough to improve model performance and to gain better results. For topic modeling, the current analysis is based on static corpus, which has certain influences on reflecting the changing trend of topics discussed in real life. In further studies, Dynamic Topic Models could be conducted to detect topic evolution over fixed time intervals.

5 Conclusion

In this study, one found that different states showed disparate interest and emotion towards covid19 situation. People in California pay more attention on government action and economy while people in New York focus on daily life during COVID-19 and Medical related events. What's more, people showed more concerns for pandemic's impact on their personal life (such as losing jobs and unemployment benefits). The unemployment trend during COVID-19 outbreaks stress and bring huge negative emotion to people in twitter, and the unemployment benefits also gives people dramatic emotional change in April. Through examination of sentiment changes and trends with related topics and themes, the government agency, health organization, business industry could gain informative understanding of addressing issues aroused during pandemic period. And the way to use transfer methods predict social media sentiment can provide an efficient and urgent view of different groups people's emotion.

References

Alwan, N. A. *et al.* (2020) 'Scientific consensus on the COVID-19 pandemic: we need to act now', *The Lancet*, 396(10260), pp. e71–e72. doi: 10.1016/S0140-6736(20)32153-X.

Betsch, C. *et al.* (2020) 'Social and behavioral consequences of mask policies during the COVID-19 pandemic', *Proceedings of the National Academy of Sciences*, 117(36), pp. 21851–21853. doi: 10.1073/pnas.2011674117.

Bonaccorsi, G. *et al.* (2020) 'Economic and social consequences of human mobility restrictions under COVID-19', *Proceedings of the National Academy of Sciences*, 117(27), pp. 15530–15535. doi: 10.1073/pnas.2007658117.

Boon-Itt, S. and Skunkan, Y. (2020) 'Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study', *JMIR Public Health and Surveillance*, 6(4), p. e21978. doi: 10.2196/21978.

Bradbury-Jones, C. and Isham, L. (2020) 'The pandemic paradox: The consequences of COVID-19 on domestic violence', *Journal of Clinical Nursing*, 29(13–14), pp. 2047–2049. doi: 10.1111/jocn.15296.

CDC (2020) *CDC COVID Data Tracker*, *CDC COVID Data Tracker*. Available at: https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days (Accessed: 26 November 2020).

Coronavirus timeline: Tracking the critical moments of Covid-19 (no date). Available at: https://www.nbcnews.com/health/health-news/coronavirus-timeline-tracking-critical-moments-covid-19-n1154341 (Accessed: 29 November 2020).

Dahal, B., Kumar, S. A. P. and Li, Z. (2019) 'Topic modeling and sentiment analysis of global climate change tweets', *Social Network Analysis and Mining*, 9(1), p. 24. doi: 10.1007/s13278-019-0568-8.

Fiorillo, A. and Gorwood, P. (2020) 'The consequences of the COVID-19 pandemic on mental health and implications for clinical practice', *European Psychiatry*, 63(1), p. e32. doi: 10.1192/j.eurpsy.2020.35.

Fitzpatrick, K. M., Harris, C. and Drawve, G. (2020) 'Fear of COVID-19 and the mental health consequences in America.', *Psychological Trauma: Theory, Research, Practice, and Policy*, 12(S1), pp. S17–S21. doi: 10.1037/tra0000924.

Galea, S., Merchant, R. M. and Lurie, N. (2020) 'The Mental Health Consequences of COVID-19 and Physical Distancing: The Need for Prevention and Early Intervention', *JAMA Internal Medicine*, 180(6), p. 817. doi: 10.1001/jamainternmed.2020.1562.

Gao, J. *et al.* (2020) 'Mental health problems and social media exposure during COVID-19 outbreak', *PLOS ONE*. Edited by K. Hashimoto, 15(4), p. e0231924. doi: 10.1371/journal.pone.0231924.

Go Alec, Richa, B. and Lei, H. (2020) *Sentiment140 - A Twitter Sentiment Analysis Tool*, *Sentiment140*. Available at: http://help.sentiment140.com/home (Accessed: 29 November 2020).

Han, X. *et al.* (2020) 'Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China', *International Journal of Environmental Research and Public Health*, 17(8), p. 2788. doi: 10.3390/ijerph17082788.

Huang, C. *et al.* (2020) 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China', *The Lancet*, 395(10223), pp. 497–506. doi: 10.1016/S0140-6736(20)30183-5.

Imran, A. S. *et al.* (2020) 'Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets', *IEEE Access*, 8, pp. 181074–181090. doi: 10.1109/ACCESS.2020.3027350.

Kruspe, A. *et al.* (2020) 'Cross-language sentiment analysis of European Twitter messages duringthe COVID-19 pandemic', *arXiv:2008.12172 [cs, stat]*. Available at: http://arxiv.org/abs/2008.12172 (Accessed: 29 November 2020).

Li, S. *et al.* (2020) 'The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users', *International Journal of Environmental Research and Public Health*, 17(6), p. 2032. doi: 10.3390/ijerph17062032.

Liu, Q. *et al.* (2020) 'Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach', *Journal of Medical Internet Research*, 22(4), p. e19118. doi: 10.2196/19118.

NBC News (2020) *Coronavirus timeline: Tracking the critical moments of Covid-19, Coronavirus timeline: Tracking the critical moments of Covid-19*. Available at: https://www.nbcnews.com/health/health-news/coronavirus-timeline-tracking-critical-moments-covid-19-n1154341 (Accessed: 29 November 2020).

Paul, M. J. and Dredze, M. (2014) 'Discovering Health Topics in Social Media Using Topic Models', *PLoS ONE*. Edited by R. Lambiotte, 9(8), p. e103408. doi: 10.1371/journal.pone.0103408.

Poria, S. *et al.* (2016) 'Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis', in *2016 International Joint*

Conference on Neural Networks (IJCNN). 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada: IEEE, pp. 4465–4473. doi: 10.1109/IJCNN.2016.7727784.

Romer, D. and Jamieson, K. H. (2020) 'Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S.', *Social Science & Medicine*, 263, p. 113356. doi: 10.1016/j.socscimed.2020.113356.

Schwartz, H. A. *et al.* (2013) 'Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach', *PLoS ONE*. Edited by T. Preis, 8(9), p. e73791. doi: 10.1371/journal.pone.0073791.

Song, Y. *et al.* (2009) 'Topic and keyword re-ranking for LDA-based topic modeling', in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09. Proceeding of the 18th ACM conference*, Hong Kong, China: ACM Press, p. 1757. doi: 10.1145/1645953.1646223.

Wang, T. *et al.* (2020) 'COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model', *IEEE Access*, 8, pp. 138162–138169. doi: 10.1109/ACCESS.2020.3012595.

Weinberger-Litman, S. L. *et al.* (2020) 'A Look at the First Quarantined Community in the USA: Response of Religious Communal Organizations and Implications for Public Health During the COVID-19 Pandemic', *Journal of Religion and Health*, 59(5), pp. 2269–2282. doi: 10.1007/s10943-020-01064-x.

Xue, J. *et al.* (2020a) 'Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter', *PLOS ONE*. Edited by J. Zhao, 15(9), p. e0239441. doi: 10.1371/journal.pone.0239441.

Xue, J. *et al.* (2020b) 'Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter', *PLOS ONE*. Edited by J. Zhao, 15(9), p. e0239441. doi: 10.1371/journal.pone.0239441.

Yin, H., Yang, S. and Li, J. (2020) 'Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media', *arXiv:2007.02304 [cs]*. Available at: http://arxiv.org/abs/2007.02304 (Accessed: 26 November 2020).