PREDICTING CUMULATIVE COVID-19 INFECTIONS AND ANALYZING COVID-19 TWEET SENTIMENTS

UNDERSTANDING THE SKEPTICISM OF COVID-19 DATA AND INFORMATION

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Marina Kun

November 2, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS
Catherine D. Baritaud, Department of Engineering and Society
John A. Stankovic, Computer Science
Vicente Ordonez-Roman, Department of Computer Science

Beginning in December 2019, a new strand of coronavirus known as COVID-19 plagued the world with millions of infections and deaths. The virus transformed into a global health crisis which introduced further devastating effects to many aspects of society. To control the spread of COVID-19, world leaders have implemented preventative measures including social distancing and mask mandates. Researchers have collected a copious amount of data to offer information and analysis about the nature and course of the disease. In addition, news sources have played a significant role in facilitating the spread of information about the state of the pandemic to the general public. There are many facets to the process of producing and sharing information that contribute to the understanding of the pandemic. It is important to understand the nature of COVID-19 under multiple contexts to ensure the safety and wellbeing of communities across the globe.

The outbreak of the pandemic has significantly impacted the United States with over 8 million cases and 200,000 deaths (Centers for Disease Control and Prevention, 2020). The American people are also divided about the severity of COVID-19 and compliance with preventative measures. The technical project aims to study the nature of the pandemic in U.S counties by modeling the projection of cumulative COVID-19 infections and analyzing tweet sentiments regarding attitudes toward prevention guidelines. Tightly coupled with the technical project, the STS research focuses on understanding the factors of skepticism of COVID-19 data and information under social, political, and technical contexts. The STS research involves the application of Pinch and Bijker's (1987) Social Construction of Technology model to identify the relevant social groups and their contribution towards the interpretative flexibility of data and information related to the virus. The technical research is conducted by Computer Science and Systems Engineering student Morgan Freiberg and myself. In addition, the project

is overseen by Computer Science professors Vicente Ordonez and Jack Stankovic. Figure 1 below presents the timeline for the technical and STS research projects.

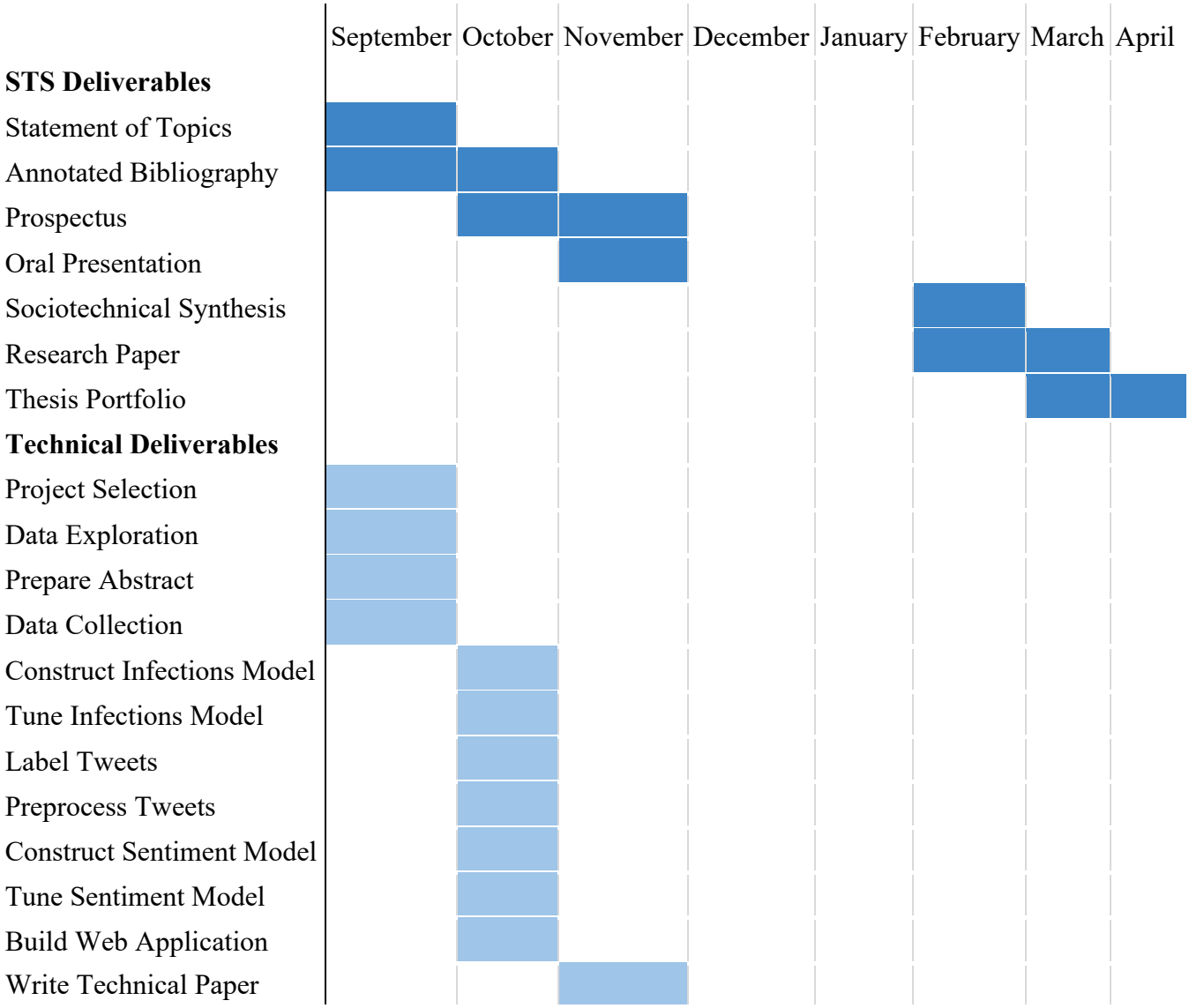

Figure 1: Timeline for STS and Technical projects. The Gantt chart shows the schedule for the STS and technical projects for the current and next semesters. (Kun, 2020)

PREDICTING CUMULATIVE COVID-19 INFECTIONS AND ANALYZING COVID-19 TWEET SENTIMENTS

The coronavirus disease is spread from one individual to another through respiratory droplets. The disease is highly contagious, so it must be closely monitored in order to control its spread. It is important to identify trends of growth or stagnation at a community level to inform individuals and local governments about the state of the virus in their area. Understanding the beginning trends of COVID-19 in a community can enable effective preventative measures that will help control the spread of the disease (Dalip & Deepika, 2020, p. 30). There are many machine learning and statistical approaches that predict and offer insight about the number of infections, deaths, and available resources. However, these accessible models and information can be widely misinterpreted by civilians and elected officials which may influence the dangerous spread of misinformation (Backhaus, 2020, p. 162). This technical project aims to mitigate the spread of COVID-19 by understanding the trends of the virus in counties of the United States with the construction of models that can forecast cumulative infections and analyze tweet sentiments related to COVID-19 prevention guidelines. The tweet sentiment analysis leverages natural language processing techniques to understand whether a tweet conveys support for the prevention guidelines. The information of the two models will be displayed in a clear and concise format that will prevent misinterpretation and further contribution towards the spread of misinformation.

The cumulative infections model intends to offer useful insights about the trends of the virus in a particular area of the United States. The initial steps to creating the cumulative infections model involves data collection. The project uses the open source COVID-19 dataset collected by The New York Times. This dataset includes live data for the cumulative number of

cases and deaths for U.S. counties. Medical researchers and computer scientists Benvenuto, Giovanetti, Vassallo, Angeletti, and Ciccozzi (2020) determined that the Auto Regressive Integrated Moving Average (ARIMA) model can be applied to analyze the prevalence and incidence trends of COVID-19 (pp. 2-3). The ARIMA model is a statistical analysis model that uses past time series data to predict future values. Expanding upon Benvenuto et al.'s findings, biomedical researchers Saleh, Ibrahim, and Ebrahim (2020) determined that the ARIMA model performs best out of other linear parametric models in predicting the spread of COVID-19 (p. 916). Thus, this project uses the ARIMA model to forecast the cumulative infections for each U.S. county. The optimal parameters for the model were determined for each county using a function that automatically finds the minimal information criterion values for an ARIMA model. The model was trained using 80% of the cumulative infections data for each U.S. county. The remaining 20% was allocated to the test set and used to evaluate the model.

The second process of the technical project conducts tweet sentiment analysis related to COVID-19 preventative measures for every county in the United States using natural language processing techniques. Although tweets do not represent an accurate sample population for an area, they provide some insight about the correlation between the county's aggregated sentiment on Twitter towards the preventative guidelines and the county's infection trend. At the Delhi Technological University in India, researchers Sethi, Pandey, Trar, and Soni (2020) created and compared multiple machine learning models that were designed to label the sentiment of a tweet related to the coronavirus pandemic. The researchers successfully constructed a Logistic Regression model to label tweets into two categories with around an 80% accuracy (p. 5). Logistic Regression uses the sigmoid function to provide probability scores in order to classify inputs to different classes (p. 4). This technical project will similarly analyze tweets using a

Logistic Regression classifier. The methodology for the tweet sentiment analysis is shown below in figure 2.

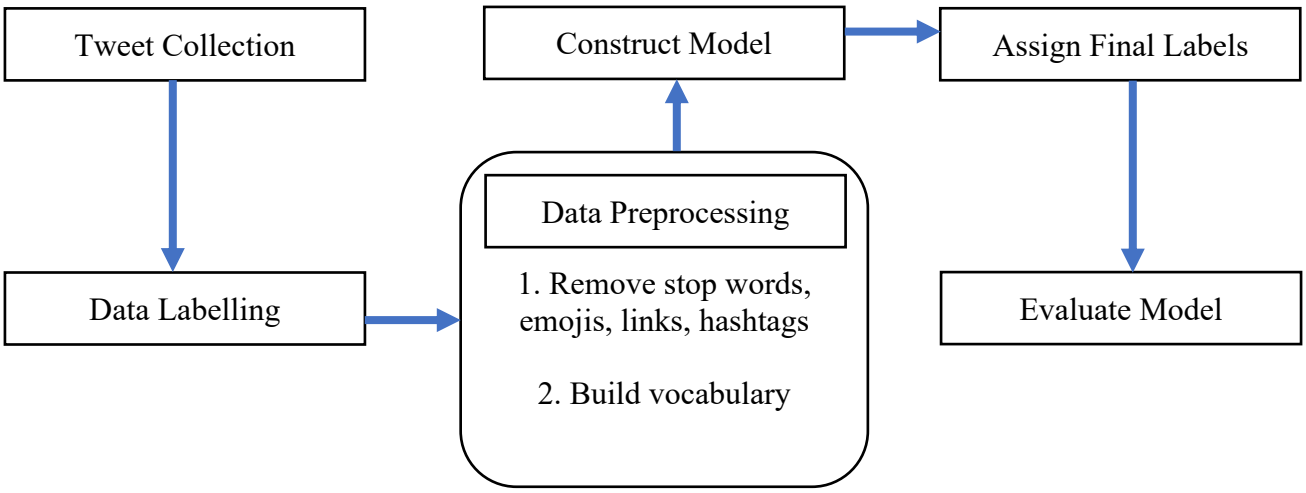

Figure 2: Tweet Sentiment Analysis Methodology. The flow chart describes the process for analyzing the sentiment of tweets regarding COVID-19 prevention guidelines. (Kun, 2020).

First, the tweets were collected from Twitter using hashtags related to preventative measures such as "WearAMask" and "MasksDontWork". Then, the tweets were labelled as "good" or "bad" to indicate whether they support the preventative guidelines. These labels were assigned based on the tweet's hashtag's association with supporting the guidelines (Hasan, Agu, & Rundensteiner, 2014, p. 2). For example, a tweet with the hashtag "WearAMask" would be labelled as "good", whereas a tweet with the hashtag "MasksDontWork" would be labelled "bad". After labelling, the text of the tweets was preprocessed by removing links, emojis, hashtags, and stop words. Stop words include very common English words, such as pronouns and articles, that do not contribute much to the analysis of the tweet sentiment. Finally, the vocabulary for the model was built using the bag-of-words representation, which involves the construction of a matrix with word counts (Ji, 2020). All the preprocessed texts and their corresponding labels were separated into training, testing, and validation sets. Every text from

each set was converted to a numeric vector with words counts, following the bag-of-words representation. Then, the Logistic Regression model was created to classify the tweets as “good” or “bad”. The model was trained using the training set and validated with the validation set. Once the model has been constructed, the analysis can be performed for each U.S. county. Geotagged tweets from the United States were retrieved using a dataset from the Institute of Electrical and Electronics Engineers which included U.S. tweet IDs related to COVID-19. The tweets will be sorted based on their affiliated counties and labelled using the constructed Logistic Regression model. The final labels of the text will be assigned based on its probability score. For instance, if a tweet has a probability score of .99 for the “good” label and .01 for the “bad” label, then the tweet will be labeled as “good”. However, if a tweet has insignificant probability scores of .5 and .5, the tweet will be labeled as “neutral”. Finally, the total number of tweets for each label of a county will be accumulated to calculate the percentage of tweets that support, do not support, and are neutral about COVID-19 preventative measures.

The final component of the technical project will summarize the results of the cumulative infections and twitter sentiment models through the creation of a web application. The application will allow users to view any U.S. county’s forecasted cumulative infections trends and the corresponding aggregated tweet analysis score. The app will explain the metrics of tweet analysis score and disclaim that it is not representative of the general population of that county. In addition, it will provide further detail about the statistical analysis and machine learning techniques used to develop the cumulative infection and tweet analysis models. The app will also report the accuracies of the two models. The cumulative infection and the tweet analysis models had validation accuracies of 87% and 89%, respectively. The entire technical project required no additional funding or physical resources. The models were developed on Google Colab, a cloud

based coding platform. Finally, the concluding paper will be written as a technical report which will include the methodologies, results, and discussion of this COVID-19 research project.

The technical project aims to provide well performing models that can help local governments and individuals make informed evaluations about the state of COVID-19 in a particular area in the United States. In addition, it hopes to produce a clear representation of the results in order to mitigate any possibilities of misinterpretation. Most of all, the research project hopes to contribute valuable information about the virus to help control the spread of infections in the United States.

UNDERSTANDING THE SKEPTICISM OF COVID-19 DATA AND INFORMATION

Since the advent of the COVID-19 pandemic, there has been a worldwide effort in tracking the virus which has led to a plethora of accessible data and information. Research regarding the virus provides different solutions and approaches to handling the pandemic. Furthermore, other than sharing factual information about the virus, news sources report conspiracy theories and anecdotes from social media (Shahsavari, Holur, Wang, Tangherlini, & Roychowdhury, 2020, p. 6). As a result of the numerous and diverse amount of information available regarding COVID-19, many people have different views about the scientific information and data analysis presented by experts. Further exacerbating the effects of the pandemic, there is a divide among many individuals and a growing controversy about the severity of the pandemic.

The technical project for modelling trends of COVID-19 is another addition to the growing number of solutions for tracking the virus. However, the project's results can become insignificant when users are skeptical about the information presented and the datasets used. So,

it is imperative that skepticism about COVID-19 data and information is fully understood in order to improve the technical project's clarity and effectiveness. This STS project aims to understand why people are skeptical towards COVID-19 data and information by analyzing the factors of skepticism defined by the roles of misinformation and data misinterpretation. Furthermore, the Social Construction of Technology (SCOT) model will be applied in order to identify relevant social groups and explore the interpretative flexibility of COVID-19 data and information. The SCOT model was selected for this analysis because it provides a holistic perspective on the various usage of the technical artifact through its emphasis on interpretative flexibility. In addition, the model can provide insight about how the multiple interpretations of COVID-19 data and information can lead to negative outcomes, including skepticism.

IDENTIFYING FACTORS OF SKEPTICISM TO IMPROVE COVID-19 PREVENTION GUIDELINES

Although the prolific amount of information and research can significantly contribute to controlling the spread of COVID-19, it can also produce side effects of misinformation which negatively impact the collective effort in combatting the virus. There is empirical evidence, provided by a multivariate model, that trusting scientific information can predict an individual's compliance with the COVID-19 guidelines (Plohl & Musil, 2020, p. 8). Thus, it is necessary to understand the reasoning behind skepticism in order to combat it and improve the compliance of the prevention guidelines. Factors of skepticism can be explored through analyzing the current views on the social, political, and technical contexts of data misinterpretation and the spread of misinformation. Figure 3 depicts the factors of data and information skepticism.

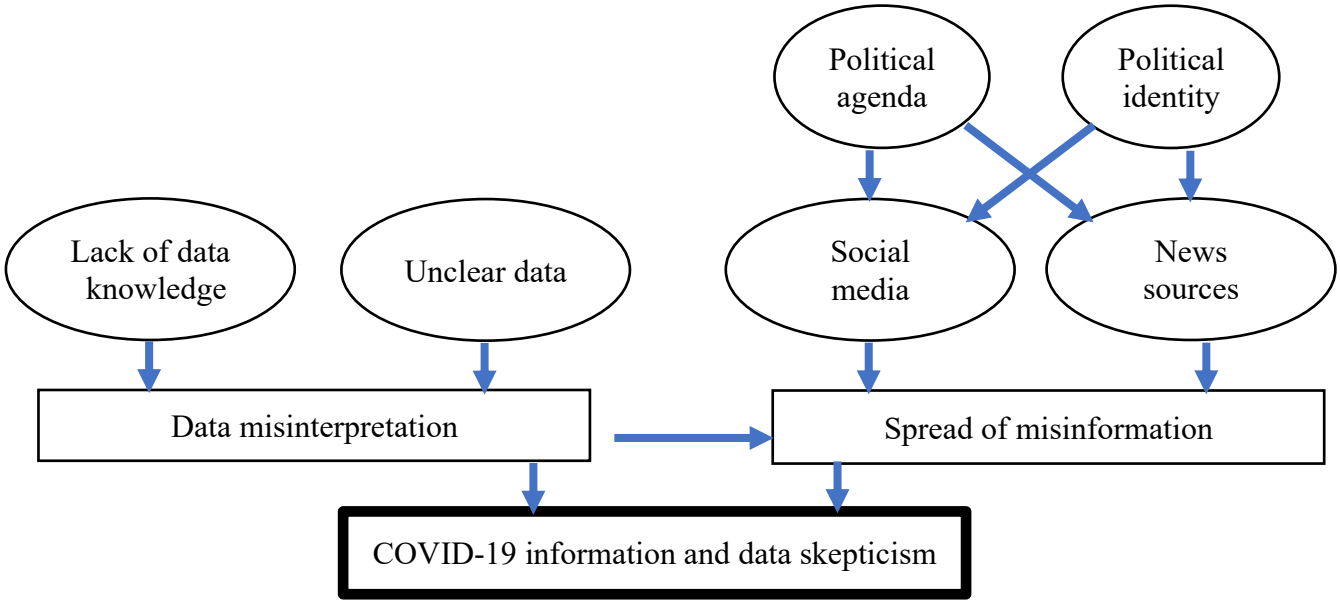

Figure 3: Factors of COVID-19 Information and Data Skepticism. The diagram visualizes the construction of COVID-19 data and information skepticism by exploring the agents of data misinterpretation and the spread of misinformation. (Kun, 2020).

Data misinterpretation and the subsequent spread of misinformation contribute directly to COVID-19 data skepticism. Misinformation is a growing concern across multiple scientific topics that is both easily and quickly distributed using social media. At New York University and Stony Brook University Jerit, Paulsen, and Tucker (2020) conclude that misinformation leads to distrust in experts and scientific evidence (p. 11). Furthermore, public health and infectious disease researchers Jaiswal, LoSchiavo, and Perlman (2020) attribute the political and social agendas of disinformation, misinformation, and mistrust in medicine as significant agents for the public health response to COVID-19. In addition, much of the information about COVID-19 involves basic statistical methods which can be widely misinterpreted. These statistical misinterpretations can mislead not only the general public but important policy makers (Backhaus, 2020, p. 163-164). A professor of epidemiology Tara Smith adds that it is important to consider multiple data sources rather than a single data point to avoid mischaracterizing the virus according to an individual's bias (as cited in Bosman, 2020, para. 30). The existence of misinformation and misinterpretation complicates the public's skepticism of COVID-19 data.

APPLYING THE SOCIAL CONSTRUCTION OF TECHNOLOGY MODEL TO ANALYZE USAGE OF COVID-19 DATA AND INFORMATION

The scientific research and data collection of COVID-19 can only be successful in controlling the spread of the virus if it is used and understood accurately by the relevant social groups of the Social Construction of Technology model. Figure 4 below shows the mapping of the situation between the relevant social groups and COVID-19 data and information.

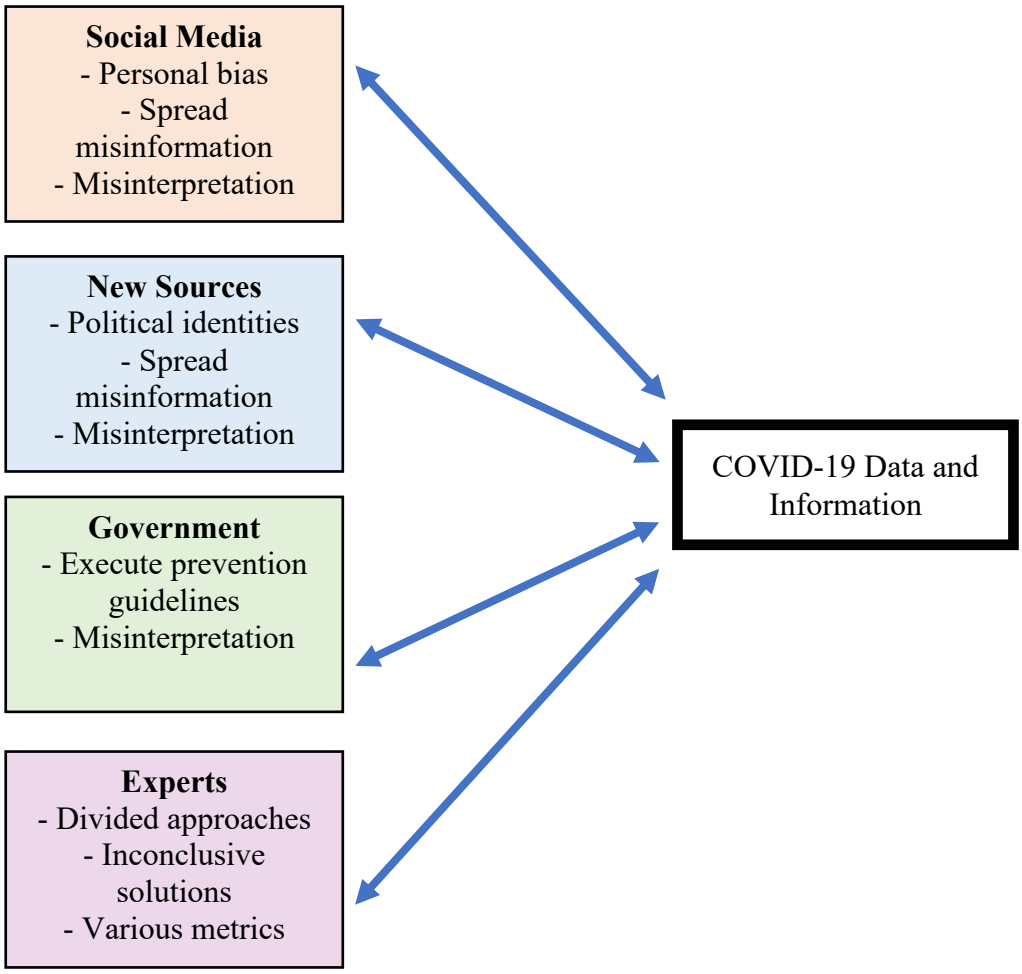

Figure 4: Social Construction of Technology Model for COVID-19 Data and Information. A depiction of the agents social media, news sources, governments, and experts who use and modify the technical artifact COVID-19 data and information. (Adapted by Kun (2020) from Carlson, 2007).

The model depicts the relationships between the technical artifact, COVID-19 data and information, and the key stakeholders which include governments, research experts, social media, and news sources. It also emphasizes the social groups' contributions to the interpretative flexibility of the artifact. The double arrows between the social groups and the technical artifact represents a constructive two-way relationship. The data and information presented to the social groups changes their understanding of the coronavirus which in turn adds to the development of the artifact. The bullets listed under each stakeholder describe the problems in their relationship with the technical artifact. The experts are the initial producers of the technical artifact. However, they can provide inconclusive and diverse results which lead to various perspectives and interpretations of the virus. Local governments must use the large and diverse amount of information to make educated decisions about controlling the spread of the virus. Social media users can misinterpret the artifact to support their personal bias and contribute to the spread of misinformation about the pandemic. Similarly, some news sources may be affiliated with a particular political party which influences them to report COVID-19 related information according to their political agendas (Jamieson & Albarracín, 2020, p. 5). Overall, the SCOT model highlights that interpretative flexibility can produce negative outcomes for the technical artifact through the facilitation of inaccurate interpretations. The model provides an understanding about how the interpretative flexibility of COVID-19 data and information can distort the true technical artifact with misinformation and data misinterpretation.

The technical project aims to provide a clear representation of the trends of COVID-19 in the United States. Most importantly, it explains the basic statistical analysis involved in the project which could help users accurately interpret the project's results as well as other data and information related to the virus. Figure 5 shows how the addition of the technical project to the

SCOT model in figure 4 will contribute to a better understanding of COVID-19 data and information.

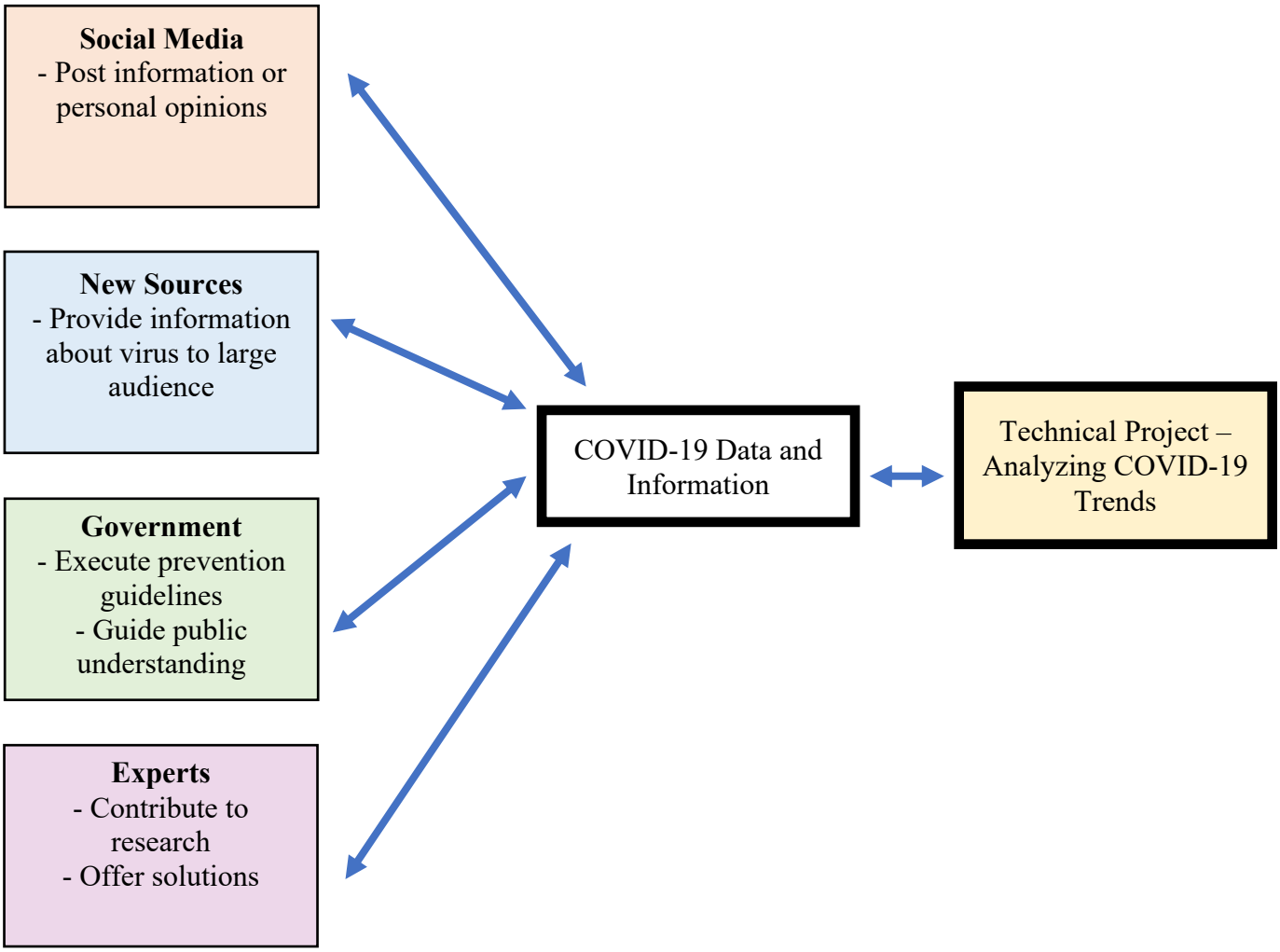

Figure 5: Social Construction of Technology Model for COVID-19 Data and Information and Technical Project Involving Analysis of COVID-19 Trends. A depiction of how the technical project can provide a positive contribution to the understanding of COVID-19 data information for the social groups social media, news sources, governments, and experts. (Adapted by Kun (2020) from Carlson, 2007).

The technical project serves as an additional technical artifact which will provide a positive contribution to the original artifact, COVID-19 data and information. The project will improve the interpretative flexibility of information about the coronavirus by preventing data misinterpretation and clarifying information to help social groups recognize misinformation from

accurate information. Consequently, experts can provide further research which can offer further solutions to the pandemic without easily misleading the public. In addition, social media users can post their views and information about the pandemic without facilitating the spread of misinformation. Similarly, news sources can present accurate information about COVID-19. Government officials can use the accurate data available to enforce effective preventative measures and share critical information to control the spread of the virus. With the positive contributions of the technical project, the interpretative flexibility of COVID-19 data and information becomes less diverse which will lead to a more accurate understanding of the virus.

This STS research will be written as a scholarly article that explores the factors of the coronavirus data and information skepticism through analyzing the social, technical, and political agents defined by the Social Construction of Technology model. Through this research process, the project hopes to characterize the spread of misinformation, data misinterpretation, and mistrust in science as well as explore the interpretive flexibility of COVID-19 data and information. The outcome of this STS analysis will lead to a better understanding about the spread of information related to the pandemic. Furthermore, it will provide valuable insight for representing and sharing accurate information which will ultimately combat illogical skepticism and improve the control of COVID-19.

CONTRIBUTING TO EFFORTS TO CONTROL THE SPREAD OF COVID-19

The COVID-19 pandemic has greatly impacted the world with dangerous health and economic concerns. The pandemic has also contributed to the growing discussion about trusting data and science. The STS topic explores the roles of misinformation, misinterpretation, and mistrust of science to characterize COVID-19 data and information skepticism. By understanding the skepticism behind data, it will help produce a better representation of the data analysis conducted in the technical project. The technical project serves to model the cumulative infections trend and analyze tweet sentiments regarding prevention guidelines. The two projects will improve the spread of accurate information and enable effective preventative measures towards controlling the spread of COVID-19.

WORKS CITED

Backhaus, A. (2020). Common pitfalls in the interpretation of COVID-19 data and statistics. *Intereconomics*, *55***,** 162–166. doi: 10.1007/s10272-020-0893-1

Benvenuto D., Giovanetti M., Vassallo L., Angeletti S., & Ciccozzi M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*, 1-4. doi: 10.1016/j.dib.2020.105340

Bosman, J. (2020, July 27). Hoping to understand the virus, everyone is parsing a mountain of data. *The New York Times*. Retrieved from https://www.nytimes.com/2020/07/27/us/coronavirus-data.html

Centers for Disease Control and Prevention. (2020, November 1). United States COVID-19 cases and deaths by state. In *CDC COVID data tracker*. Retrieved from https://covid.cdc.gov/covid-data-tracker/#cases_casesinlast7days

Dalip, & Deepika. (2020). AI-enabled framework to prevent COVID-19 from further spreading. In A. Joshi, N. Dey, K. Santosh (Eds.), *Intelligent systems and methods to combat Covid-19* (pp. 29-36). Singapore, Singapore: Springer Verlag.

Hasan, M., Agu, E., & Rundensteiner E. (2014). Using hashtags as labels for supervised learning of emotions in Twitter messages. Retrieved from http://web.cs.wpi.edu/~emmanuel/publications/PDFs/C25.pdf

Jaiswal, J., LoSchiavo, C., & Perlman, D. C. (2020). Disinformation, misinformation and inequality-driven mistrust in the time of COVID-19: Lessons unlearned from AIDS denialism. *AIDS and behavior*, *24*(10), 2776–2780. doi: 10.1007/s10461-020-02925-y

Jamieson K. H., & Albarracín D. (2020). The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US. *Harvard Kennedy School Misinformation Review*, *1*(2), 1-23. doi: 10.37016/mr-2020-012

Jerit J., Paulsen T., & Tucker J. A. (2020). Confident and skeptical: What science misinformation patterns can teach us about the COVID-19 pandemic. doi:10.2139/ssrn.3580430

Ji, Y. (2020). Bag-of-words representations [Lecture notes and audio file]. In Y. Ji, *Text classification (1): Logistic regression*. Retrieved from http://yangfengji.net/uva-nlp-course/slides/lecture-02.pdf

Kun, M. (2020). *Factors of COVID-19 information and data skepticism*. [3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Kun, M. (2020). *Social construction of technology model for COVID-19 data and information*. [4]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Kun, M. (2020). *Social construction of technology model for COVID-19 data and information and Technical Project Involving Analysis of COVID-19 Trends*. [5]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Kun, M. (2020). *Timeline for STS and technical projects*. [1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Kun, M. (2020). *Tweet sentiment analysis methodology*. [2]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Plohl, N., & Musil B. (2020). Modeling compliance with COVID-19 prevention guidelines: The critical role of trust in science. *Psychology, Health & Medicine*. doi:10.1080/13548506.2020.1772988

Saleh A. I, Ibrahim A. A., & Ebrahim A. A. (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. *Journal of Infection and Public Health*, *13*(7), 914-919. doi: 10.1016/j.jiph.2020.06.001

Sethi, M., Pandey, S., Trar, P., & Soni, P. (2020). Sentiment identification in COVID-19 specific tweets. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems: ICESC 2020*. doi:10.1109/icesc48915.2020.9155674

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T., & Roychowdhury, V. (2020). Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*. doi: 10.1007/s42001-020-00086-5