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

A Research Paper submitted to the Department of Computer Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

Team Member: Morgan Freiberg

By

Marina Kun

November 10, 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
John Stankovic, Department of Computer Science
Vicente Ordonez-Roman, Department of Computer Science

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

Marina Kun
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
mk3qa@virginia.edu

Morgan Freiberg
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
msf6wk@virginia.edu

ABSTRACT

The continual rise of COVID-19 cases globally has highlighted the need for modelling its spread at a more fine-grained level so communities can create measures that fit their specific needs.

We aim to create a web application that uses machine learning techniques and statistical analyses to predict cumulative infections for U.S. counties. In addition, the application will analyze the general sentiment about the pandemic for every county to provide more insight about the context for each projection. Previous research has found success in using statistical models, such as the Auto Regressive Integrated Moving Average (ARIMA) model, to predict global COVID-19 trends. Additionally, a Logistic Regression model was found to be successful in conducting sentiment analysis on tweets relating to COVID-19. Similarly, we used the ARIMA model to predict cumulative infections in U.S. counties with an 87% accuracy. Next, we applied a Logistic regression model to analyze COVID-19 related tweet sentiments with 89% accuracy. Understanding the infection trends and the social context in specific areas of the United States will help identify clusters and can influence local health measures to control the spread of COVID-19.

1 Introduction

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. 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 disease is highly contagious, so it must be closely monitored in order to control its spread. Researchers have collected a copious amount of data to offer information and analysis about the nature and course of the disease but the public is left to their own devices to sort through many opinions and news sources surrounding COVID-19. To control the spread of COVID-19, world leaders have implemented preventative measures including social distancing and mask mandates. 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). However, the public is divided about the severity of COVID-19 and their compliance with preventative measures. 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 at a county level in the United States with the construction of models that can forecast cumulative infections and analyze tweet sentiments related to COVID-19 prevention guidelines. The information of the two models will be displayed in a clear and concise format that will prevent misinterpretation and further contribute towards the spread of misinformation. To make these findings accessible to and easily understandable by the general public a web application will be created to display the results.

2 Background

In order to predict future cumulative case counts, an autoregressive integrated moving average (ARIMA) model was used. This statistical model is commonly used for time series analysis and makes predictions fully on previous

values of the predictand. Sentiment analysis was completed by applying scikit-learn's CountVectorizer to preprocessed tweets to produce a sparse representation of token counts that were then fed into a Logistic Regression model.

3 Related Works

Medical researchers and computer scientists Benvenuto, Giovanetti, Vassallo, Angeletti, and Ciccozzi (2020) applied the Auto Regressive Integrated Moving Average (ARIMA) model to analyze the prevalence and incidence trends of COVID-19 (pp. 2-3). The researchers implemented their research using the John Hopkins COVID-19 epidemiological dataset which includes cases worldwide. The global context of Benvenuto et al.'s research does not provide specific and sufficient information to analyze the impacts of the virus on smaller communities. 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). These researchers applied the model to forecast the spread of the disease within four weeks in Saudi Arabia. Although the scope of Saleh et al.'s project was much smaller than Benvenuto et al.'s, it still does not provide specific information to different areas of a country. Furthermore, their project is limited by the set length of time they are forecasting the spread of the virus. In addition, both groups of researchers Benvenuto et al. and Saleh et al. did not make their research accessible to the public's use. Computer scientists Sethi, Pandey, Trar, and Soni created a model that can classify tweet sentiments regarding COVID-19. They collected and preprocessed data from Twitter, extracted important tweet features, and scored each feature. Then, the researchers developed machine learning models including Logistic Regression to classify the tweets. The tweets were labelled as positive, neutral, or negative regarding their general sentiment towards the pandemic. A positive tweet indicates that the author of the tweet generally feels hopeful about the pandemic. The labels are not specific to particular aspects of the pandemic that can improve the spread of COVID-19.

4 System Design

4.1 Cumulative Prediction

The prediction of cumulative COVID-19 cases is based on data collected and aggregated by the New York Times, which details daily case counts by county. This data was then split into separate sets based on the Federal Information Processing Standard (FIPS) county codes, which is an identifier set by the Census Bureau and is unique to each county. In order to combat missing or insufficient data, time series interpolation was applied to each county data set and counties with few data points were filtered out. A thorough analysis of each county's data was not possible given the large number of models that needed to be created, thus presenting a unique challenge in selecting the best model for a given county with minimal data analysis. Ultimately, the models were selected using a stepwise auto-selection technique with mean squared error (MSE) as the deciding metric. If an ARIMA model was not able to be applied to a specific county due to the existence of seasonality in the data, the selected model for the county was defaulted to the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. A SARIMA model is an extension of ARIMA that explicitly models seasonality within data. Each model was then used in the creation of prediction charts to clearly display the previous data the model used to make predictions and the forecasted cumulative counts.

4.2 Sentiment Analysis

4.2.1 Training. In order to give a user a clearer picture of the state of COVID-19 in a specific region, the prediction of cumulative cases in a county was supported by an analysis of a region's sentiment regarding support or resistance to preventative measures. The training of a supervised sentiment analysis model requires a labeled dataset. This presented a difficult challenge as data mapping tweets to their sentiment regarding the measures to prevent the spread of COVID-19, such as wearing masks and social distancing, did not exist. Thus, a list of hashtags that were found to be largely indicative of a tweet's opinion regarding such measures was compiled following research conducted on the hashtag's use on Twitter. After gathering a large amount of tweets that included these hashtags, 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. The hashtag that was used to create a tweet's label was removed in order to prevent the model from creating a simple mapping from hashtag to label as the goal of the analysis was to identify a more general sentiment within the tweet itself so the model could be applied on tweets that did not have these hashtags. Finally, these preprocessed tweets were split into training and testing sets and a CountVectorizer was fit to the training

data in order to produce a vocabulary and construct a sparse representation of the counts of the tokenized data that a model could be trained on (Ji, 2020). Every text from each set was converted to a numeric vector with words counts, following the bag-of-words representation. A Logistic Regression model was then fit to the transformed training dataset and the testing set was used to validate the model based on the mean absolute percentage error (MAPE).

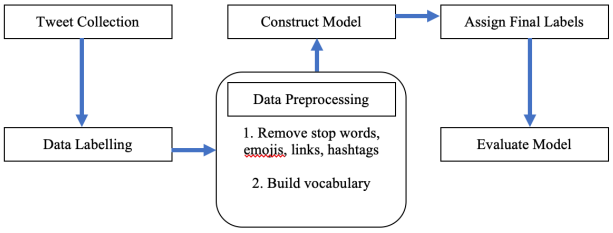

Figure 1: **End to end construction of Logistic Regression model to perform sentiment analysis.**

4.2.2 Predicting. To form predictions, a collection of geo-tagged, COVID-19 related tweet IDs created by a researcher at the School of Computer and Systems Sciences, JNU, New Delhi and available on the Institute of Electrical and Electronics Engineers website was used (Lamsal, 2020). These tweet IDs were rehydrated using the Twitter API and preprocessed in the same way as the training data, The CountVectorizer fit to the training dataset was used to convert the processed tweets to a numeric vector with words counts, following the bag-of-words representation before creating predictions with the Logistic Regression model. The handling of the output of the Logistic Regression model created a distinctive issue as not every COVID-19 related tweet discussed or provided a sentiment on the measures being explored. Thus, the model was used to predict the class probabilities of supporting or not supporting the measures and strict rules regarding the probabilities were applied in order to bin the predictions into supporting, not supporting, and neutral/unrelated categories. Simple summary statistics, such as the counts and proportion of labels within counties and states were calculated in order to relay this information to a user in a simple and understandable manner.

4.1 Web Application

The web application, built using React Native, compiles the deliverables of the cumulative predictions and twitter sentiment analysis into an easily digestible format for users. By specifying a county, a user gains access to the chart of previous and forecasted cumulative COVID-19 cases and summary statistics about the sentiment of available tweets regarding the support or resistance to health measures. Since not all counties have enough twitter data available, if a user asks for a county with a small pool of tweets the state wide statistics will be displayed. Additionally, an information page is available on the web application to explain exactly what the user is seeing. The page also describes the high level description of the methodology used and topics that may help a user better understand the information displayed.

5 Procedure

The system is available to stakeholders through an interactive web application. The simple design of the web application allows users to understand how to use the system. There are three tabs, which are composed of a landing page, a search page that lets users interact with the models, and an information page that provides information about the models. Using the search page, users can look up any county in the United States to receive information about the county's projected cumulative COVID-19 infections and the Twitter sentiment towards preventative measures of a certain area.

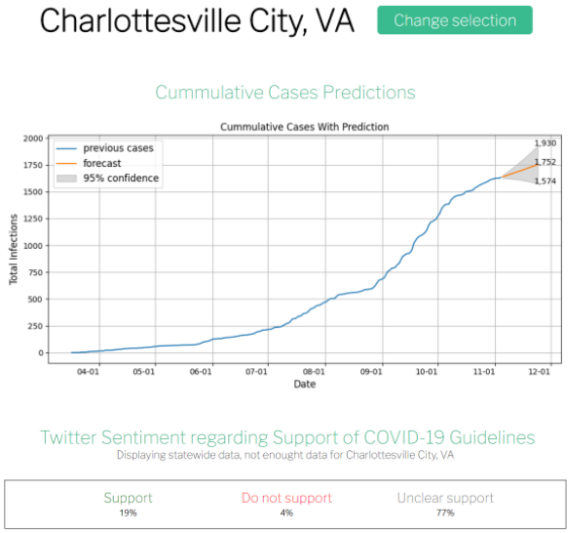

Figure 2: **Layout of information resulting from searching for a county on the web application.**

The application defines the limitations of the general Twitter sentiment to prevent users from generalizing the results of the Twitter sentiments as they are not completely representative of the sample population of a county. In addition, the information page provides more detailed information about the models, so the users can make their own conclusions about the reliability of the system. Besides ordinary individuals, other stakeholders include local

government officials or institutions. They can apply the models to assess the state of the virus in a particular area of the United States to improve their enforcement of preventative measures. Similar to ordinary individuals, these organizations can determine the significance of the results by assessing the models' performance and data on the information page. Another key group of stakeholders includes organizations and individuals who contribute to the spread of valid information and misinformation. These stakeholders include social media users and news organizations. They can use the application to report information about the impacts of the virus. They may also describe the correlation between the results of the infections and Twitter sentiments models. The application's information page hopes to guide these users in presenting accurate information that will not mislead others.

6 Results

The cumulative infection model achieved an average validation accuracy of 87%. Similarly, the sentiment analysis model resulted in a validation accuracy of 89.44% and a training accuracy of 98.36% . The web application allowed users to gain access to forecasted case predictions and some level of understanding of a region's view towards safety measures quickly, efficiently, and in a manner that is easily understandable.

7 Conclusion

This 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. The cumulative infections model was constructed to forecast infections in U.S. counties in order to provide insight about the state of the virus in any local area. The model achieved an 87% validation accuracy in predicting future cumulative infections. The Twitter sentiment model was designed to classify COVID-19 related tweets regarding their support of prevention guidelines. The model achieved an 89% validation accuracy. Ultimately, the sentiment model proposes to improve the application of COVID-19 preventative measures and understand the effectiveness of their enforcement. Finally, the web application hopes to produce a clear representation of the results in order to mitigate any possibilities of misinterpretation and prevent further spread of misinformation. 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.

8 Future Work

The cumulative prediction may be improved by exploring data transformations and further tuning of the model parameters that would allow for more accurate predictions. Improvements to the sentiment analysis methodology would include investigating additional preprocessing techniques that could be applied to the text of a tweet and performing a thorough hyperparameter tuning process on the Logistic Regression model in order to choose parameters that would further improve performance. Finally, the web application may be improved by automating the process of data collection and the creation of predictions in a manner that could be run on the web application in order to update the information daily. Additionally, the incorporation of interactive charts would greatly improve a user's experience and understanding of the information presented.

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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.

[5] 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

[6] 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

[7] Kun, M. (2020). Tweet sentiment analysis methodology. [2]. Technical Report: Predicting Fine -grained Cumulative COVID-19 Infections and Analyzing COVID-19 Tweet Sentiments (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

[8] Lamsal, R. (2020). Coronavirus (COVID-19) geo-tagged tweets dataset. Institute of Electrical and Electronics Engineers (IEEE) Dataport. doi: 10.21227/fpsb-jz61.

[9] 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

[10] 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