**Predicting COVID-19 Twitter Sentiment:**

**Executive Summary**

Peter Harmazinski

Western Governors University

C772: Data Analytics Graduate Capstone

Dr. Emelda Ntinglet

July 4, 2020

**The Problem & the Hypothesis**

COVID-19 is ravaging human life. People are dying, economies are struggling, and the world is changing. 14 years ago, a social media platform, Twitter, was created that enables people around the globe to express themselves. This study examined if COVID-19 metrics have an impact on the sentiment of related tweets. The null hypothesis is COVID -19 metrics have no effect on tweet sentiment, while the alternative hypothesis is COVID -19 metrics do influence tweet sentiment. This study is not the first of its kind. There has also been effort to predict COVID-19 outbreaks based on Twitter data (Jahanbin & Rahmanian, 2020).

**Data Analysis Process**

While the data originate from Twitter and the COVID-19 Tracking Project, the datasets were retrieved from Kaggle. First the Twitter data were filtered to include only U.S. English tweets. Then stop words and non-words, such as URLs and Twitter handles, were removed. For the COVID data set, all quantitative null values were imputed with 0, under the assumption that null values are not erroneous and indicate a lack of data.

Since the data is nonparametric, a Kruskal-Wallis H test was used to identify group differences. The Kruskal-Wallis test did indicate that there is a significant difference in sentiment among the states at a .05 level. While Kruskal-Wallis does have the advantage of not requiring normality, it is also not as powerful as the equivalent parametric test: ANOVA (“Kruskal-Wallis Test”, n.d.). A Dunn post hoc test indicated that the only significantly different state is Alabama.

To improve model performance, feature interactions were identified using a random forest. Of the significant features identified at the .05 level, the 100 most important features were selected for input into the models. For logistic regression, the dimensionality was further reduced using PCA with 2 principal components. PCA yielded better results for logistic regression, but not random forest and KNN.

Sentiment analysis was conducted using VADER, which is a model designed to predict “microblog-like contexts” (Hutto & Gilbert, 2015). There are three potential classes: negative, neutral, and positive sentiment. Considering there are only 236 neutral tweets, they were removed, and the problem simplified to binary classification. Therefore, three classification models were implemented: Random Forest, Logistic Regression, and K Nearest Neighbors. Each has its respective advantages. For example, Logistic Regression is fast and simple, and Random Forest is not prone to overfitting, and KNN has few hyperparameters to tune. Likewise, each model has disadvantages: Logistic Regression requires careful feature selection, Random Forest can be difficult to interpret, and KNN can be computationally intense (Varghese, 2018). In addition to these three, a dummy classification model was also implemented with a uniform prediction for a baseline comparison. Models were compared with a goal of obtaining the highest AUC, while maximizing recall. The analysis and visualizations were completed with Python.

**Study Findings**

This analysis shows that there is some merit to investigating the impact of COVID-19. Of the 3 models compared, random forest had the best performance. It achieved a minimally acceptable AUC of .7 (Mandrekar, 2010). It also achieved the highest F1 score, accuracy, and second highest recall. A high recall value is critical for this model due to the importance of being able to identify states with potential mental health trauma (Rosati, 2020).

**Study Limitations**

The low AUC may be attributed to the dataset limitations, of which there are a few. First, the Twitter data timespan is approximately a month. Second, sentiment analysis is not yet a precise science. There also linguistic factors to consider. Third, COVID-19 testing and tracking was not being well-maintained in April. Lastly, both datasets have a significant number of null values.

**Proposed Actions**

While the final model developed in this study does not excel at classifying sentiment based on COVID metrics, it does exhibit the possibility of such an analysis being successful, given better data. Future studies may benefit from expanding the time period of data used by including international data, and by integrating data from other studies on COVID-19’s effects on mental health. An analysis with more robust data will likely yield better results.

**Expected Benefits**

Although the result of this study leaves room for improvement, it does provide potential benefits. Considering that the COVID-19 outbreak could lead to increased stress and mental health problems (Coping with Stress”, 2020), this analysis can improve mental health awareness while also serving as a foundation for future, expanded studies, as described previously.

References

*Coping with Stress* (2020, April 20). Centers for Disease Control and Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/managing-stress-anxiety.html>.

Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. ResearchGate. <https://www.researchgate.net/publication/275828927_VADER_A_Parsimonious_Rule-based_Model_for_Sentiment_Analysis_of_Social_Media_Text>.

Jahanbin, K. & Rahmanian, V. (2020, March). *Using twitter and web news mining to predict COVID-19 outbreak*. ResearchGate. <https://www.researchgate.net/profile/Kia_Jahanbin2/publication/339770709_Using_twitter_and_web_news_mining_to_predict_COVID-19_outbreak/links/5e84d4db4585150839b508b7/Using-twitter-and-web-news-mining-to-predict-COVID-19-outbreak.pdf>.

*Kruskal-Wallis Test* (n.d.). StatisticsSolutions. <https://www.statisticssolutions.com/kruskal-wallis-test/>.

Mandrekar, J.N. (2010, September). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology. Volume 5, Issue 9. <https://www.sciencedirect.com/science/article/pii/S1556086415306043>.

Rosati, G. (2020, January 4). *Precision, Recall and Predicting Cervical Cancer with Machine Learning*. Towards Data Science. <https://towardsdatascience.com/precision-recall-and-predicting-cervical-cancer-with-machine-learning-367221e70538>.

Varghese, D. (2018, December 6). *Comparative Study on Classic Machine learning Algorithms*. Towards Data Science. <https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>.