**Predicting COVID-19 Twitter Sentiment**

Peter Harmazinski

Western Governors University

C772: Data Analytics Graduate Capstone

Dr. Emelda Ntinglet

June 9, 2020

**Abstract**

**With COVID-19 devastating lives and economies around the world, social media has been an outlet for both frustration and encouragement. In this analysis, binary classification techniques are used to determine the feasibility of predicting U.S. Twitter sentiment based on COVID-19 metrics. Twitter data is leveraged with the goal of helping to identify states that may have a higher likelihood of negative mental impacts due to the pandemic. The modeling techniques used for classification are Random Forest, Logistic Regression, and K-Nearest Neighbors. Of the three, only Random Forest resulted in an acceptable AUC of at least .7. The results are likely impacted by the limited data collection time period, sentiment analysis limitations, and inconsistent COVID-19 reporting. Despite the data limitations and the AUC being only “acceptable”, the analysis shows that COVID-19 metrics do have some consequence on Twitter sentiment. As COVID-19 reporting improves, this analysis may serve as a building block for future studies with expanded scopes of timeframes and locations.**

**Research Question**

Do COVID-19 metrics affect the sentiment of COVID-19 tweets?

Hypotheses:

* H0: COVID -19 metrics have no effect on tweet sentiment.
* H1: COVID -19 metrics do influence tweet sentiment.

Context:

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. Since the data is not normally distributed, nonparametric methods are used for the analysis. This study is not the first of its kind. There has also been effort to predict COVID-19 outbreaks based on Twiiter data (Jahanbin & Rahmanian, 2020).

**Data Collection**

**Data:** Datasets are retrieved from Kaggle, but the data originates from Twitter and the COVID-19 Tracking Project.

Twitter

* <https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april>
* <https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april>

COVID-19 Tracking Project

* <https://www.kaggle.com/sudalairajkumar/covid19-in-usa>

U.S. States

* <https://www.kaggle.com/stansilas/us-state-county-name-codes>

The Twitter datasets have 22 features and 14,607,013 observations. 18 of the features are qualitative and 4 are quantitative. The COVID-19 dataset has 27 features and 3,993 observations. 6 features are qualitative and 21 are quantitative.

There are several limitations to the dataset. First, the Twitter data timespan is approximately a month. Next, sentiment analysis is not yet a precise science. There are also linguistic factors to consider. Second, bCOVID-19 tracking has not been precise due to the nature of the disease. Lastly, both datasets have a significant number of null values. Most of the Twitter features are irrelevant to the sentiment analysis. However, country\_code is relevant, and it has 12,950,294 null values. Since this analysis is location-specific, those observations will be excluded. For the COVID-19 dataset, any null values are assumed to have a value of 0 and are imputed with the same.

**Data Extraction and Preparation**

**Twitter Data**

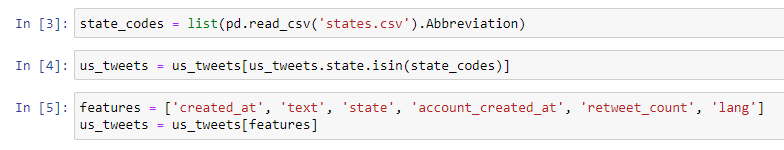
First, I read in the Twitter data from CSV files, stored them in Pandas DataFrames, and unioned the DataFrames together. Pandas allows for less code writing and efficient handling of large volumes of data (“6 Advantages of Pandas”, 2019).



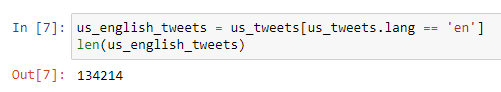


Then I filtered the data to include U.S. tweets only. I read in a CSV files of U.S. states with their abbreviations, which I used to filter the tweets to exclude incorrect abbreviations and U.S. territories.

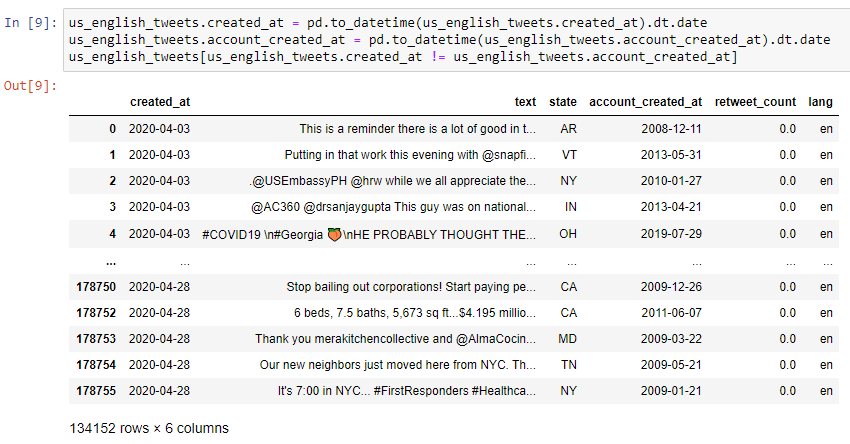




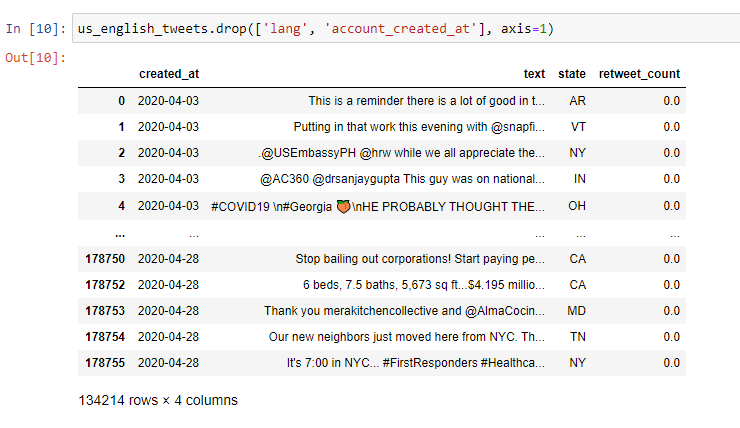
Next, I filtered the tweets to include only English tweets.



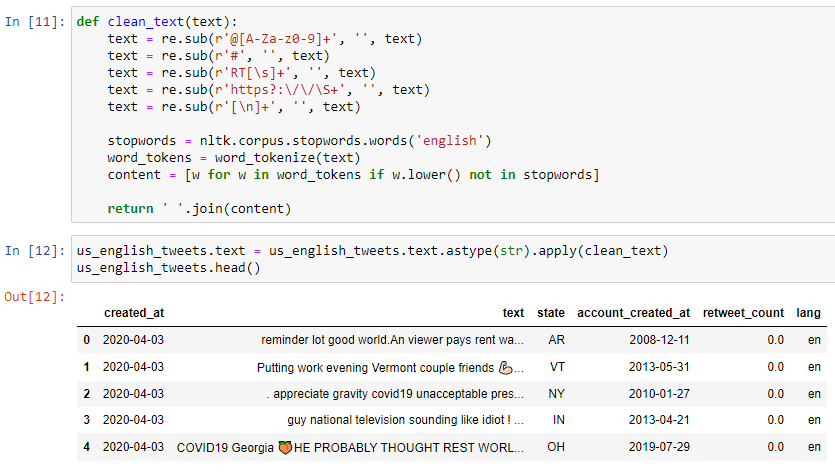
I converted the created\_at and account\_created\_at features to a Python date for easy data manipulation and filtered out any tweets that occurred on the same date as the account creation, in an attempt to exclude a subset of tweets from fake accounts.



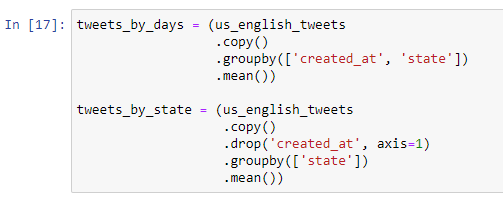
I dropped the lang and account\_created\_at features, since they are not needed for the analysis.



I removed non-word and stop word tokens from the text and added a sentiment column through NLTK’s Vader algorithm (Hutto & Gilbert, 2015).



Then I created two new DataFrames, one grouped by state and date, and the other by state alone. The former is used for classification, while the latter is used for a Kruskal-Wallis H test.

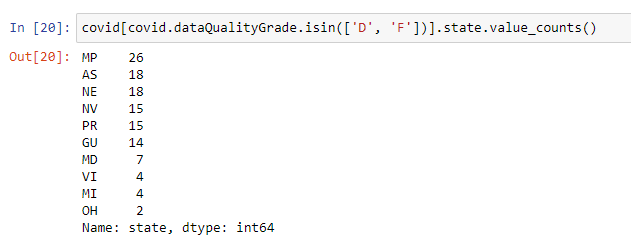


**COVID-19 Data**

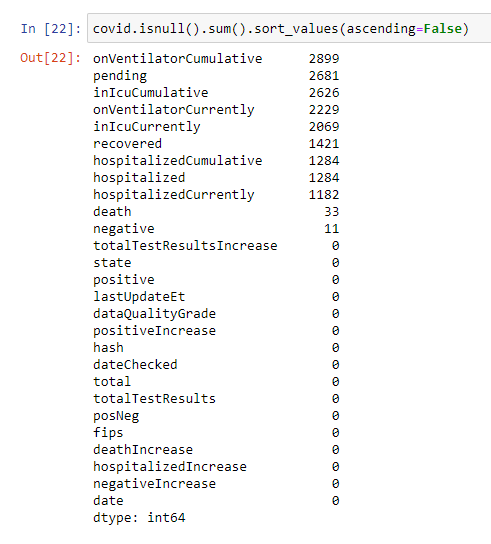
I read in the U.S. COVID-19 data from a CSV file into a DataFrame.

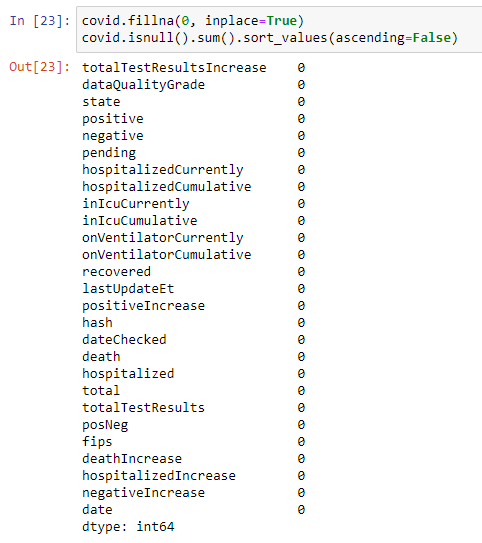


I removed observations with poor quality, denoted by the dataQualityGrade feature.



There were 11 columns with missing values. Considering the novel nature of COVID-19 data collection, it’s unsurprising that there are missing values. I assumed the missing values indicated no numbers reported and imputed with 0’s (LeDoux, 2019).

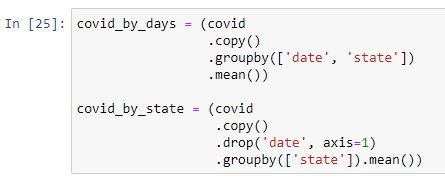




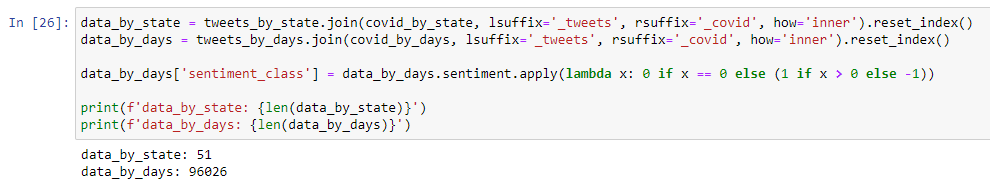
I dropped columns that were unrelated to the analysis: dataQualityGrade, lastUpdateEt, hash, and dateChecked.



Like the Twitter data, I grouped the COVID-19 data into two new DataFrames, by state and date, and by state.



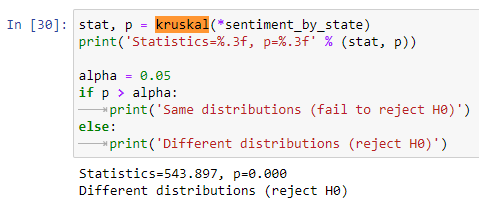
With the datasets cleaned, I inner joined them to create the DataFrames for analysis.



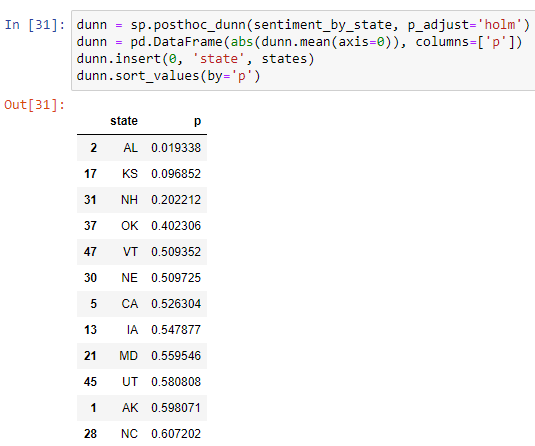
Python was used for this data extraction and preparation. Python is a free, open source programming language with a robust ecosystem of libraries and support. It’s also easy to understand and scalable if a production deployment is required (DataCamp Team, 2020).

**Analysis**

Since the data is nonparametric, a Kruskal-Wallis H test is 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 significance 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.).

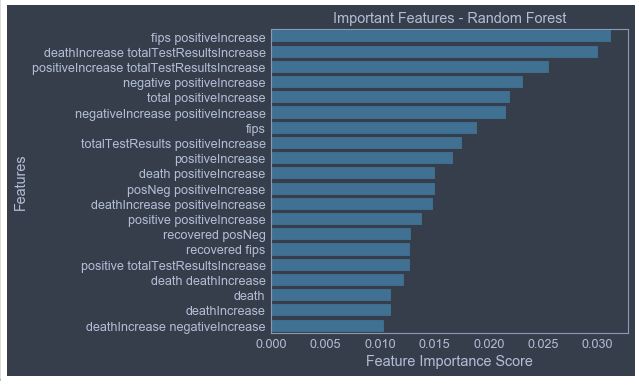


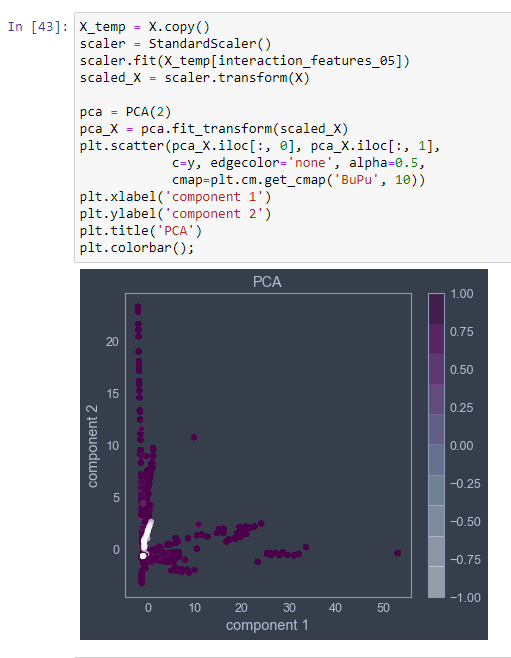
A Dunn post hoc test indicated that only significantly different state is Alabama.



Interaction terms were identified, and dimensionality was reduced through PCA.



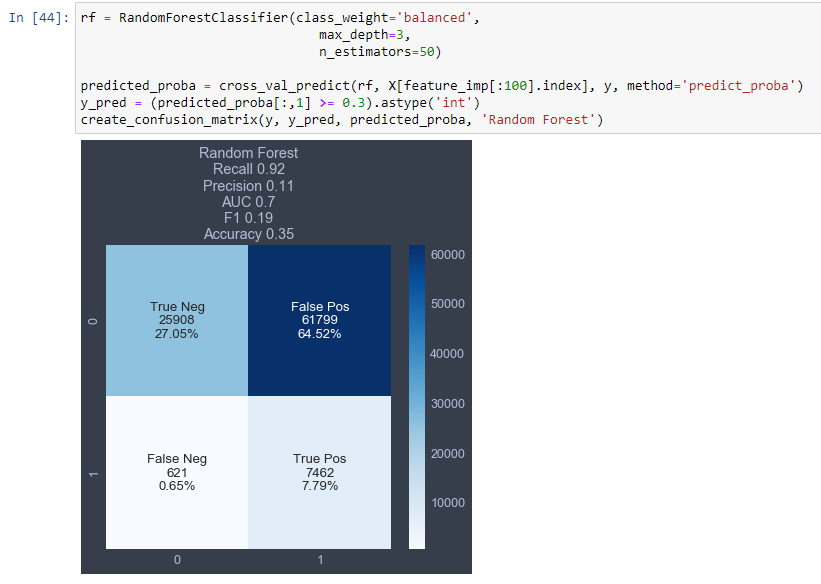




There are three potential classes: negative, neutral, and positive sentiment. Considering there are only 236 neutral tweets, I removed the neutral tweets and simplified the problem to binary classification.

Then I implemented three classification models: 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, I also implemented a dummy classification model 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.

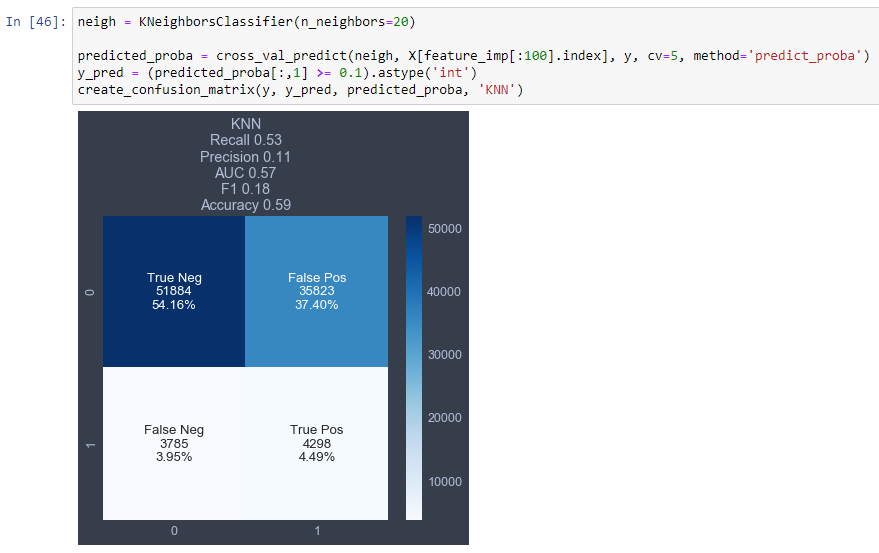
Random Forest



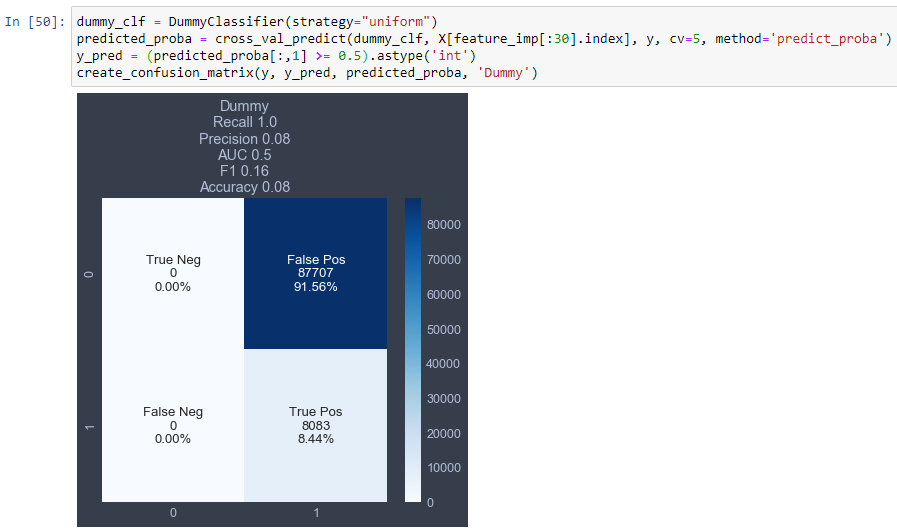
Logistic Regression



KNN



Dummy



**Data Summary and Implications**

Despite the limitations of only a month’s worth of data, the imperfect nature of sentiment analysis, and incomplete COVID-19 data collection, 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).

Considering that the COVID-19 outbreak could lead to increased stress and mental health problems (Coping with Stress”, 2020), it could be beneficial to take this study further. 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.

References

*6 Essential Advantages of Pandas Library – Why Python Pandas are Popular?* (2019, May 1). DataFlair. <https://data-flair.training/blogs/advantages-of-python-pandas/>.

*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.

DataCamp Team (2020, January 9). *Choosing Python or R for Data Analysis? An Infographic*. DataCamp. <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>.

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/>.

LeDoux, J. (2019, June 1). *Impute Missing Values*. <https://jamesrledoux.com/code/imputation>.

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>.