## Applied Machine Learning Project (IST707)

Inside the Python notebook checkpoints folder includes checkpoints for different Python notebooks created throughout this project. They can be accessed and viewed in Jupyter Notebook, VS Code, or your preferred Python notebook environment. The final Python notebook that was actually submitted with this project is located within the zip folder called Experiments\_Sentiment\_Analysis\_Final. This zip folder contains a Python notebook file called Experiments\_Sentiments\_Analysis (5). In this code a variety of Python packages and libraries such as Pandas, numpy, seaborn, matplotlib, scipy, wordcloud, nltk, beautiful soup, Sklearn, fuzzywuzzy, imblearn, mlxtend, and prettytable are imported. To ensure that everything in this code file runs, make sure that you have each of these libraries downloaded. In this file a variety of tasks are done. Machine learning models like Naïve Bayes, Support Vector Machines, k-Nearest Neighbors and Decision trees. Additionally, different methods of vectorization like word2vec, tf-idf and bag of words are conducted. Different versions of the two data sets, one from ZINDI, and one from our professor as used as training and testing data sets in this file. Additionally different functions are defined to remove stop words, hashtags, and emojis. This file can be accessed via Jupyter Notebook, VS Code, or your preferred Python notebook environment.

There are eight Excel files, four of which are csv and four of which are Excel worksheets. The csv files are cleaned\_test, cleaned\_train, Test, Train\_Set. Test and Train\_Set are the original training and testing data sets from the competition website ZINDI. Cleaned\_test and cleaned\_train are the same data sets but with cleaned text removing stop words, hash tags and other unnecessary text. The four excel files are Manually\_Label\_1002-1501, Manually\_Label\_11000-11500, Vaccine\_clean\_20k (1), and Vaccine\_clean\_20k. The two manually labeled files contain 500 rows of manually classified Tweets done by members of our group, these were used as testing data sets on the professor’s data. The two Vaccine Clean data sets are the Twitter/X data provided by our professor with the “(1)” file having more columns. These files can be viewed in Microsoft Excel.

The PowerPoint file, IST 707 PRESENTATION, is the presentation delivered by my group summarizing our goals, methods, data, results, and conclusions. This can be viewed in Microsoft PowerPoint.

The pdf file, Final Report IST707 Group 2, is a more detailed description of how we approached this research task, the methods we employed, the preprocessing steps, feature engineering steps, machine and deep learning modeling, and conclusions. This can be viewed in Adobe Acrobat or your preferred place to view pdf files (i.e., Chrome).