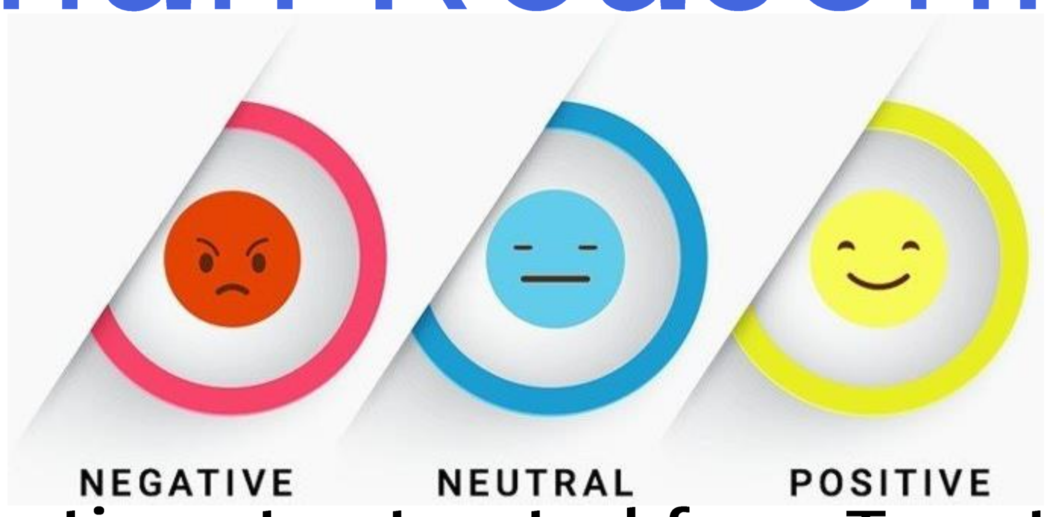
Data Mining Human Reasoning

Taha Kashaf

Advisor: Prof. John Lawrence

Vaccine Hesitancy in the USA



The purpose of this project was to use sentiment extracted from Tweets made by people in individual states within the United States of America to build an understanding of possible reasons for Vaccine Hesitancy.

This was done using Machine Learning Sentiment Analysis tools, and Natural language Processing (NLP) techniques.

Data Collection

Data was collected using the Twitter API. This was done using both the 2.0 and 1.1 Endpoints

Data “Cleaning”

Data returned by the API contained many extra details and needed to be “cleaned” before it was usable. A single “tweet object” returned by the API contains just under 100 unique fields, of which only 2 were needed

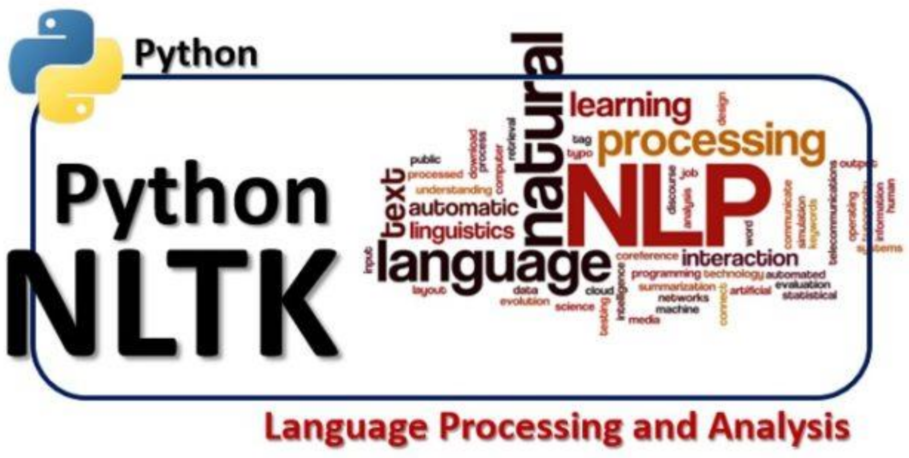
Data “Formatting”

Data needs to be formatted in a certain way for the program to be able to use it. This involves splitting tweets into individual files, and structuring them in a two-level folder structure.

Data “Pre-Processing”

Reg[ular]
Ex[pression]*

Before the model can “understand” the data, it needs to be further processed in a number of ways. To make the data more useful to the model, all special characters, numbers, unwanted spaces, etc. are removed using RegEx. The data is then turned into numbers, as statistical machine learning models can’t deal with raw text.



Classifying

Finally, the Machine Learning Model can be trained and used for classification. Machine Learning was done using Python and Scikit-Learn

python™ scikit learn React

University of Dundee

Computing Honours Projects 2022