Anjan Khatri

IT 320

Dr. Nathan Green

Date: 10/07/2020

Project

Topic: Analyzing the false information related to COVID19 using Natural language Processing

Method and hypothesis

The dataset contains false information related to COVID19 in the text and CSV format with the headlines of news and the curated outcome of the data. The tools that were used to analyze are word distribution using tokenization, part of speech tagging, named entity relation, matplotlib, and polarity of the sentence. The hypothesis of this project was to find the word that would influence the outcome and distribution of that world that would change the sentiment of the text. The analysis of repeating words including bleach, hoax, covid19 vaccine was mostly used that would affect the end result.

Process:

First, the dataset was downloaded from the internet in CSV and later converted to a text file. Since we are analyzing the false news related to COVID19, the data set is curated to provide only false information. We read the data using, and since the data is not cleaned, we used word tokenization to make each token out the world. Tokenization will separate the words along with the punctuations. The tokenized word consists of common English words that were repeatable and could impact the result. We use Stopword Corpus from NLTK. Stopword corpus contains common English words, so we use it to take out those command words from our tokenize sentence. We also filter the punctuation that was not necessary for the analysis. We use a part of speech tagging to attaches a part of speech to each word, which indicates the grammatical categories and helps in text analysis. I use the name entity recognition for the pos tagged word. NER does not have much influence on this analysis, but it would help in the further analysis that helps to search for chunks corresponding to an individual part of speech. Since we have already filtered the data and assigned the part of speech, we now calculate the frequency of words that were used in our data set. By using the frequency of the words, we plot the log graph to visualize the distribution of the words with the words on the x-axis and the frequency on the y axis. This graph gives the most frequently used word at the top of the graph. In the last section of this project, we use TextBlob for the sentiment analysis with the polarity. The polarity gives the simple classification of text based on positive, negative, and neutral sentiment. The result of the polarity from this project shows the neutral sentiments with the polarity of being 0.03.

Additional work:

Since the data set is classified as a piece of false news, only then, when we introduce given data and analyze it uses naïve Bayes using sklearn then we find the accuracy to be nearly 95%. Since all the news data are treated as false news, then we can see that the machine has truly predicted the given set of data. We use TfidfVectorizer which analyze the term frequency and measure the significant term and the count vectorizer to use the textual data for predictive modeling. Then we split the dataset to train the use naïve Bayes model to predict accuracy.