Tal Sternberg
Final Project Detailed Methodology
COSC72 21S

(1)
I have decided to modify my project slightly. I am going to create a COVID19 chatbot (based off the ELIZA program we did at the beginning of the course. This bot will be able to respond to questions and commentary about vaccinations as well as a generalized response for non vaccine related input.

(2)
Using basic sentiment analysis from the Lexicon Sentiment Analysis, I will determine the sentiment of the input. I will be using the same sentiment analysis we used in the week three short assignments for this purpose. More information about the NRC Emotional Lexicon and another applied example (with graphical analysis) that I will use are linked below. I will take the values of [anger, disgust, fear, negative, sadness] to have negative associations and [anticipation, joy, positive, surprise, trust] to have positive associations. For user input I will score each of these and add up the positive and negative values. If the positives outweigh the negatives, the user input is positive and visa versa. If the positive and negative values are equal, the input is neutral. For each response type (vaccine question, vaccine statement or general), there will be a response structure for a case with positive, negative, and neutral associations (determined by the sentiment analysis). I will also use REGEX to form sentence structure and pull words from user input, as well as tell apart statements from questions (by use of key question words and mostly question marks). The input will be typed user text (theoretically with correct punctuation) and the output will be a relevant string response that corresponds to the type and sentiment of the input.

Links:
https://investigate.ai/upshot-trump-emolex/nrc-emotional-lexicon/
https://github.com/raffg/harry_potter_nlp/blob/master/sentiment_analysis.ipynb

(3)

Because my chatbot is based around vaccine sentiment, I will be testing it on the Kaggle dataset of covid vaccine tweets. I will test the first 50 tweets as input (metrics to be explained in the next section) to ensure that the bot is working the way I expect it to for the most part. If there are not enough questions in the Twitter dataset, I will Google "commonly asked questions about the covid vaccine" and use 10 of those in addition to my dataset for testing purposes. Other than this possible case, the dataset should not need much modification.

(4)

I will test the first 30 tweets as input for the bot and score my bot on the following:

+1 point if it correct identifies a question or statement and the responds accordingly

+1 point if it correctly identifies a positive, negative or neutral sentiment and responds accordingly

No partial points

If the bot scores an average of 1.75 or above on overall accuracy, I will treat the algorithm as successful and functional.

(5)

A paper by Suryadi et al.:

https://iopscience.iop.org/article/10.1088/1757-899X/1077/1/012042/pdf

Summary: The team used lexicon-based sentiment analysis on covid-19 related tweets from multiple different countries with translation APIs. It was found that for the Indonesian-speaking country, the dominant emotions were sadness and trust, while in the English-speaking countries, the dominant emotions were fear and trust. The team used Emotion Mix, Emotional Intensity, and Daily Dominant Emotion to score and compare the analysis on these tweets from country to country. The highest number of tweets from a single news source was 19,357 and the lowest was 2,032. While the number of total tweets analyzed was not given, this gives us a sense of the magnitude of the operation.

A paper by Kundi et al.:
https://www.researchgate.net/profile/Dr-Muhammad-Asghar/publication/283318830_Lexicon-Based_Sentiment_Analysis_in_the_Social_Web/links/5632e78d08aefa44c3685cda/Lexicon-Based-Sentiment-Analysis-in-the-Social-Web.pdf
Summary: This paper uses a more simplified lexicon based framework: classification of a tweet as positive, negative or neutral. It is also trained to detect and score internet based slang. They used a dataset of 308,316 tweets from three different types of products (iPhone, Nokia, Samsung). Their algorithm was 92% accurate for binary classifications and 87% accurate for multi-class classifications.

A paper by Denecke et al.:
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9000924&casa_token=NpYsAcRvvwkAAAAA:qDHF845lC8fc2O9SD-qyglQe8fabIY2Jdc1hmXph9qNQ0E2Sh1rNsRJCpwQYvYg-8fm3XW99QA
Summary: This team created an emotion-regulatory chatbot using lexicon based sentiment analysis to identify the users' emotional state. Using recursive deep models for sentiment analysis, the bot achieved 80% accuracy on pinpointing specific emotions present in users. There is no direct data or measurements of sources they used to train and test, but they do mention using data and sources found using Google Scholar.