ISyE 6740 – Spring 2021
Project Proposal

Team Member Names: David Kaczmarkiewicz
Project Title: Sentiment Analysis of COVID Vaccine Tweets

Problem Statement

To understand public sentiment towards covid vaccines and understand that sentiment by country, and by time throughout the pandemic. If possible look at the demographics or characteristics of each country and see if there are any characteristics that tend to lead the public to positive towards the vaccines or be negative about them. Ideally would like to be able to go deeper into understanding the sentiment of the tweets beyond positive, negative, or neutral and be able to classify emotions such as excitement, hopefulness, skepticalness, distrustfulness, or frustration. I would also like to understand if any of these sentiments changed throughout the pandemic and possibly understand why or look to events that may have caused those changes.

Data Sources

All COVID-19 Vaccines Tweets.
https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets

Multiple possible labeled and established datasets to train for a supervised approach at classification:
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank
Twitter US Airline Sentiment
Sentiment Dictionaries for WordStat Content Analysis Software
Emoji sentiment data

Methodology

Preprocessing: If necessary, remove tweets not pertaining to COVID vaccines. Remove stop words, remove punctuation, move hashtags and emojis to separate fields separated from the text. Create categorical features from hashtags and emojis. Process user-generated location information into categorical locations.

Feature extraction: Two possible paths we could take include implementing a bag of words approach or Tweet vectors. The bag of words approach will let us define a vocabulary of known words and also measure the presence of those words. All structure of the sentence or order of words is removed. Tweet vectors make it possible for similar words to have similar representations which may improve the performance of the classifiers. May need to utilize a dimensionality reduction process like PCA.

Classification:

Defining neutral tweets: Objective vs Subjective. There will be plenty of tweets about covid vaccines relaying new information but that do not offer a sentiment or opinion towards the vaccine. These need to be grouped in their own classification. This is a highly critical step, we want to analyze the subjective tweets and classify them, we're trying to analyze people's feelings and opinions so the objective tweets must be separated.

Unsupervised Techniques:

Utilize only the COVID tweet database in order to classify the tweets in distinct categories. Some of the algorithms I plan to implement include the following:

Naive Bayes: Utilize Bayes's theorem to predict the categories.

Support Vector Machines: Utilize the tweet vectors that represent points in an n-dimensional space in order to separate them by some distance cost function.

K-Nearest Neighbors: Similar to SVM by using the tweet vectors and group tweets by distance to each other. Determining a good K value may be difficult. Possibly use a radius-based technique in order to determine "neighborhoods."

Logistic Regression: I can imagine two techniques to make logistic regression work. We could remove all the neutral tweets by another classification technique and then gauge how strong the sentiment is by how close the tweets returned value is to 0 or 1. Or we could leave the neutral tweets in the database and set up a zone around .5, say .3 - .7, and define those tweets as neutral, points closer to 0 and 1 as negative and positive sentiments.

Supervised Techniques:

A Rule-based approach applied to the bag of words data. Create a list of polarized words. Count the number of negative vs positive words classify the tweet as negative or positive or neutral. Expand this idea for different sentiments and emotions to classify into more than 3 categories. Will utilize the WordStat dictionaries here.

Apply supervised techniques (Decision trees, KNN, Naive Bayes, Logistic Regression, Neural Networks) by training the model on a labeled set of data. I will need to ensure there's no special language over-represented in COVID vaccine tweets that are not represented in the training data.

Evaluation and Final Results

I expect to see overwhelmingly positive sentiment towards COVID vaccines. The more interesting data will be trying to determine how and why vaccine negativity trends across the pandemics' time frame. Ideally, I'd like to show line graphs of the change in sentiment over time for all of Twitter and subsets by location. I would also like to produce word clouds that show the language used to describe the positive and negative sentiments towards the vaccine.

Sources and Literature

Opinion Mining and Sentiment Analysis.
Bo Pang and Lillian Lee
https://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank
Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

Sentiment Analysis of Twitter Data
http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf

NLTK Book
https://www.nltk.org/book/

Sentiment Analysis: A Definitive Guide
https://monkeylearn.com/sentiment-analysis/