ISyE 6740 – Spring 2021
Final Project

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

Problem Statement

Reviewing social media and the news one can see there are mixed emotions and varying public sentiments in regards to the various vaccines produced to combat the COVID-19 virus. The CDC advises that vaccinating a large portion of the population will help slow and stop the pandemic currently affecting the world. It is important and would be helpful to officials tasked with overseeing vaccination efforts to understand this public sentiment and the emotions of the population in greater detail.

If we could understand the changes happening over time we might be able to tie events to the changing sentiment. There would also be value in understanding the overall sentiment towards each of the particular vaccines such as Pfizer-BioNTech, Moderna, Johnson and Johnson's Jansenn, Oxford-AstraZeneca, Sputnik V, Covaxin, Sinopharm, and Sinovac.

Public sentiment analysis generally deals with classifying texts as positive or good, negative or bad, or neutral. If we could understand in finer detail though such as Ekman's proposed six basic human emotions of: fear, anger, joy, sadness, disgust, and surprise (Ekman) we could better understand what the public is feeling overall.

Data Sources

The primary dataset (Preda) used in this analysis is a collection of almost 70k tweets from Twitter between 12/12/2020 and 4/14/2021 that were compiled by searching for the most relevant search term for each of the following vaccines: Pfizer-BioNTech, Moderna, Oxford-AstraZeneca, Sputnik V, Covaxin, Sinopharm, and Sinovac.

The dataset includes the tweet, the user name, user provided location, user provided description, user create date, the number of user followers, friends, and favorites, if the user is verified, the date of the tweet, the hashtags within the tweet, and how many times the tweet was retweeted.

The data was obtained from kaggle.com and is included in the citations.

There are some limitations to the dataset that were discovered upon analysis. A number of the longer tweets do not capture the full tweet, the ending of the tweets is cut off and missing.

Johnson and Johnson's Jansenn vaccine has not been included in a similar way as the other vaccines which were specifically searched for and included. It does appear in the dataset in much more limited numbers than the other vaccines when taking into account its popularity in the United States and Twitter's overall United States centric nature. I believe the inclusion is most likely due to the dataset creator searching for general COVID-19 vaccine terms which overlapped with a limited number of tweets regarding Johnson and Johnson's Jansenn vaccine.

Another dataset used in this analysis of the primary dataset is GloVe's pre-trained twitter word vector array. (Pennington et al.) It provides an embedding of aggregated global word to word co-occurrence statistics from a corpus including 2 billion tweets representing linear substructures of the word vector space.

Methodology

I employed three different approaches to attempt to solve different parts of the problem statement.

In the first I utilized the NLTK (Natural Language Toolkit) sentiment package and specifically the sentiment analyzer module and the sentiment intensity analyzer models in order to classify each of the tweets as positive, negative, or neutral and the intensity that the model believes represents those three sentiments. I then group the data by date and present the findings as well as take a closer look at individual tweets. I will refer to this as the SIA method.

In the second I implement the concept of word vectors and a pre-trained embedding provided by Stanford referred to as GloVe (Global Vectors for Word Representation). I then map the corresponding vector to each word in a tweet, and average the values across the tweet. I then compare each tweet against the six different emotions: fear, anger, joy, sadness, disgust, and surprise. Once the scores are obtained I analyze the data and my findings.

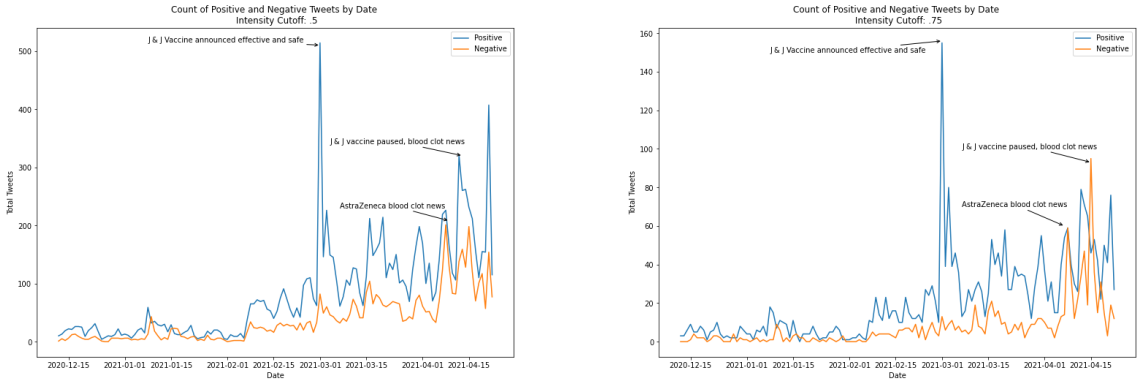
In the third and last approach implemented I attempt to train a set of word vectors from the primary dataset and look at what the model believes are similar words as well as ask the model to solve a number of analogies in it's newly trained domain. An example of an analogy this type of model can solve on a larger more robust dataset is "man is to king, as woman is to…" The model will return results similar to queen, or princess. I refer to this method as Word2Vec based on the library used.

SIA Method

The NLTK's sentiment package relies on the VADER (Valence Aware Dictionary for Sentiment Reasoning) model. Using the dictionary the model can create scores from tweets by looking at their grammar and construction. The model can interpret positive words such as love, happy, joy, excited and assign a positive emotional intensity, the same is true for negative words. The model is also intelligent enough to understand "*am not happy*" and classify it with a negative

score. The model creates a score for each word and sums up the total to come to a final classification.

An advantage to this method is the minimal amount of preprocessing needed to apply it to a dataset. The model takes into account words that are entirely capitalized as well as punctuation and then applies modifiers taking into account the special emphasis placed on the word. Because of this minimal preprocessing was done. Emojis and hyperlinks were removed, while hashtags were converted to drop the leading '*#*' and leave the word in place.



Above are two plots representing the count of positive and negative tweets for each day of the dataset determined by a cutoff factor of .5 and .75 from the model. I have highlighted a few events that happened on specific days where spikes are shown in the data, both positive and negative. On 2/28/2021, the CDC announced the Johnson and Johnson's Jansenn vaccine was determined to be safe and effective while also having the advantage of only receiving one shot. The enormous amount of positive tweets can be seen in both plots.

The two other highlighted points represent news being released of blood clots possibly being caused by AstraZeneca vaccine and the Johnson and Johnson Jansenn vaccine. The negative tweets increased substantially in this case as one would expect. There is a difference between the two plots, the negative sentiment is far greater when we only consider the most extreme intensity scores. Curiously, with a lower intensity score, more positive tweets can be seen, but the event does cause a spike in activity.

There is a limitation of the data in this information. The author of the dataset did not specifically include Johnson and Johnson's Jansenn vaccine and two of the major events in the time period directly relate to their vaccine. I would be curious to see the results with that data included in the model.

Some of the most positive tweets from the analysis include the following:

User Name	Tweet	Compound Sentiment Score
Alessandrina	Safe and Effective, Safe and Effective, Safe and Effective, Safe and Effective, Safe and Effective, Safe and Effec…	0.9847
JanPaasa	It's easy to be brave sometimes but it's a courageous thing to be strong. Cheers to all strong and independent yout…	0.9712
Nicola Facciolini	@mod_russia GOD BLESS HOLY #RUSSIA VICTORY TRIUMPH JUSTICE. GOD BLESS PRESIDENT #PUTIN. THE #ITALY PEOPLE WANTS…	0.967
Michelle R Cloud	My beautiful daughter and I got our first dose of the vaccine today!! So happy, relieved, grateful, thankful to the… https://t.co/i0AS4naA5A	0.9602
Aimee Giese	Vaccine Number 2, CHECK! The sense of relief and joy is amazing. Thank you health care workers, front line workers…	0.9516

Some of the most negative tweets from the analysis include:

User Name	Tweet	Compound Sentiment Score
Raj Meister	#SputnikV causes NO Harms NO Issues NO Health Warning NO Health Complications NO Blood Clotting NO Deaths Why isn… https://t.co/TNGmruPka2	-0.9682
Jeremiah Marsh PeoplesLobby	STOP THE MURDER!! THIS IS INSANE!!! #ModeRNA is a killer #vaccine. https://t.co/UkBWJShtPu	-0.9599
Workout Solutions	@dockaurG While Canada is forced to stay at home home, no family, no business, no school, drug abuse, violence, une… https://t.co/tiOLYwIZB9	-0.9509
Hannibal Khoury ⏳	The @EU_Commission's criticism of #SputnikV is the most pathetic, insecure shit I've ever seen.	-0.9305
Rhonda S.	I hate Tylenol but had to get up and take it cause my arm was excruciating pain from the second shot.,I also cancel… https://t.co/7AcWHliXw8	-0.9294
xZavior	#johnsonandjohnson #vaccine KILLS \n#PfizerVaccine KILLS #Moderna KILLS the government just tells you "shut up and… https://t.co/ppwQpIuN2d	-0.9287

The results seem to show the algorithm is performing fairly well. All the positive tweets have a very positive sentiment and the negative a negative sentiment. Some pitfalls of this type of analysis though is the model is simply judging the sentiment of the actual tweet and we can't be sure the sentiment is in regards to our subject, the vaccines such as in Hannibal Khoury's tweet, "*The @EU_Commission's criticism of #SputnikV is the most pathetic, insecure shit I've ever seen*." This user is most likely in favor of the Sputnik V vaccine but the tweet is included in the dataset, and from context directing his negative sentiment towards the EU Commission not approving the Sputnik vaccine at the time.

We could improve this limitation by filtering out data from the dataset not pertaining to the vaccines, this could be the subject of an entire other project on topic identification in natural language processing.

Word Vectors with GloVe Method

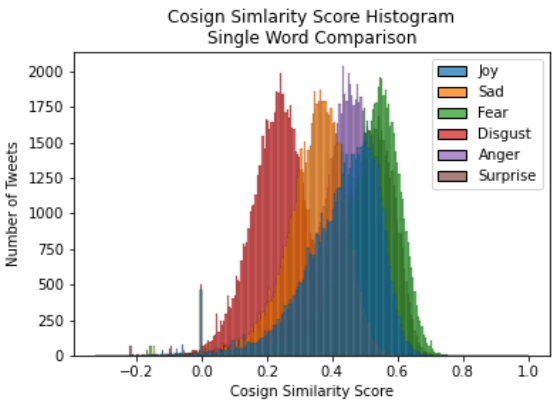
In the second method I implement the concept of word vectors and a pre-trained embedding provided by Stanford referred to as GloVe (Global Vectors for Word Representation). I then map the corresponding vector to each word in a tweet, and add them up. I then compare each tweet against the six different emotions: fear, anger, joy, sadness, disgust, and surprise. Once the scores are obtained I analyze the data and my findings.

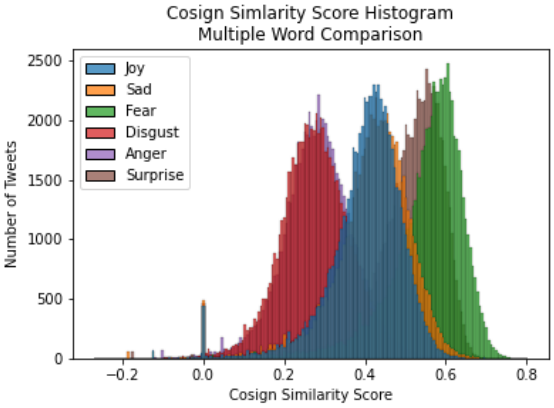
GloVe is an unsupervised learning algorithm for obtaining vector representations of words. The model used in this approach was pre-trained by the creators of the package on a Twitter dataset of two billion tweets, with over one million unique words. A word vector is a vector of numbers that represent the meaning of a word. The idea behind the concept is to place each similar word in a similar vector space. For this analysis I chose to use the 100 dimensional vector space representation.

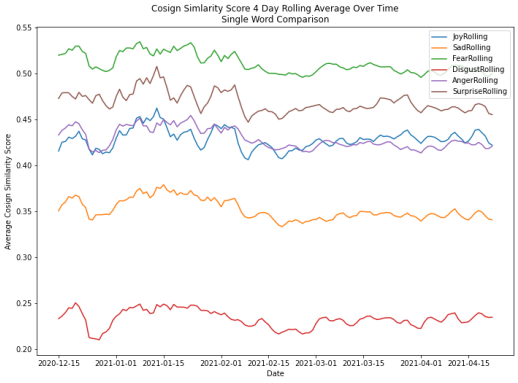
Our primary dataset needed to be preprocessed in order to be evaluated by our pre-trained model. All of the punctuation, hyperlinks, all mentions of other users ie. '@davekacz', and common stop words such as "a", "the", "is", "are" needed to be moved. All of the text needed to be converted to lowercase similar to the training data.

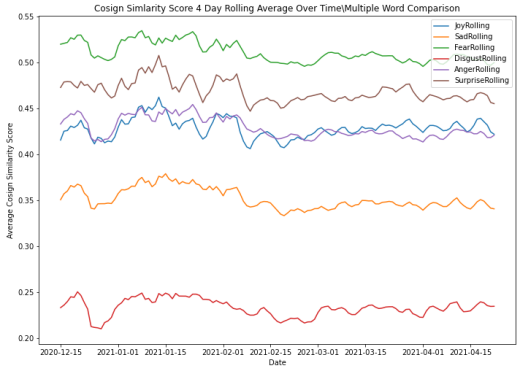
My first attempt at utilizing the model was to compare each tweet to the six emotions defined earlier: fear, anger, joy, sadness, disgust, and surprise. Each tweet was mapped to its average vector space by adding the sum of each of the processed tweets and dividing by the number of words. These sums were then divided by the number of words in the tweet to arrive at the average vector space point of the sentence. I then took the cosine similarity score of each of the tweets against the words: joy, sadness, fear, disgust, anger, and surprise.

I also compared each of the tweets to the average vector representation of a list of words meant to convey the emotion in a more broad way as opposed to a single word. For example the list of words that composed the sadness included: 'sadness', 'unhappiness', 'sorrow', 'dejection', 'regret', 'depression', 'misery', 'despair', 'desolation'.









The plots above show both the single word comparison and the multiple word comparison. The histograms show the overall distribution of the cosine similarity scores for each of the six emotions. The line plots show the average cosine similarity score for all tweets for each date of the dataset.

If we look at the similarities of the distributions on the histogram plot we notice that *disgust* is much less represented than any of the other emotions, but fear has much higher scores across the dataset. Both of these findings make logical sense on what one would expect from our dataset. If we look at the differences though, we find that the *joy* distribution has shifted to a lower score while surprise has shifted to a higher score. Perhaps some of the additional words in each of those emotions captured a slightly different meaning than the single word.

The line plots all show very similar patterns in terms of movement and each of their mean scores although varied between each other hold consistent across time. My impression is the model is not working as intended. Perhaps instead of trying to capture the six different emotions in a meaningful way the model captured that strong emotions were present. Judging by the similar spikes and valleys in each of the emotions when news came out that caused an emotional reaction in some way, the model responded with higher scores across all categories.

Taking a closer look at the scores for each individual tweet we can see the over representation of fear where we would not expect it. The first table is a sampling of tweets our first classifier SIA classified as very positive, we can also see that *surprise* represented very highly.

	So honoured to be part to this exciting team administering in primary care #Covid_19… https://t.co/gWDTDMGoF0	Take a breath and celebrate what happened today. Democracy won. Our constitution won. #PfizerBioNTech gave us hope… https://t.co/W3mdtevk26	Day two of #CovidVaccination with #TeamMFT. Great to see so many people starting to smile as hope for a great #2021… https://t.co/Awu927l4CS	I was so fortunate and PROUD to get my #CovidVaccine yesterday!!! I am so thankful for everyone who made this histo… https://t.co/aaCL9uCkf1	Finally! What a fantastic Christmas gift! Please get vaccinated if your health care provider recommends it when it… https://t.co/1M0jTsYZqm
Joy	0.5263	0.6104	0.6214	0.5887	0.5645
Sadness	0.3688	0.4773	0.4586	0.4421	0.3908
Fear	0.5414	0.6453	0.6001	0.5403	0.5288
Disgust	0.2129	0.3439	0.2712	0.3369	0.2020
Anger	0.4100	0.5314	0.4913	0.4486	0.4322
Surprise	0.6175	0.6260	0.6199	0.6117	0.6423

The table below shows some of the most negatively scored tweets from the SIA analysis. *Fear* is highly represented here as well, but at an even higher degree than the previous tweets, which follows logically with what we would expect, the tweets seem to exhibit fear. *Joy* is also highly scored in these tweets which is not what one would expect.

	Tragedy. Another life taken by corporate greed and the evil pharmaceutical companies. #vaccination #vaccine… https://t.co/1YcShyQpR6	@OANN Awful. God have mercy on those being manipulated into getting this hellish unethical, dirty, aborted fetal ce… https://t.co/ePDn9T13dQ	U MAY die if u become covid positive. But #PfizerBioNTech shot will DEFINITELY kill you😭😭 Stuck between devil and s… https://t.co/U8yz2l1oSc	Fuck @AstraZenecaNL with their arrogance, lies and bad result for protection agains #COVID19 . Stop buying that… https://t.co/8lsJ47UxqZ	European, American and Indian also was brutally killed every year by terrorist all over world who comes a community… https://t.co/qLJLk9hNTM
Joy	0.5053	0.4764	0.6101	0.5632	0.5115
Sadness	0.4570	0.4710	0.4859	0.5047	0.4052
Fear	0.6945	0.6229	0.6756	0.7441	0.6022
Disgust	0.3582	0.4401	0.2729	0.4750	0.2865
Anger	0.5752	0.5248	0.5519	0.6418	0.4728
Surprise	0.4523	0.5011	0.6052	0.5502	0.5463

I also attempted a number of unsupervised clustering methods with the tweet vectors as assembled. I then found the word with the highest cosine similarity score as the center of each of those clusters curious what the model would report. Unfortunately even testing a variety of algorithms and cluster centers the models provided nothing worthwhile to report.

Word2Vec Method

In this method of analysis instead of using a pre-trained embedding for each word and comparing it to our primary dataset, I will train a vector representation using our primary dataset and then try to extract information from what the model has learned by asking it for similar words or asking the model to respond to analogies. I utilized the Word2Vec package from the Gensim library to train the model.

The data was preprocessed by removing all of the URL's, mentions (@davekacz), emojis, smileys, numbers, and punctuation. All of the mentions of the specific vaccines: Pfizer-BioNTech, Moderna, Oxford-AstraZeneca, Sputnik V, Covaxin, Sinopharm, and Sinovac were modified to be represented by the same word in order for the model to be able to connect more data to the appearances of those words.

Below are the vaccine names and the top 10 words that the model believes are similar to them based on their singular represented word and their cosine similarity scores. As you can see, what the model learned was not sentiment towards any of the vaccines, but it did pick up on topics or words that were frequently discussed with those vaccines.

The model picks up quite a few words that with more time I would clean up so that they are all represented by the same word such as in *oxford* and *zeneca* appearing in the Oxford-AstraZeneca column. The model does provide some interesting information though with *blood* and *clots* being

represented also under Oxford-AstraZeneca as well as many of the same terms regarding China in both the Sinopharm and Sinovac vaccines.

Oxford-AstraZeneca	Pfizer-BioNTech	Moderna	Sputnik V	Sinopharm	Sinovac	Johnson and Johnson
oxford , 0.81	johnson , 0.71	pfizer , 0.51	ema , 0.62	sinovac , 0.86	sinoph , 0.86	jampj , 0.85
az , 0.69	jampj , 0.63	microchips , 0.50	austria , 0.62	chinese , 0.73	chinese , 0.71	biontech , 0.73
uk , 0.62	biontech , 0.58	florida , 0.45	france , 0.60	china , 0.69	coronavac , 0.63	pfizer , 0.71
blood , 0.61	astra , 0.58	brian , 0.45	eu , 0.60	nationals , 0.62	hong , 0.61	janssen , 0.69
zeneca , 0.60	pause , 0.57	vibes , 0.45	russias , 0.60	pakistan , 0.59	china , 0.61	jnj , 0.66
johnson , 0.60	phizer , 0.57	jampjs , 0.45	germany , 0.60	chinas , 0.59	chinas , 0.60	novavax , 0.65
clots , 0.60	oxford , 0.53	woo , 0.44	europe , 0.59	beijing , 0.59	kong , 0.58	pause , 0.64
pfizer , 0.58	novavax , 0.53	updating , 0.44	russian , 0.58	uae , 0.59	nationals , 0.57	jandj , 0.63
paused , 0.57	uk , 0.52	welp , 0.44	sneaked , 0.58	manufact , 0.57	hongkong , 0.56	phizer , 0.62
oxfordvaccine , 0.56	moderna , 0.51	yall , 0.44	merkel , 0.58	coronavac , 0.56	beijing , 0.56	conservatives , 0.60

The table below shows the top 10 outputs when posing the model the following analogy, “Thankful is to AstraZeneca, as the other vaccines are to…” Excited appears frequently across all the vaccines as can be expected. Dolly Parton makes an appearance quite frequently in the Johnson and Johnson’s vaccine responses. Her receiving the vaccine that she helped fund must be well represented in the dataset.

Pfizer-BioNTech	Moderna	Sputnik V	Sinopharm	Sinovac	Johnson and Johnson
excited , 0.72	excited , 0.72	excited , 0.54	gift , 0.60	phil , 0.56	dolly , 0.69
grateful , 0.70	happy , 0.66	happy , 0.47	phil , 0.59	excited , 0.53	dollyparton , 0.67
relieved , 0.70	super , 0.66	finally , 0.47	excited , 0.57	emmerson , 0.52	blessed , 0.66
dodgers , 0.69	yay , 0.66	kissing , 0.47	happy , 0.54	dodgers , 0.52	grateful , 0.65
blessed , 0.68	fauciouchie , 0.66	gift , 0.46	proud , 0.54	phnom , 0.52	dodgers , 0.65
happy , 0.67	dolly , 0.65	proud , 0.46	dodgers , 0.54	frontliner , 0.52	dollypartonvaccine , 0.65
incredibly , 0.67	blessed , 0.65	blessed , 0.45	emmerson , 0.53	wilkins , 0.50	science , 0.64
hopeful , 0.64	covidvacccine , 0.64	traveling , 0.45	paf , 0.52	paf , 0.50	excited , 0.63
fauciouchie , 0.63	relieved , 0.64	dodgers , 0.45	thanx , 0.52	jose , 0.50	incredibly , 0.63
science , 0.63	vaxxed , 0.64	venkatesh , 0.44	liberation , 0.52	happy , 0.50	covidvacccine , 0.63

A final analogy asked to the model was looking to find the model's sentiment towards rapid covid tests as opposed to the pcr tests. The model was asked, “PCR is to safe, as rapid is to…” and the model provided the following responses.

rapid
ineffective , 0.65
lactating , 0.64
sourcing , 0.62
risky , 0.60
protected , 0.60
cheapest , 0.58
deadly , 0.58
real , 0.57
covishild , 0.57
pregnant , 0.57

There are a few words in the response from the model that one would expect to find such as *ineffective*, *risky*, *cheapest*, *deadly*, but there are a number that do not make much sense such as *lactating*, and *pregnant*.

Evaluation and Final Results

Overall there was some success in trying to better understand the public’s sentiment when it comes to the different COVID-19 vaccines. There were also some areas that were limited by time and scope and the data I had to work with that I hope to address later.

The SIA method proved to be quite robust by being able to gauge public sentiment to be either positive, negative, or neutral and changes were shown in that sentiment at times one would expect the change in sentiment to occur with the news. I believe the model works as intended and could be used to monitor ongoing sentiment towards each of the vaccines.

There was less success with the word vector methods. Some relationships were shown with both methods, but they are much harder to interpret. It is also difficult to understand what might be the right or an interesting question to ask the model once it has been trained in the Word2Vec case.

More time and effort could be put into the word vector methods to make them more robust. The dataset also created a limitation without the full extent of the tweet in some cases. This loss of language may have affected the model negatively. In time I plan to gather my own dataset and see if these methods can be improved on with better data, a more robust preprocessing method, and even utilizing some target identification methods to better inform the sentiment analysis.

In the end we must also consider where we started. All of the models and methods in this analysis were derived from tweets from Twitter about COVID-19 vaccines. Twitter is not the real world, and although it has influence on public sentiment, it is almost certainly not representative of the entire population's opinions or sentiment.

Sources and Literature

Bird, Steven, et al. *Natural Language Processing with Python*. Creative Commons, 2019. *NLTK*, https://www.nltk.org/book/

Ekman, P. *Are there basic emotions?* 1992. *APA PsycNet*, https://doi.apa.org/doiLanding?doi=10.1037%2F0033-295X.99.3.550

Pang, Bo, and Lillian Lee. *Opinion Mining and Sentiment Analysis*. 2008. *Opinion Mining and Sentiment Analysis*, https://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf

Pennington, Jeffrey, et al. *GloVe: Global Vectors for Word Representation*. 2014. *GloVe: Global Vectors for Word Representation*, https://nlp.stanford.edu/projects/glove

Preda, Gabriel. “All COVID-19 Vaccines Tweets.” *Kaggle*, 2021, https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets Accessed 02/ 05/ 2021.

Packages Used

---

pandas, numpy, scikit-learn, nltk, glove, preprocessor, regex, gensim, word2vec