Melissa Vang, Raheyma Khan, Shivani Garlapati, Sreni Sureneni, Deeksha Poornashri  
Group 3 - COVID-19 vaccine  
Term Project Proposal

1. Project title      
   
Discussions and attitudes about COVID-19 vaccine and 2020 election on twitter: A text mining analysis?

2. Project background   
  
 The coronavirus disease (COVID-19) pandemic which appeared at the end of 2019 has continued to spread globally. As COVID-19 is a new disease and an outbreak worldwide, there is ongoing development of COVID-19 vaccines. According to Ned Temko Correspondent (2020), there have been 160 candidate vaccines around the world, but there is no way of knowing how effective, safe or any of them will prove. Thereon, with the global spread of the COVID-19 infection, activities on social media applications like Twitter began to increase (Abd-Alrazaq et al., 2020).

Along with the COVID-19 pandemic outbreak, there is the upcoming U.S. presidential election scheduled for November 3, 2020. “All members of the U.S. Congress, one-third of Senators, and numerous state and local elected officials are up for election at this time” (Johnson et al., 2020, p. 253). With the seriousness of the pandemic situation, it is likely to have an impact on the 2020 election. This research is important because there is little research covered on the public’s response related to COVID-19 vaccine on Twitter as well as tweets related to COVID-19 vaccine and the 2020 election.

3. The research questions  
  
1. What are people's attitudes towards the COVID-19 vaccine? How have COVID-19 affected people’s lives?

2. How is the COVID-19 vaccine related to the 2020 election? What are the most related tweets regarding the COVID-19 vaccine and the 2020 election?

4. The goal of the project (What you are going to do in this paper to answer the research questions)  
  
COVID-19 vaccine is the most discussed topic on Twitter in recent times. The high quality data is to be collected from Twitter. The goal of our project is to study the tweet texts related to COVID-19 vaccine and analyze the people’s attitude and reaction towards the vaccine and also analyze how it has affected and continues to affect the elections, economy and people’s life. Comparing the positive and negative tweets on COVID-19 vaccine. Also, with the ongoing pandemic situation, it will have an impact on the 2020 election. Therefore, both hashtags “covidvaccine” and  “election2020” would be collected during the month of September 2020.

5. Methodology (including how to collect the data, how to analyze the data, how to present the results, etc)  
  
Question 01: What are people's attitudes towards the COVID-19 vaccine? Are people generally fearful or hopeful?

Covid-19 pandemic has brought the entire world to a stop. People are eagerly waiting for the discovery of a vaccine so that they can get back to their normal lives. However, the race for a vaccine has led to many people questioning whether the vaccine found will be safe or not. Twitter is a place where people have been voicing their hopes and fears about the Covid-19 vaccine.  
 In order to solve our first research question of people’s attitudes towards the Covid-19 vaccine, we will conduct a thorough sentimental analysis of tweets collected during the month of September 2020. The tweets will be collected using Python library ‘Tweepy’ which we can query Twitter’s API for tweets with hashtags “covid-19vaccine”, “COVID-19Vaccine”, “covidvaccine”, and “coronavaccine”. The tweets are people’s opinions so they will most definitely contain many errors and will need to be cleaned using the Text Preprocessing techniques.  
 After we have our cleaned data, we can apply Python packages and machine learning algorithms to do a Sentimental Analysis.  
 Our approach towards Sentiment Analysis uses the Machine Learning algorithm, Naive Bayes Classifier. The classifier assigns a probability to each tweet whether it’s positive or negative. We will use a prelabeled set of tweets (e.g. Niek Sanders’ Corpus of over 5000 hand-classified tweets), build a word feature vector, and use it to train our NB Classifier on this training data. Our covid-19 vaccine tweets will form our test set, and we will use our NB model to classify them as positive and negative (Al-Masri, 2019).   
 We are also considering using another Machine Learning Algorithm: Maximum Entropy. Since we have a binary classification problem, this method is the same as Logistic Regression. We will again use a prelabeled set of tweets to train our model. We will have to vectorize our prelabeled tweets first because ML algorithms need numbers not text. We will vectorize using tf-idf. After our Logistic Regression model is trained, we will use it to classify our covid vaccine tweets. (Singhal, 2020). Our final decision on which model to choose for Sentiment Analysis of covid vaccine tweets will depend on which model gives the best accuracy.

Question 02. How is the COVID-19 vaccine related to the 2020 election? What are the most related tweets regarding the COVID-19 vaccine and the 2020 election?

To tackle our second question, we will conduct a topic analysis and a word co-occurrence analysis on tweets collected during the month of September 2020. The tweets collected will have both hashtags “covidvaccine” and  “election2020” together. We will perform preprocessing tasks so that the tweets are ready for analysis.

For topic analysis, we will convert our tweets into vectors using tf and build a sparse matrix. We will make sure to remove rare and common words because they will not be helpful in determining the topics in our tweets. Then we will fit the LDA (Latent Dirichlet Allocation) model to our tf vector, with number of topics between 2 and 5. We will output the words in each topic along with their weights. After interpreting the topic words we will be able to have an idea about the relation between covid vaccine and election 2020 (Coding Club, n.d.).

For word co-occurrence analysis, we will use bigrams in nltk. We will make sure to remove rare and common words. Bigrams will help us identify all two words that appear together in our tweets. We will store all the bigrams in a flat list and count the number of times each bigram appears in all our tweets. By selecting the 20-25 most commonly occurring bigrams we will be able to get an idea about the information stored in our tweets relating to covid vaccine and election 2020 (Morrissey, et al., 2018).

References Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020, April 21). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7175788/>

Al-Masri, A. (2019, February 13). Creating The Twitter Sentiment Analysis Program in Python with Naive Bayes Classification. Retrieved from [https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in- python-with-naive-bayes-classification-672e5589a7ed](https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-%09python-with-naive-bayes-classification-672e5589a7ed)

Coding Club. (n.d.). Topic Modelling In Python Unsupervised Machine Learning To Find Tweet Topics. Retrieved from <https://ourcodingclub.github.io/tutorials/topic-modelling-python/>

Johnson, A. F., Pollock, W., & Rauhaus, B. (2020). Mass casualty event scenarios and political shifts: 2020 election outcomes and the U.S. COVID-19 pandemic. *Administrative Theory & Praxis, 42*(2), 249-264. <https://doi.org/10.1080/10841806.2020.1752978>

Morrissey, M., Wasser, L., Diaz, J., & Palomino, J. (2018, February 05). Analyze Co-occurrence and Networks of Words Using Twitter Data and Tweepy in Python. Retrieved from [https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to- apis/calculate-tweet-word-bigrams/](https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-%09apis/calculate-tweet-word-bigrams/)

Ned Temko Correspondent. (2020, August 11). The COVID “vaccine war” that everybody loses. *Christian Science Monitor*, N.PAG

Singhal, G. (2020, July 01). Building a Twitter Sentiment Analysis in Python. Retrieved from <https://www.pluralsight.com/guides/building-a-twitter-sentiment-analysis-in-python>

6. Your timeline to work on the project.

**Oct. 7** - Drafted the proposal for term project/ created the PowerPoint presentation.

**Oct. 8** - Continued drafting the proposal for submission (focused on the research questions/ methodology section).

**Oct. 9** - Briefly gone over the draft/ research questions/ methodology so that everyone is on the same page.

**Oct. 11** - Term Project Proposal due at 11:59 - Melissa will submit the assignment for the group.

**Oct. 14** - Revise Term Project and collect all data for the month of September in 2020 (split sections among group members to work on).

**Oct. 21** - Continue working on the Term Project (project background/introduction, methodology, discussion, conclusion, limitation/future scope).

**Oct. 28 -** Group meeting/catchup (go over the Term Project paper/ Term Project Powerpoint).  
\*Side note: Application deadline on Monday, November 2, 2020 for submissions for iConference Student Symposium, don’t forget to register (optional).

**Nov 2-8** - Continue working on the Term Project (no class meeting), wrap-up Final Report and Term Project PowerPoint; divide slides among group members.

**December 2** - Class summary. Term project progress. Prep for Term Project Presentation.

**December 9** - Term Project Presentation | Term Project Final Report Due on December 8 Midnight.