Sentiment analysis of COVID-19 vaccine Tweets

# **Gourav Verma DSC-680 302, Spring-2021** [*Portfolio Page*](https://garv3007.github.io/GouravVerma/)

Please find project proposal on Github link: <https://github.com/GARV3007/COVID-19-vaccine-Tweets>

* **Any surprises from your domain from these data?**

These days Natural Language Processing (NLP) is the breeding ground of research in Data Science. Sentiment analysis is one of the most common divisions of NLP. This domain has diversified the way businesses work due to extensive application usage in creating market strategies, opinion polls, chatbots, etc. For sentiment analysis, NLP has made the processing of thousands of text documents in seconds, which will take hours to process manually. For word embedding, unsupervised and semi-supervised techniques are more popular, however, many sentiment analyses use handcrafted features. The major work in this project is to clean the text data and train the model to understand the language of Twitter. As tweets contain emoticons and short words, it is difficult to correctly recognize the context of the tweets. Tweet sentiments are classified into three types Positive, Negative, and Neutral.

* **The dataset is what you thought it was?**

The data present in the Kaggle for COVID-19 vaccine tweets is getting updated daily. However, the text seems to be incomplete, as many tweets end with ‘…’ proceeding, with a link for the tweet. I need to explore more if the actual sentiments can be extracted from it. I have requested the author of the data to use *“tweet\_mode=extended”* in his extraction process to get complete tweet text. As of now, I am proceeding with the data in hand. This dataset contains tweets from all over the world. To train the model, I am thinking to use the tweets dataset with existing sentiment classifications, then after training the model, it will be applied to the COVID-19 vaccine tweet dataset. It will help the model to understand the tweet language.

* **Have you had to adjust your approach or research questions?**

No, not yet. The research is going as per the plan. I am focused on the usability of machine learning to identify sentiments with high accuracy using tweet texts. However, as I mentioned earlier, it will be helping to get higher accuracy by using the existing tweet sentiment dataset to train the model. I am still not sure on that part. It will introduce some challenges as not all the texts are complete. Overall, my goal would be to create a highly accurate, unbiased model as far as possible.

* **Is your method working?**

I started with cleaning the data by removing URLs, tokenizing text, removing emails, newline characters, single quotes, & punctuations signs, lowercase all text, and detokenize the text. In the end, convert all the text to NumPy array for model input. Some exploratory data analysis will also be performed. Before model building, numerical features will be extracted. I will be using SingleLSTM, Bidirectional LSTM, and 1D CNNs to train the dataset. For validation of the best model confusion matrix will be used. The final model will be applied to the whole dataset to get the sentiments of each tweet. The outcome dataset will be used to plot time series plots to see the variation in emotions.

* **What challenges are you having?**

The biggest challenge is to complete the project on time and the next if I need to continue with the same dataset or not. I am afraid after doing all the work, if the results are not satisfactory, then I might need to change the dataset. Also, another point, how to train the model to understand the twitter language? Whether to use another sentiment dataset or just train on the dataset I have for the vaccine.