**Hello and welcome everyone. Thank you for you joining me. I am Brian Han I will be presenting Covid Vaccine Card sentimental analysis using NLP. But first a little about me is that I am a recent SFU grad in Criminology, and am a guy who loves his dog too much. As for my hobbies, I have taken quite the interest of crypto, stocks, and golf over the past year due to covid. Furthermore, in my previous job, I was working with different sets of data, which got me intrigued about data science and, now it has led me through this boot camp.**

**Moving on, Twitter nowadays has become a platform for people to freely express their opinion; therefore, with the new enrollment of the vaccine card in BC, I wanted to figure out what the sentiment towards that issue was.**

**The presentation will cover how I have obtained the data, how I cleaned the data, which models I have utilized, and the results, and finally the sentiment.**

**I was trying to gather tweets by using twitter api; however, I came across a complication to where I couldn’t get any more tweets once I exceed my limitation. Therefore, my initially data wasn’t quite sufficient.**

**However, I used a tool called TWINT, where I was able to scrape tweets without the limitation of the twitter api. TWINT works by manually scraping through twitter, gathering all the respective information about the tweet, and compiles them into a dataframe. So it does take quite a long time for all the information to be scraped. When querying through TWINT,**

**I used the keyword ‘VACCINE CARD’ as the foundation of my data. Twint, will manually go through twitter and scrape any tweet that has the key words ‘Vaccine Card’; and then will compile all of them together.**

**The data was gathered from sept 7th when the vaccine card was first announced, all the way to Sept 28, three weeks after the initial news. I gathered tweets from BC due to the news being the most relevant. Therefore I specified the query by geo-tagging the location, which allowed me to get BC specific tweets. Because of that, I was able to compile 20,000 tweets for my analysis.**

**To prepare the data for my model testing and sentimental analysis, I had to preprocess it. Which mean, I need to clean the dataset by removing URLs, Punctuation, Mentions, Stopwords, and Numbers. Then I tokenize the words, which splits the text into individual words, and then applied stemming and lemming which returns the words to their original state. Then I use an algorithm called TF-IDF which runs through the data to measure the importance of the word given its frequency.**

**To conduct my analysis, I utilized a tool in python library called TEXTBLOB which is a tool for processing textual data. Here is one of the results where we can see that the sentiment is majority is positive and minority being the negatives.**

**Next we have our word cloud for the most commonly found 250 words in all the tweets.**

**Now coming back to model testing, I wanted to test which classification model is the best to determine the accuracy to my current dataset. Which is why I split the dataset into a train/test model by allocating 70% as training, 30% as testing. To train my model; however, I had to remove the NEUTRAL classification from my dataset because the model we will be solely training to classify a positive or a negative sentiment correctly.**

**As for the models, I wanted to compare the accuracy between multiple models; therefore, I compared Naïve Beyes, Logistic Regression, and SVM to determine the best fit model for this dataset.**

**As a result, it shows that Logisitic Regression has the highest accuracy with a score of 89% and ROC score of 0.86. Therefore we can conclude that logistic regression is the best model for this dataset in comparison to the other models.**

**I will be able to share more details on the results per different model in my booth after the presentation.**

**As my final remarks, this analysis gives us an insight as to the public’s opinion regarding the vaccine card, and thankfully knowing this, it is looking quite hopeful in the application, as the majority seem to have a positive sentiment towards it. Projects such as these, can provide a different forms of insight on monitoring responses from action plans; therefore, companies can take reactionary or responsive measures in respect to the sentiment.**

**In addition, I plan on revising the project by utilizing VADER as one of my tools for my analysis because VADER can analyze emoji’s to calculate the sentiment and polarity behind the text, which may provide different results. Thank you for listen!**