Data Visualisation Project 2020

Sean Gavin - A00251388

Date 11/12/2020

STUDENT PLAGIARISM DISCLAIMER FORM

PLAGIARISM DISCLAIMER

STUDENT NAME: Sean Gavin

STUDENT NUMBER: A00251388

PROGRAMME: Bachelor of Science (Honours) in Software Design with Cloud Computing

YEAR: 4th Year

MODULE: Data Visualisation 4

LECTURER: Dr David Scott

ASSIGNMENT TITLE: Data Visualisation Assignment

DUE DATE: 11/12/2020

DATE SUBMITTED: 14/12/20

ADDITIONAL INFORMATION:

I understand that plagiarism is a serious academic offence, and that AIT deals with it according to the AIT Policy on Plagiarism.
I have read and understand the AIT Policy on Plagiarism and I agree to the requirements set out therein in relation to plagiarism and referencing. I confirm that I have referenced and acknowledged properly all sources used in preparation of this assignment. I understand that if I plagiarise, or if I assist others in doing so, that I will be subject to investigation as outlined in the AIT Policy on Plagiarism.
I understand and agree that plagiarism detection software may be used on my assignment.
I declare that, except where appropriately referenced, this assignment is entirely my own work based on my personal study/or research. I further declare that I have not engaged the services of another to either assist in, or complete this assignment.

Signed: Sean Gavin

Dated: 11/12/20

Table of Contents

INTRODUCTION **4**

SYSTEM OVERVIEW **4**

ARCHITECTURE 4
DESIGN DECISIONS 5

TOPIC SELECTION AND PRESUPPOSITION **9**

POST-PROCESS ANALYSIS **9**

CONCLUSION **12**

WHAT YOU WOULD DO DIFFERENTLY 12
FUTURE WORK 12

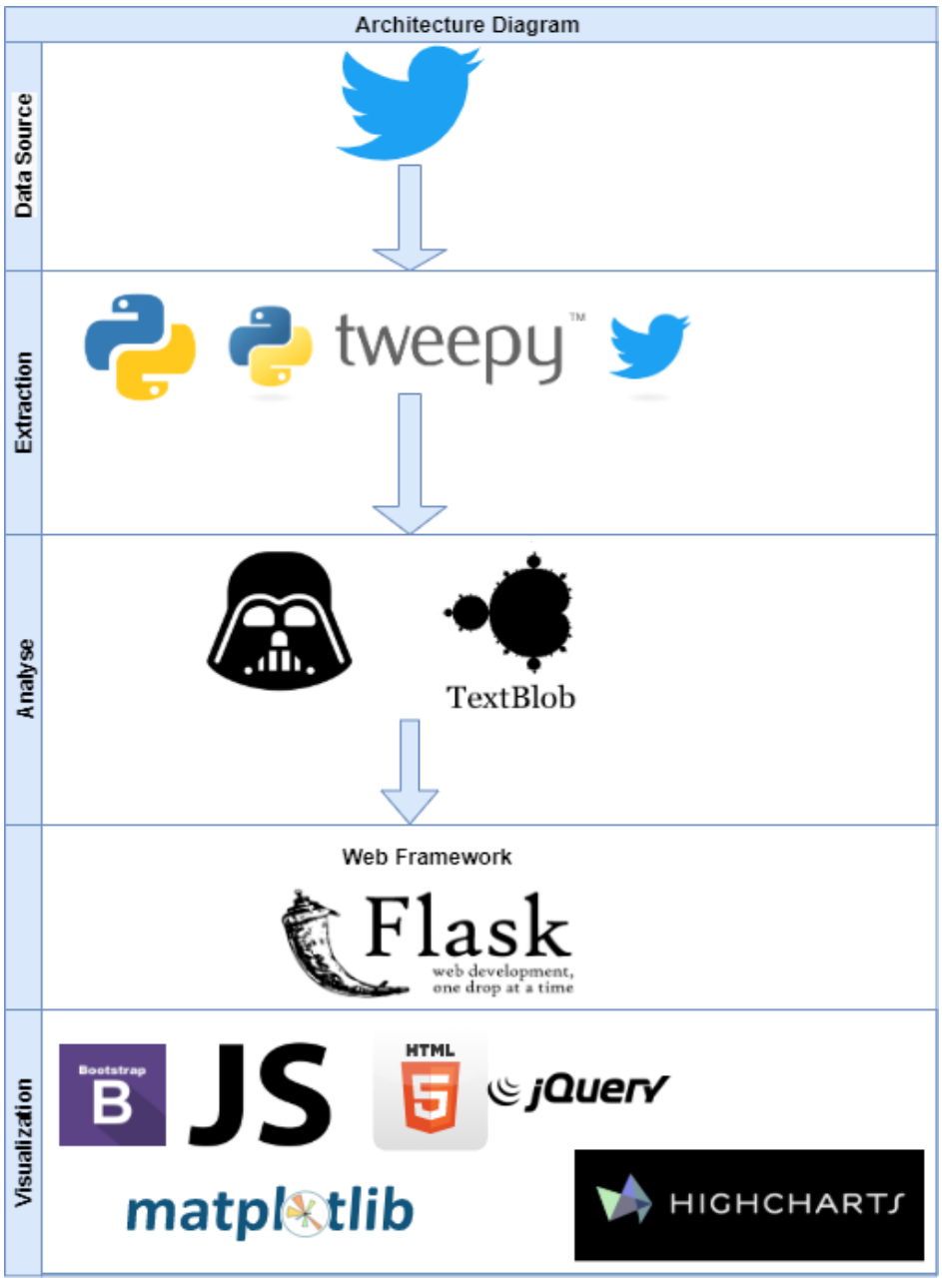
Introduction

Our goal during this assignment was to consume tweets from the TwitterAPI, utilising the TweepyAPI, and to hold out sentiment analysis on those tweets while also providing the utilization with a graphical based visualisation of the analysis through the use of live graphs.

Ideally, we'd be consuming and updating the graphs in real-time with live data and providing a drill-down to permit the user to achieve more insight than simply the tweet sentiment. Sentiment analysis is essentially opinion mining, and it is the automated process of analysing text data, and classifying consumer opinions as positive, neutral or negative.

System Overview

Architecture



Design Decisions

The application is mainly developed using python. Initially, that set the tone for the base technology that needed to be used for the application and plenty of my design decisions stem from that. It had been suggested that we use the Tweepy API and so I also quickly adopted this into my design approach. I implemented TextBlob and Vader as my sentiment analysis tool. Initially I attempted to stream the tweets to a CSV file and then read them from there when I needed them, but this became too difficult. Instead, I just read in data in the object and parsed the data or manipulated data in order to be presented on the webpage.

I wanted to draw the user's attention specifically to the chart which display the sentiment analysis. This is the core aspect of the application as it contains all of the sentiment data contained in the application. In the app we have two pie chart graphs, one for vader and one for textblob to show the sentiment breakdown for the chosen topic pfizer. I also have a bar chart showing the breakdown of device types where the tweets are coming from and drilldown which shows the breakdown of the sentiment for each device types depending on what two topics which the users picks. The stream of tweets is also displayed on the webpage when they are received.

Navigation is pretty straight forward as the user only has two input fields and one submit button that allows them to choose the two topics they wish to visualise on a chart. This gives the user the ability to compare and contrast two separate or connected topics. The user also has the option to export the data which is streamed to the webpage by clicking export csv button.

Textblob vs Vader Sentiment Break Down:

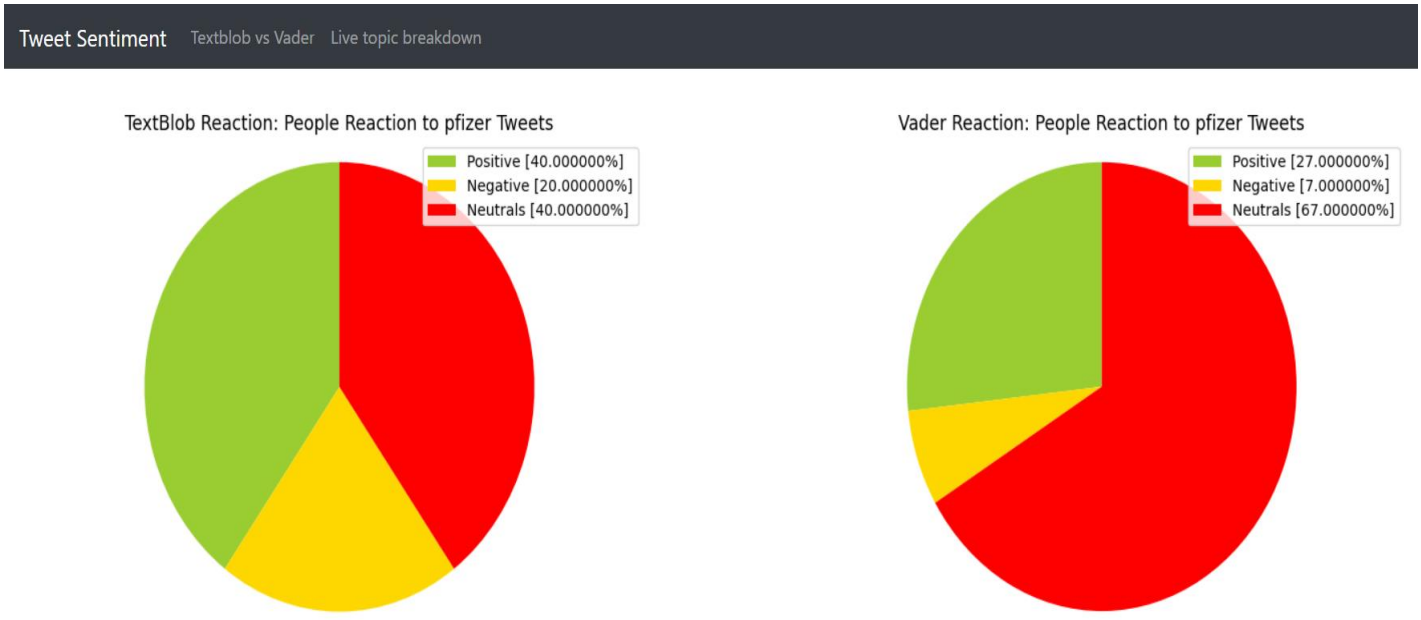


Figure 1 TextBlob vs Vader

Table breakdown for the selected topic “Pfizer”:

Profile Picture	User Name	Location	Tweet	Blob Sentiment	Vader Sentiment
	Katei2020	United States	@ChuckModi1 covid pfizer vaccine Washington DC protest Proud Boys Antifa march for Trump stabbed Trump rally Trump... https://t.co/2pRLEeFLz8	positive	negative
	CraigMirage	Madchester	So the CCP Chinese Communist Party has infiltrated many governments and companies all over the globe 🌎 including va... https://t.co/7wp4raY1XD	positive	positive
	PoliticalSense1	United States	https://t.co/ZHAhuHfeY5	neutral	neutral
	BritishVogue	London	This week, the first person in the world was given the Pfizer Covid-19 vaccine in the UK. https://t.co/DVAnMKlAtd	positive	neutral
	diana4candor	California, USA	@MMacky4 @MattMcCayAgent @freep @pfizer @FedEx @UPS I can't EVEN with that guy. No educating a guy that thinks he k... https://t.co/CNKfXYvUCQ	neutral	negative
	WallyfromC		@80sgirlforever @janbear105 @TomPodolec @Weathermetre @pfizer @UPS_Canada @UPSAirlines @TorontoPearson Major courie... https://t.co/N3HdkQs2iH	positive	neutral
	petitlarcenous	United States	@ThomMarshalljr @jonsie9109 @jampony25 @withinthemonth @chadmbol @petemuntean https://t.co/HuoNa5BerO	neutral	neutral

Figure 3 Table breakdown

Live Search Topic Breakdown bar chart for topic one “Pfizer” and topic two “Moderna” :

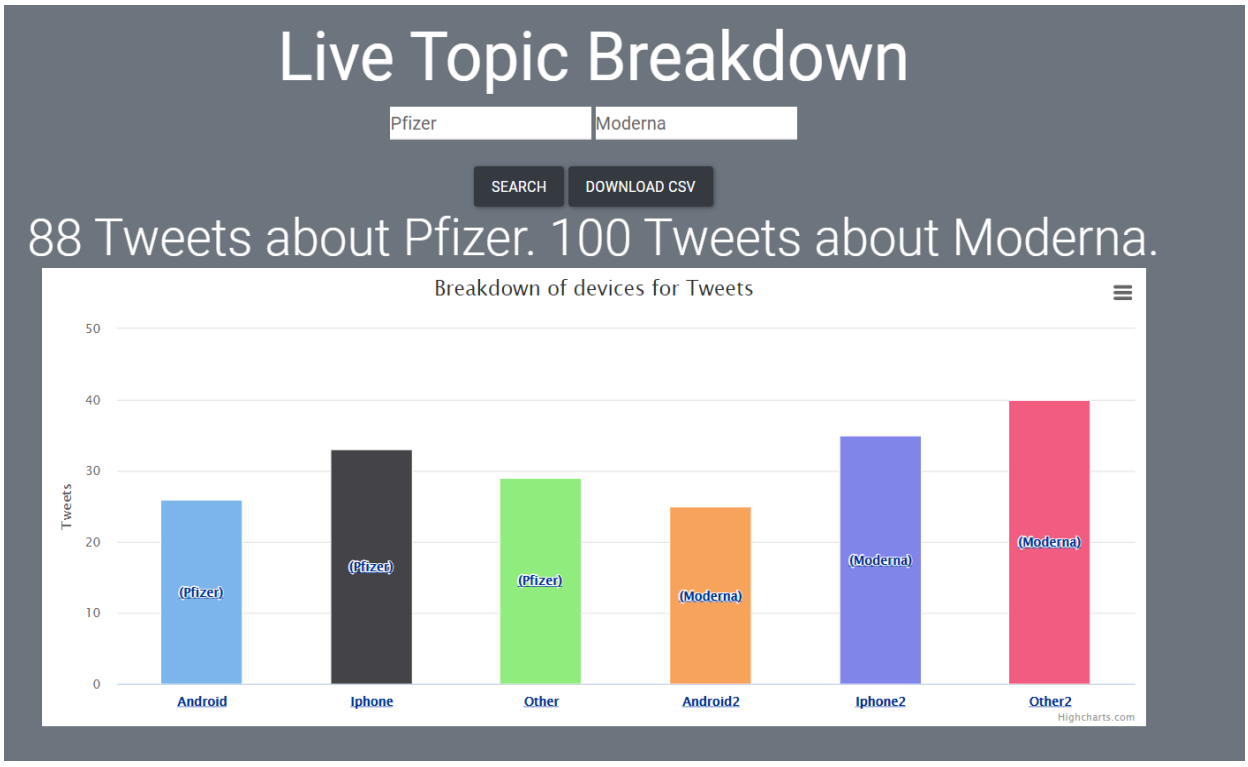


Figure 2 Bar chart Live Topic breakdown

Drilldown for android device sentiment for “Pfizer” topic:

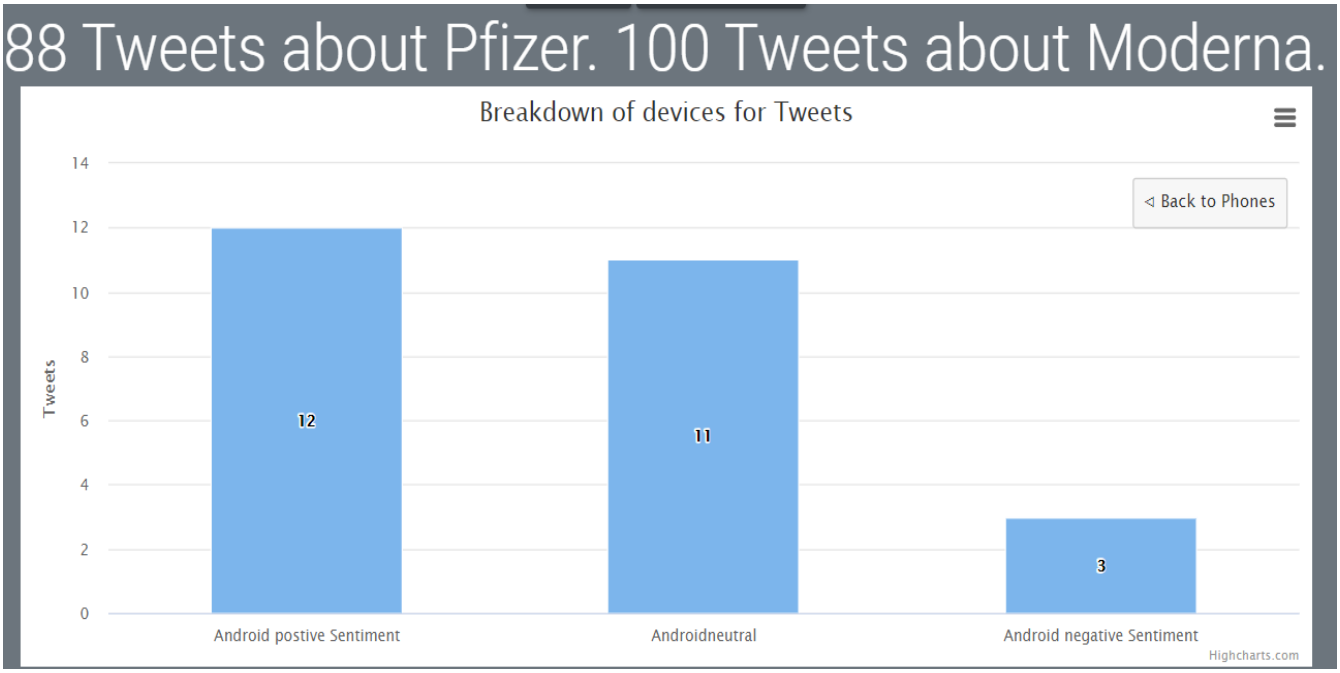


Figure 4 Drilldown for topic one

Drilldown for android device sentiment for “Moderna” topic

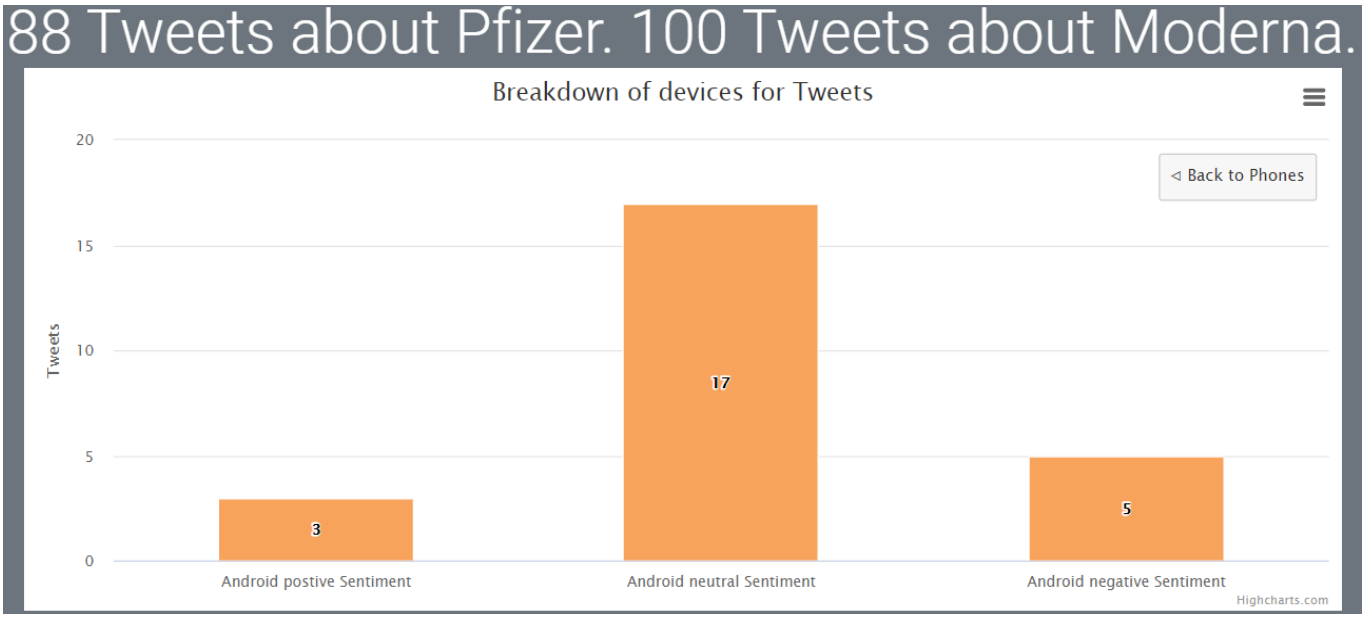


Figure 5 Drilldown for topic two

The live tweets for the first search term are shown below in figure 6. The colour of the row shows the deemed sentiment for the tweet with blue as neutral, green as positive and red as negative.

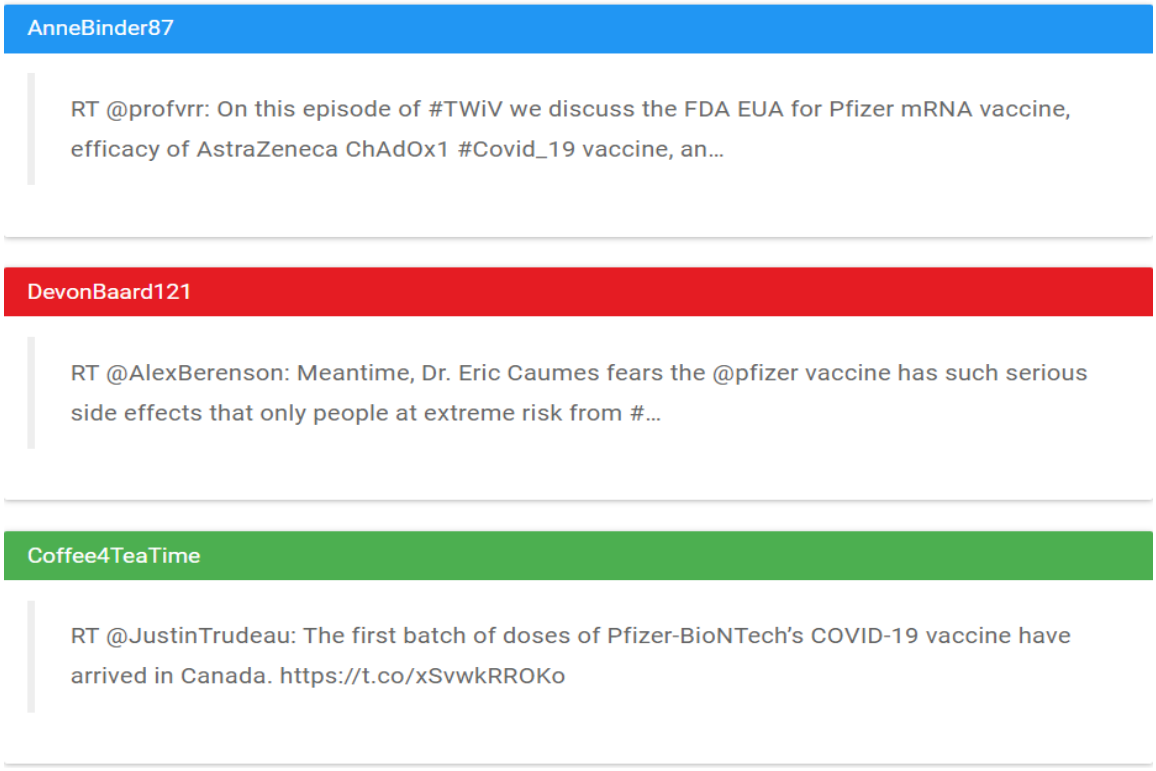


Figure 6 Table for topic one

The live tweets for the second search term are shown below in figure 7. The colour of the row shows the deemed sentiment for the tweet with blue as neutral, green as positive and red as negative.

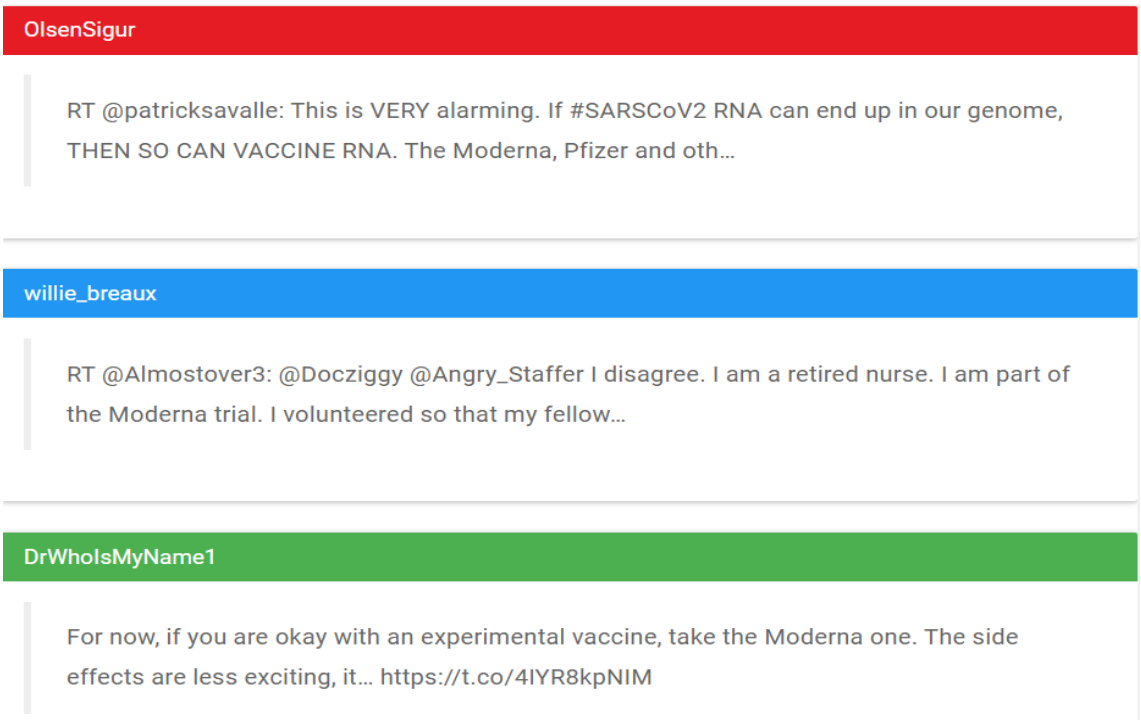


Figure 7 Table for topic two

Topic Selection and Presupposition

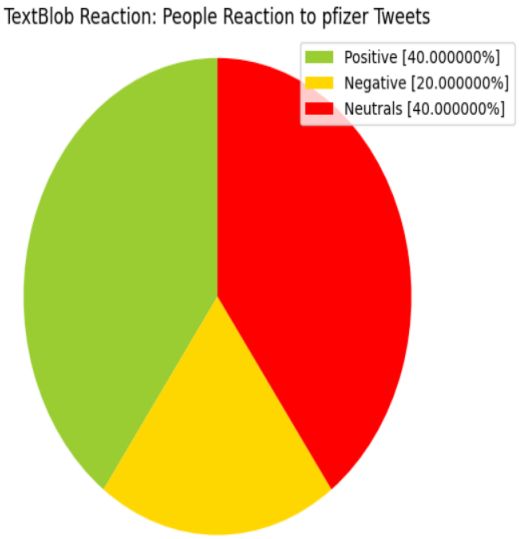
The topic which I selected is “pifzer”. The reason why I chose this topic is due to the controversy and interest that surrounds the development of vaccine for covid-19. The general publics negativity about how quick the vaccine is developed should make for an interesting result. The pre-conceived notions from my point of view would be that we have very few neutral tweets and more so positive and negative as I believe that the vaccine made by Pharmaceutical company may be a hate or love topic.

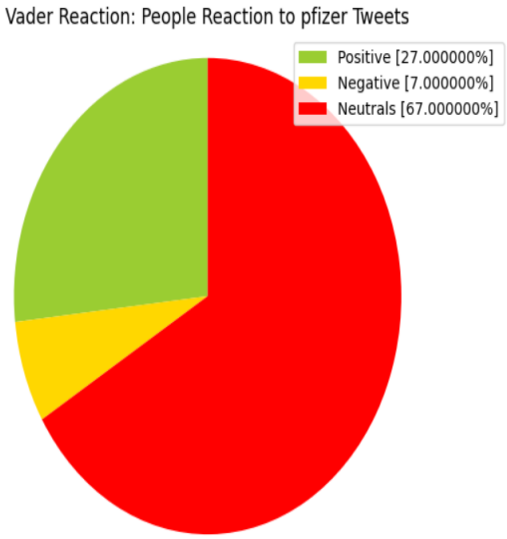
In terms of preconceived notions for drilldowns for the live topics on the device type. I’m not sure what to expect really with no prior knowledge on the foremost popular device used to tweet. I do not imagine the fact that Pfizer is the topic selection will have much effect on the device type. I think, however most people access twitter through their mobile device using the twitter app.

In relation to comparing Vader verses TextBlob, I think Vader will be the better sentiment analyser due to the fact it reads sarcasm better than Text Blob. I’m aware that some sentiment analysing may well be read as positive while it should be a negative tweet. This would be because people try to optimistic even though the tweet was mainly talking about the negative effects that the vaccines have.

Post-process Analysis

The image below is a breakdown of Textblob and Vader sentiment:



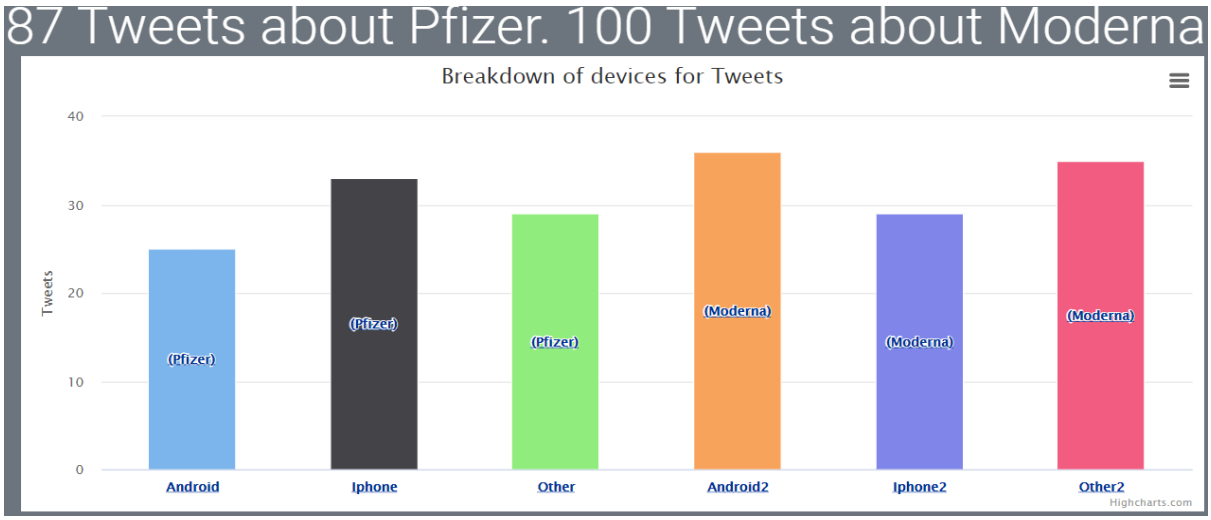


From analysing the graphs, is an overwhelmingly difference between the positive, negative and neutral sentiments. There is at least 12 -27 percent difference between the textblob reaction to the tweets compared to the Vader reaction. Vader reaction seem to be largely neutral, while in the textblob reaction seem to be 40 percent each for positive and neutral. These results don’t verify which sentiments analysis tool performs better and I think a third analysis tool is needed in order to confirm this. It difficult for me to reflect on the topic or even have an opinion on it due to the strange sentiments score being predicted.

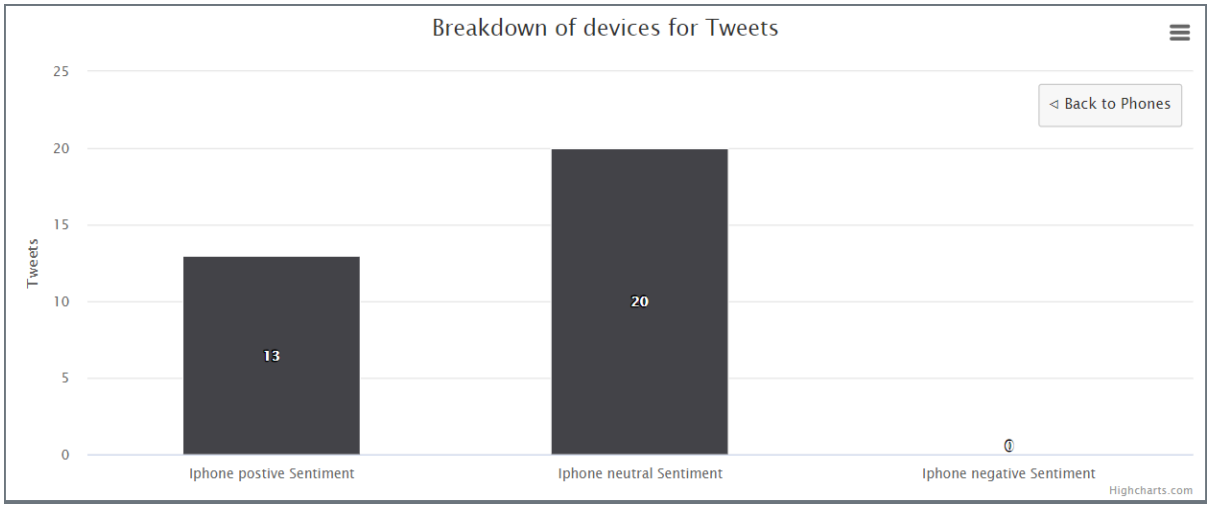
From analysing the table breakdown of sentiment on “Pfizer” maybe reflects that the tweets just don’t contain enough words with a strong sentiment value to determine the sentiment of whole tweet in general. The sentiment analyser for vader and textblob may not be strong enough to detect whether a sentence is positive or not and there’s also the possibility of it not being able to deal with sarcasm well which is quite strong on twitter.

	LouisTenken	@HuskyFan90 @WSJ I don’t think so ,it’s Pfizer hoax	neutral	negative
	mkozberg	@DereckCoatney Pfizer wick? That still isn’t important enough! How about the Tesla wick? Now that would be traumatic!	positive	negative
	theywillnotsil1	@realDonaldTrump He didn't do 💩.Pfizer received no gov funding to make sure the development of the vaccine did not... https://t.co/g5LvJrdIeE	positive	neutral

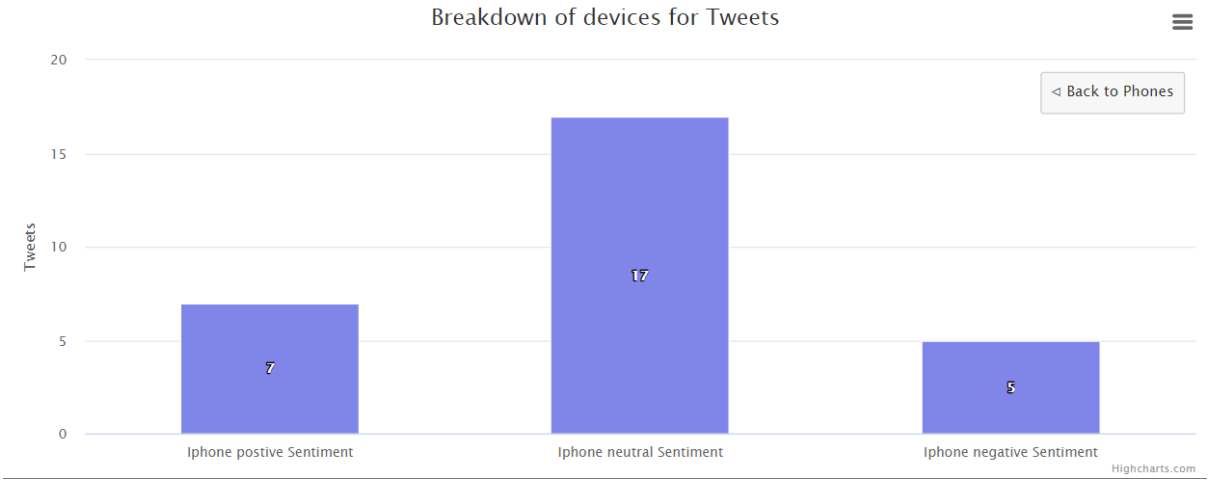
From analysing the bar chart, we see the share of devices is kind of evenly distributed between the three categories. Preconceived notions would have pointed to dominance of mobile phones over other devices which has proven to be true, but I’m still fairly surprised by the share of tweets from other devices which I didn’t expect to be as high as it was. Android and iPhone would share the mobile market evenly so that result isn’t as surprising. As to why other devices share such a large percentage is probably due to people working from home on their work laptop or pc.



Drilldown for Iphone device sentiment for “Pfizer” topic:

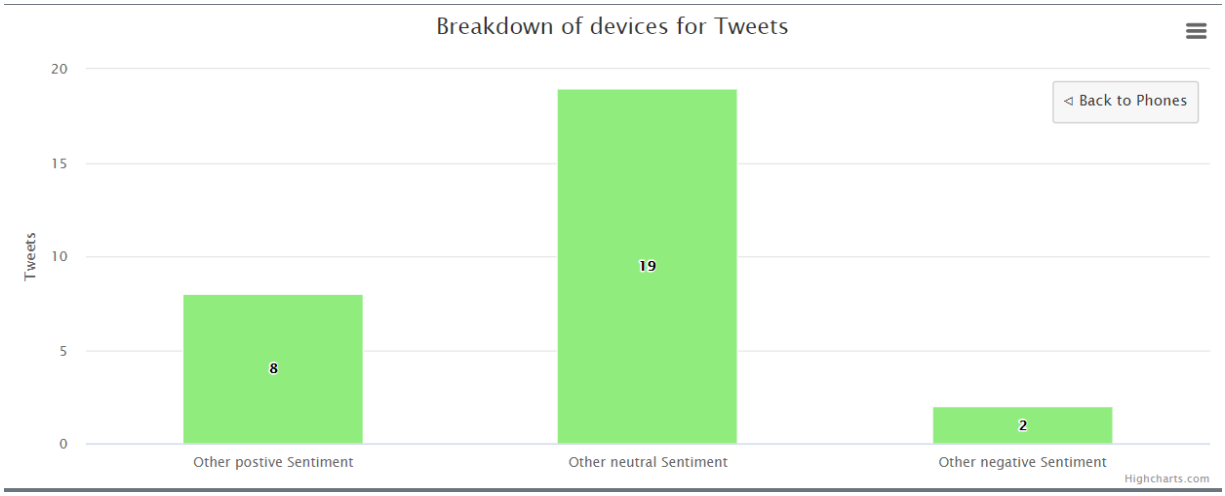


Drilldown for iPhone device sentiment for “Moderna” topic:

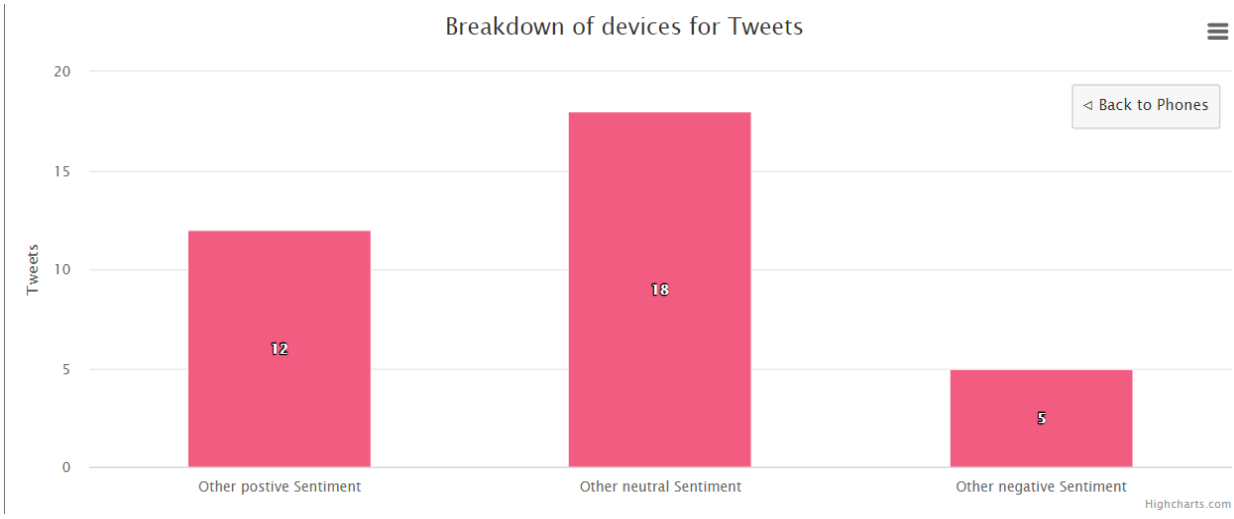


Iphone device sentiment for “Pfizer” seems to be mostly neutral, however they seems to be no users who are negative about the topic. In relation to the “Moderna” topic for iPhone users results seem to be neutral just like the Pfizer topic.

Drilldown for other device sentiment for “Pfizer” topic



Drilldown for other device sentiment for “Moderna” topic



Other device sentiment for “Pfizer” and “Moderna” seems to be mostly neutral. From analysing the two live topics the user tweets seem to be quite positive about the covid-19 vaccines.

Conclusion

What you would do differently

When comparing Sentiment types on the chosen topic I implemented the data using mathplotlib library in python. The major downfall of using this library is I couldn’t expand the graph in order to drill down the data due to matplotlib not being interactive. I should have implemented highcharts like in the live topic section.

In the live topic breakdown graph I used the TextBlob to analyse the sentiment, I used Vader Sentiment in the topic section of the assignment and I felt should gave the ability to the user to choose the Sentiment type for breakdown of the devices. There is no design benefit or loss between these as they are essentially interchangeable, however, I would just like to compare the sentiment analysis between the two to see is one more sensitive than the other in the live topic section e.g. as done in the chosen topic graphs.

I think next time I would use the Tweepy stream listener as I felt the approach, I took with ajax is a hack. I would also implement threading so the application is concurrent with the latest tweets. I might also implement a design pattern like the Model, View Controller pattern to permit new developers a higher chance to grasp what is happening and to boost maintainability of the application. Another thing that I could have done would have been to use a database to store the data as it may have been easier to access the demographic data we were storing.

Future work

I would possibly overhaul the UI/UX of the application. As of right now it’s not bad but I definitely think that it could be more appealing to users. However, this is often a minor detail and users are able to navigate and understand the application sufficiently with its current design so this would not be a high-priority update. I might also seek to use a different charting library, like D3 charts. This might not significantly improve performance, but it would be nice to be able to directly compare the UI and UX properties of one library against another.

Next time around I would implement testing as if this was to be deployed in a production environment then that would be of significant importance to confirm that the application is definitely reparable should one of the services go down. We would also like to add a section that displays the highest trending topics on Twitter in order that the user can search these trends to develop an understanding of the sentiment associated with it. This may be useful, especially for media companies, to grasp public perception on these topics before publishing articles.