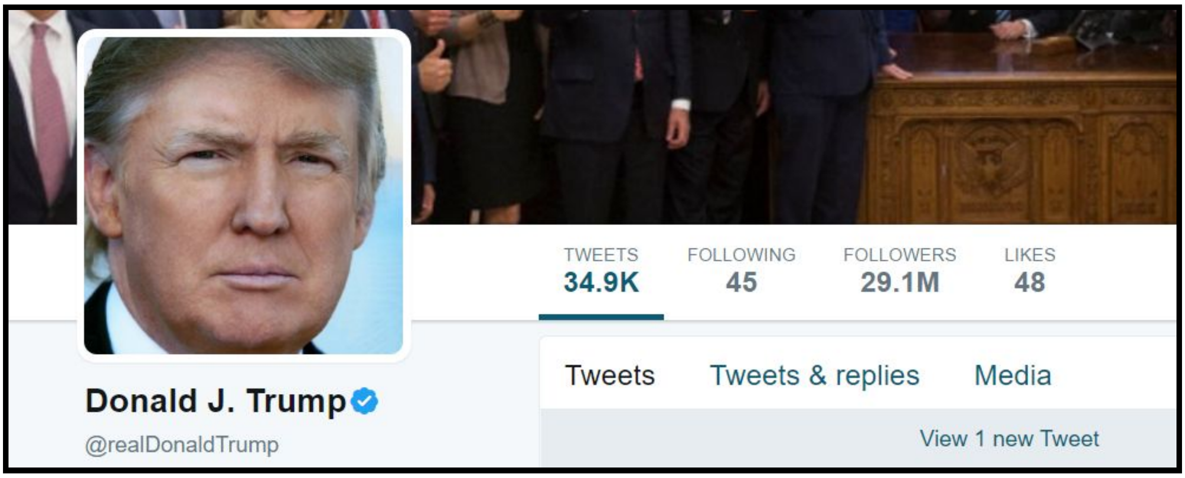
Phase 3: News Analysis

Dataset:

The dataset that we analyzed for the visualization assignment consists of tweets by President Donald Trump's twitter handle @realDonaldTrump. The motivation behind picking Trump's tweet dataset is the number of people interested in every tweet which is evident by looking at the number of followers and retweets. Also, the frequency at which tweets are posted using this twitter handle has been high throughout the presidential election campaign.



TWEETS	FOLLOWING	FOLLOWERS	LIKES
34.9K	45	29.1M	48

Donald J. Trump

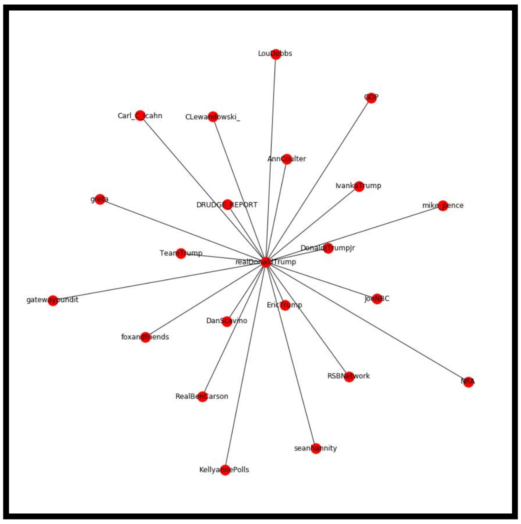
@realDonaldTrump

Tools Used: JavaScript, Python, Tableau, R

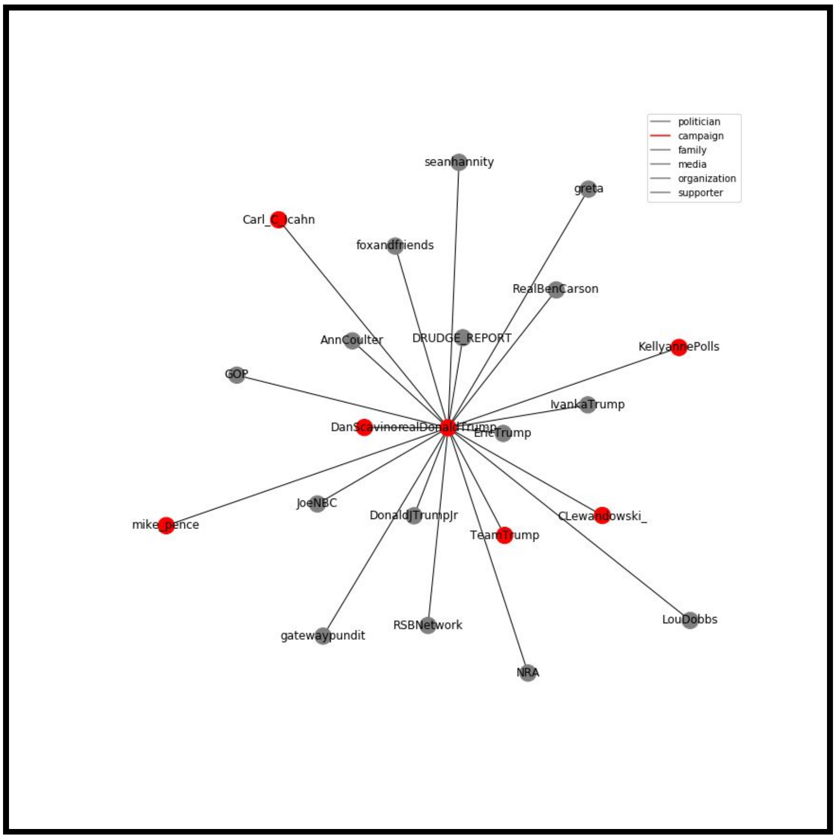
The visualizations developed using the tweets are categorized as mentioned below:

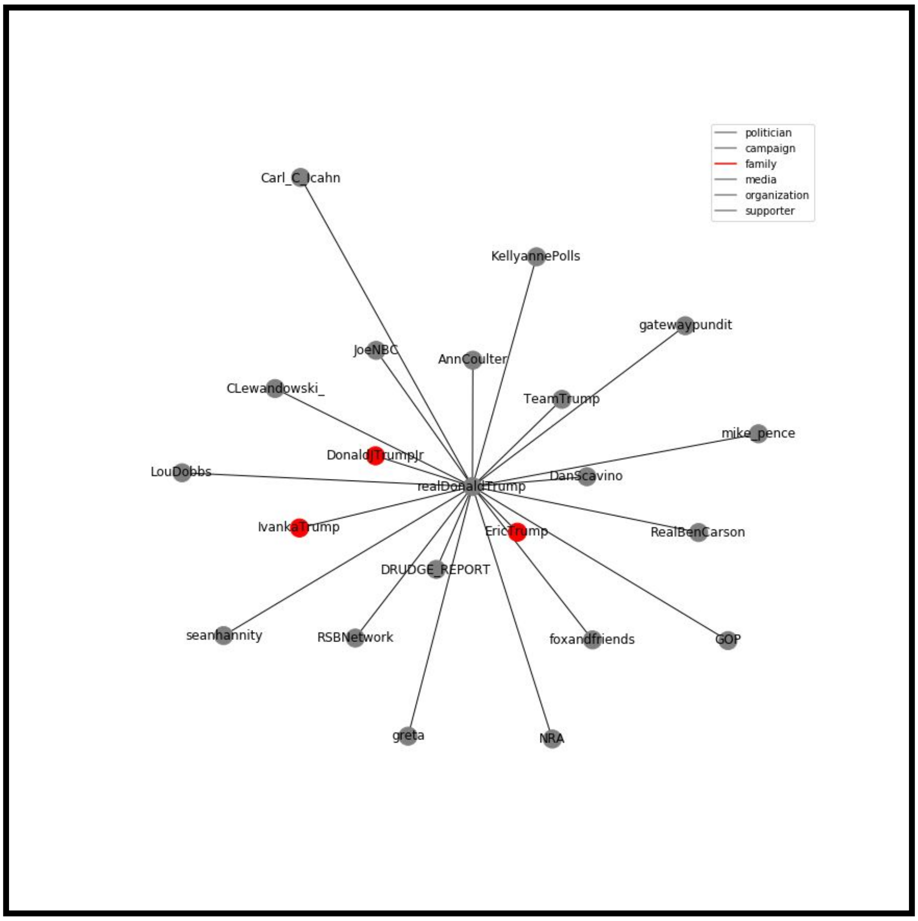
Network Analysis:

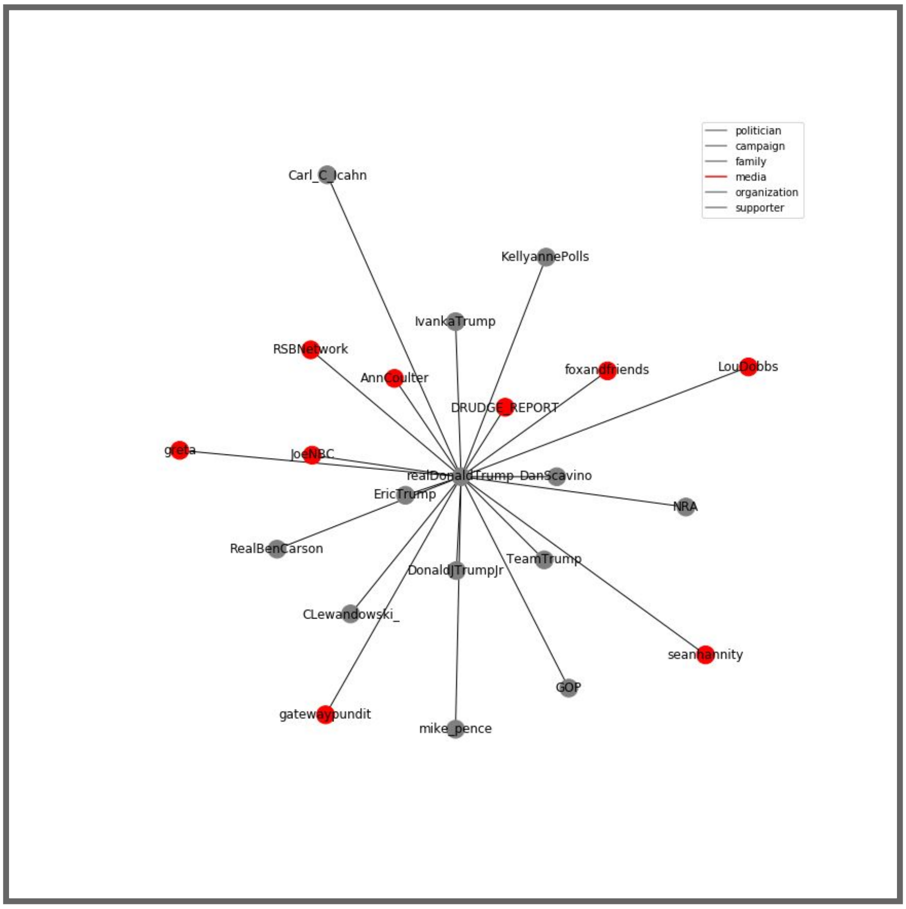
After collecting the tweets, we obtained the twitter handles and number of retweets by President Trump and used the networkx package in python to visualize the network of retweets for Trump. We further classified each of the twitter handle into three categories viz. Media, Family and Campaign.



In the visualization, lesser the distance from the center denotes more the number of retweets by Mr. Donald Trump and hence, closer the node to center, more is the importance and relevance. We decided to highlight **specific categories** to analyze trump's **media bias**, his **family network** to check specific family members Trump retweets more than the others and his **campaign network** to check whether Trump favored a particular twitter handle during the presidential campaign. We obtained the following visualizations:

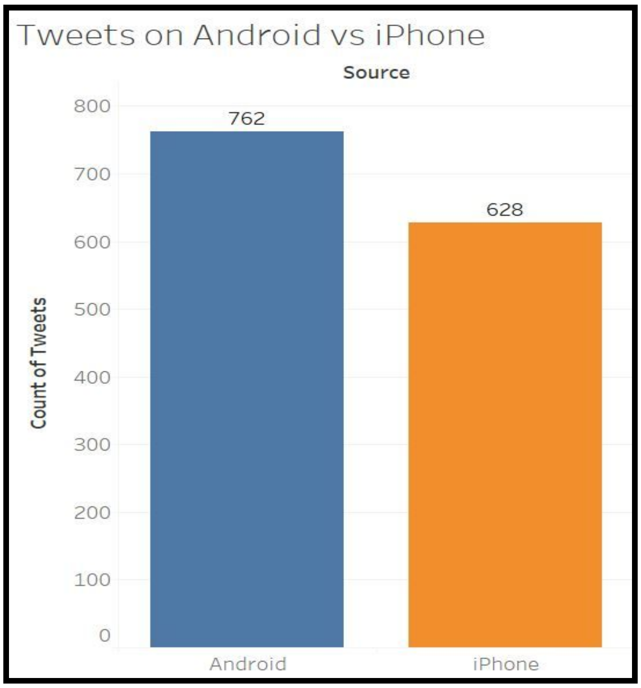



politician
campaign
family
media
organization
supporter
Carl_C_Icahn
KellyannePolls
gatewaypundit
JoeNBC
AnnCoulter
CLewandowski_
TeamTrump
mike_pence
DonaldJTrumpJr
LouDobbs
realDonaldTrump
DanScavino
IvankaTrump
EricTrump
RealBenCarson
DRUDGE_REPORT
seanhannity
RSBNetwork
foxandfriends
GOP
greta
NRA

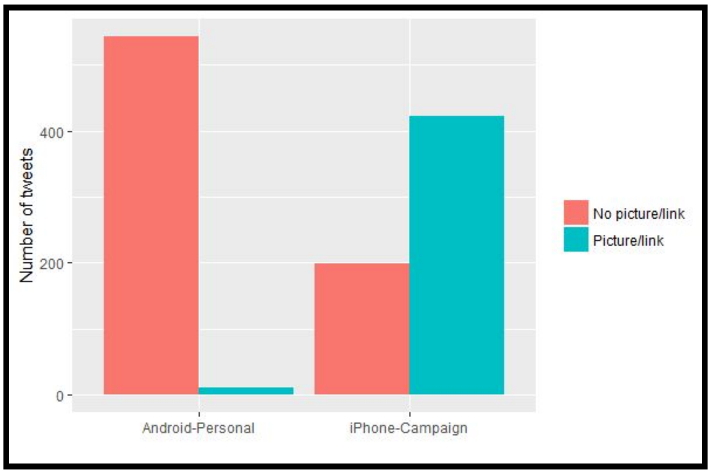

politician
campaign
family
media
organization
supporter
Carl_C_Icahn
KellyannePolls
IvankaTrump
RSBNetwork
AnnCoulter
foxandfriends
LouDobbs
DRUDGE_REPORT
greta
JoeNBC
realDonaldTrump
DanScavino
EricTrump
NRA
RealBenCarson
TeamTrump
DonaldJTrumpJr
CLewandowski_
seanhannity
GOP
gatewaypundit
mike_pence

Exploratory Analysis - Tweet Source

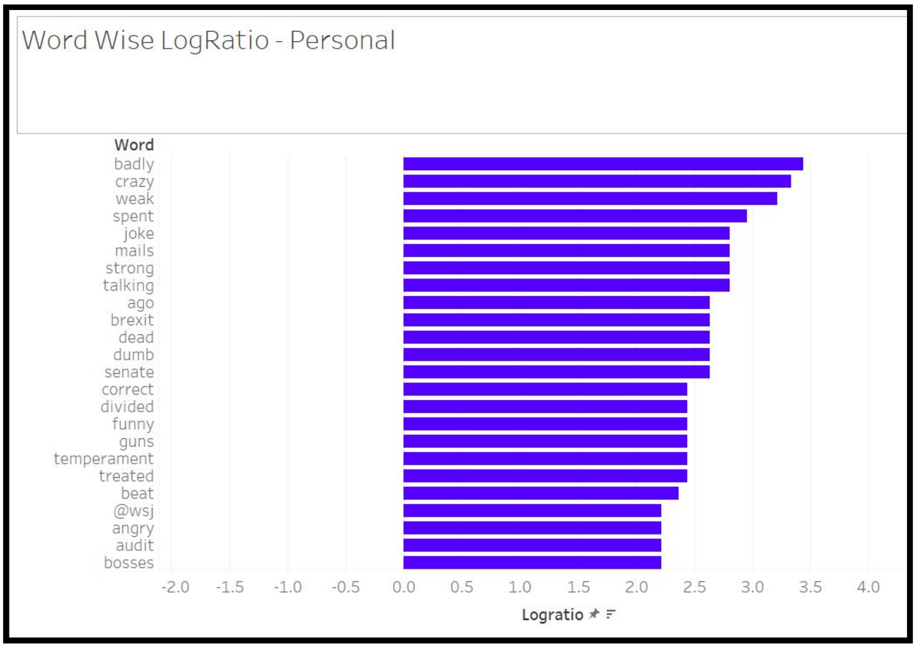
We observed that every tweet has an attribute called the source which was either Android or iPhone. We analyzed the tweets and speculated that Mr. Trump's personal device is an Android one and the campaign uses an iPhone. We first plotted the histogram to check whether the dataset is skewed based on the source for the tweet. We concluded that the dataset has approximately the same number of tweets from android and iPhone.

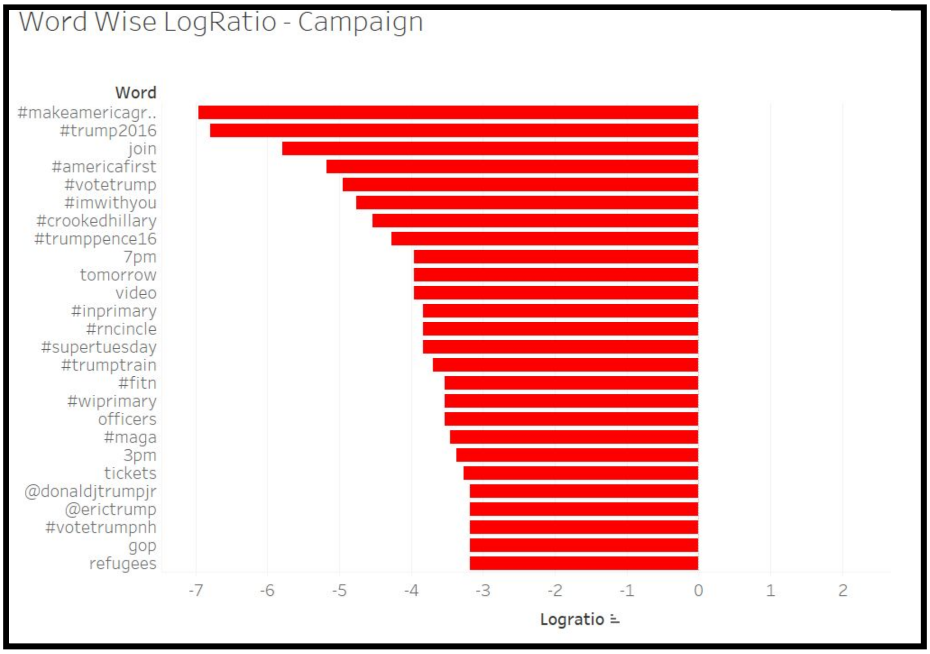


There were tweets with links and pictures related to campaigns and so to support the speculation we tried to analyze these tweets source-wise. We observed that the number of tweets with pictures and links were shared more using the iPhone device.



We also visualized the most frequent words tweeted using both the devices. The visualization clearly shows that the Android device used words like badly, crazy, weak, joke that are commonly used by individuals whereas the iPhone account had more usage of hashtags promoting the campaign like #makeamericagreatagain, #trump2016 and #americafirst.

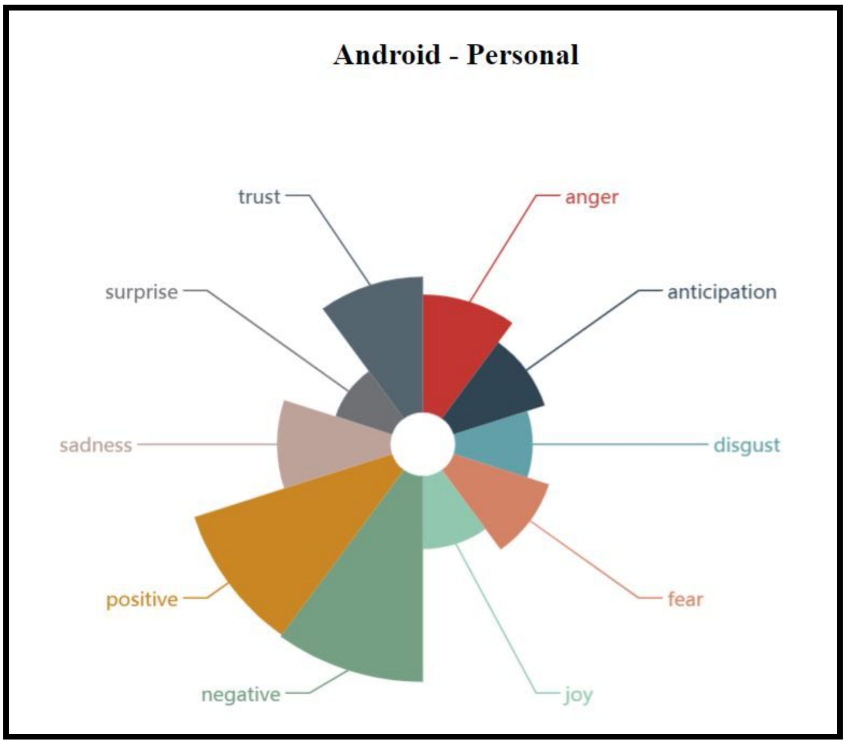


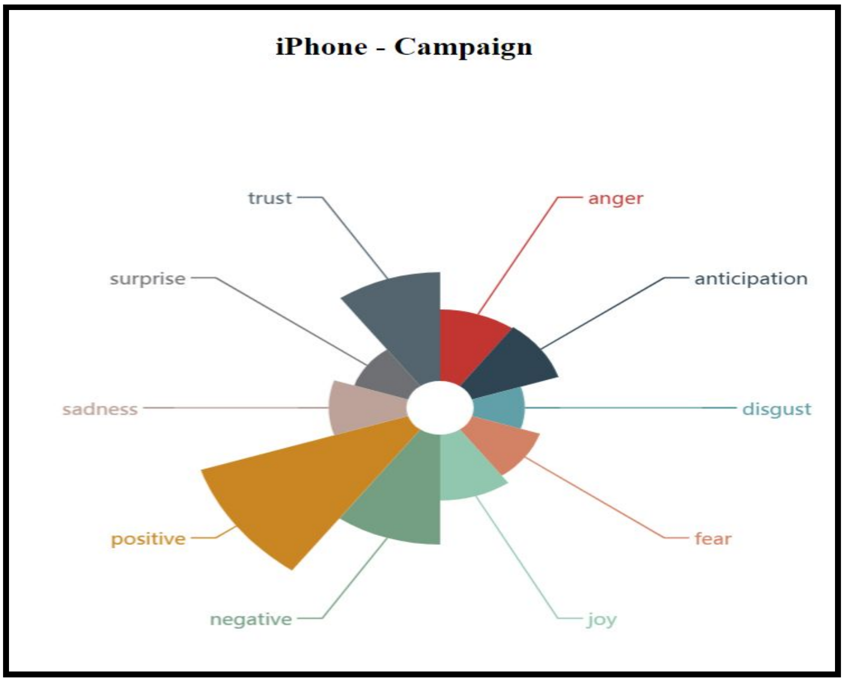


Visualizing Sentiment Distribution

We also wanted to do a sentiment analysis on the content of the tweets to support our claim and so we used the lexicon from the tidytext package for R. We created an interactive visualization using the echarts library in JavaScript and visualized the sentiments for the devices. We observed that the positive to negative sentiment ratio for the android tweets were more or less the same. On the other hand, there was considerably large amount of tweets that belong to the positive and trust category for the iPhone device that belongs to the campaign. The visualization can be accessed here.

Note- Make sure JavaScript is enabled in your browser. If it's enabled, then a popup will show up on loading the page, click "load unsafe transcripts" and the visualization will load.





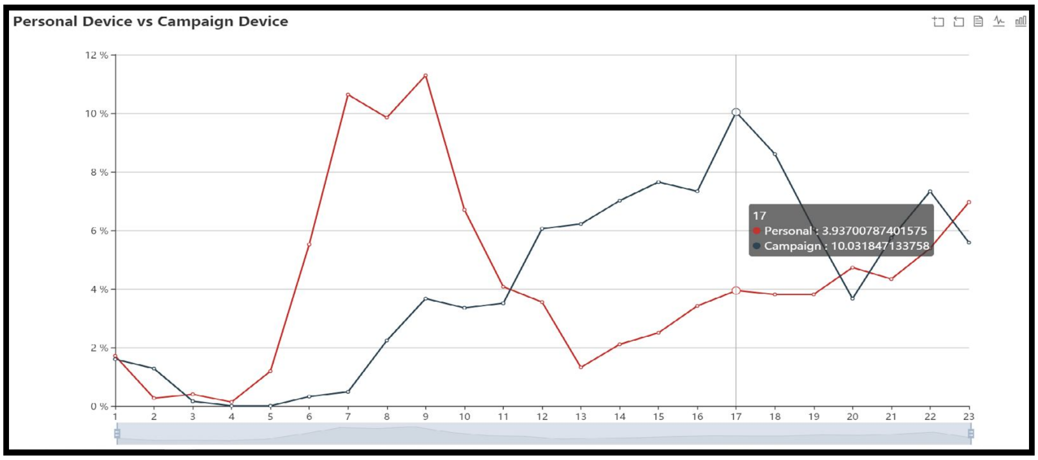
Visualizing Tweets Temporally

We wanted to look at how the number of tweets changed from the two sources (iPhone and Android) as the day progressed. So, we created an interactive visualization using the echarts library in JavaScript which allowed us to view how the number of tweets were changing. A screenshot of the visualization is shown below. The visualization can be accessed here.

Note- Make sure JavaScript is enabled in your browser. If it's enabled, then a popup will show up on loading the page, click "load unsafe transcripts" and the visualization will load.

In the visualization, the x-axis is the hour of the day and can be varied using the slider below the axis. The y-axis represents the percentage of total tweets.

As can be seen in the visualization, tweets from the personal device (android) peak at 9 a.m. in the morning while tweets from the campaign device (iPhone) peak at 5 p.m in the evening.



Methodology

1. I collaborated with Agent Phoenix for the phase 3 analysis. First, we individually went through the data sets to familiarize ourselves with the data. Since the data was available in csv format, it was easy to access and go through the data set.

2. We used **R** to clean and manipulate our data set. For the network analysis part, we added a new column to the data set which classified the twitter handles which trump was retweeting into different categories.

3. Then, we made multiple visualizations (as shown above) to gather insights from the data set. We used a **Python and Tableau** for the network analysis and exploratory tweet source analysis respectively. We also used **Javascript** to create two interactive visualisations which we hosted on terpconnect.

4. Finally, since neither of us had used the HIVEMIND ability till now, we created a google doc to create a common report for this phase.