Group 11
Charlottesville Tweets

Rashad Seaborne, Alex Turse, Cici Brown, Rohan Ravulapalli, Prayank Srivastava

Context

- Data from tweets with/in response to #charlottesville
- From August 15th – 18th, 2017
- A few days after the Unite the Right rally in Charlottesville, VA.

Q1: Qualitative thoughts

- As young people living in the United States, most of us were in high school when this event took place in Charlottesville. It really opened our eyes to parts of the political climate of the country in 2017.

- Due to this, we would like to gain an analytical understanding of what people were talking about right after the rally in Charlottesville, as well as summarize public sentiment right after the event.

- Was it mostly skewed one way or the other? Or was there an even mix of emotions represented in the tweets that we will be looking at?

Data

Q2: Charlottesville Tweets Data Set

Variables:

- Text
- Date
- Tags

Observations:

- Tweets

Data Management & Cleaning

Isolate text and variables (listed above) from 25 original variables.

Remove stop-words, special characters (emoji), convert to lowercase

Randomly sample 500-1000 tweets*

*This step is not always necessary, due to processing resources though we decided it best to sample the data set.

Data Summaries

Summary 1: Term Frequency

Based on a random sample of 500 tweets from August 15th
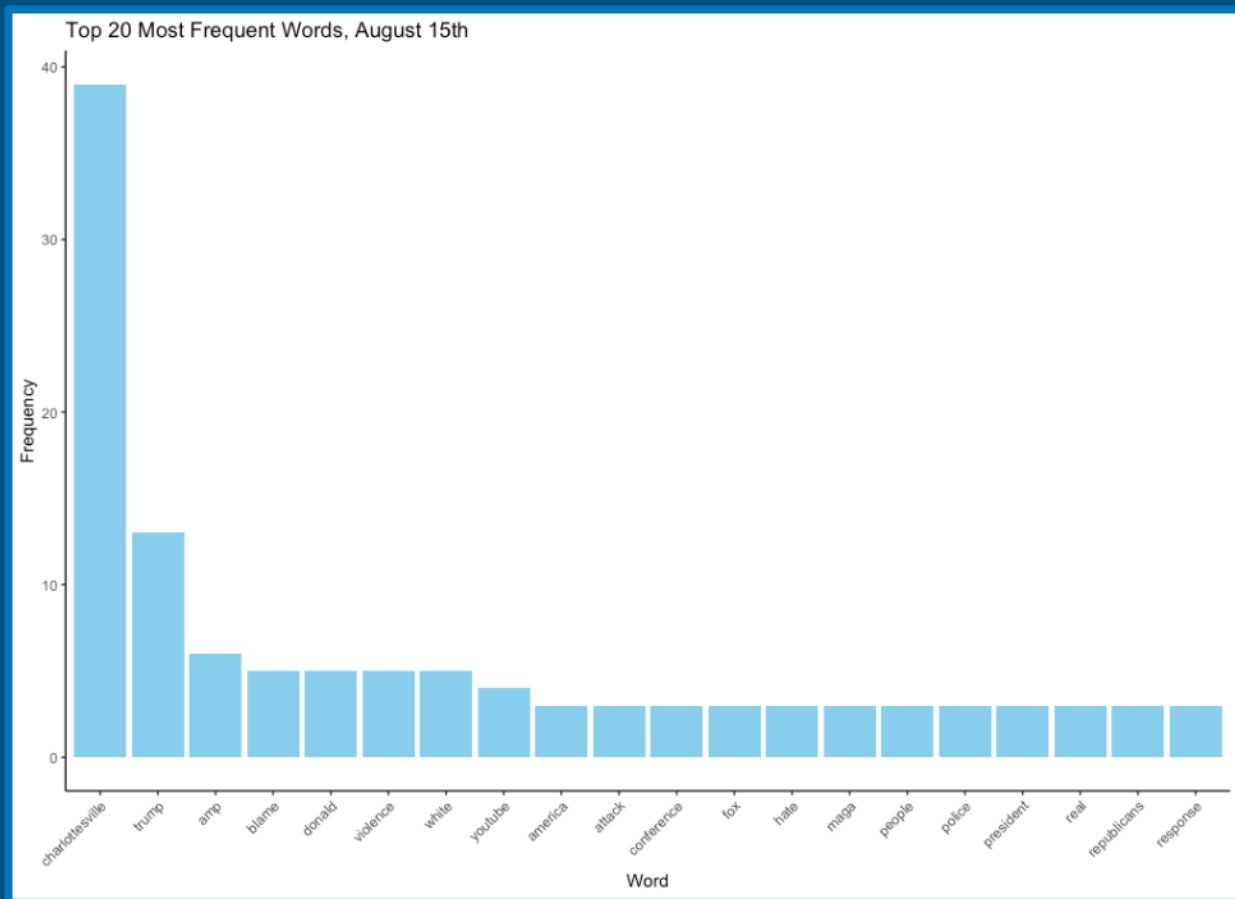
Other than **"Charlottesville"** many tweets contained the following text with their corresponding percentage* of the total text:

"Trump" : 16.4 %
"Violence" : 6.5%
"Hate" : 4%

*It should be noted that these percentages are of the total words that are not "charlottesville" in the top 20

Top 20 Most Frequent Words, August 15th

Summary 2: Sentiment Analysis

Based on a random sample of 500 tweets from August 15th

The **majority** of tweets with #Charlottesville had a **negative** or **neutral** sentiment, suggesting that many users were **upset** during that time.

Sentiment Distribution of Tweets from #Charlottesville on August 15th
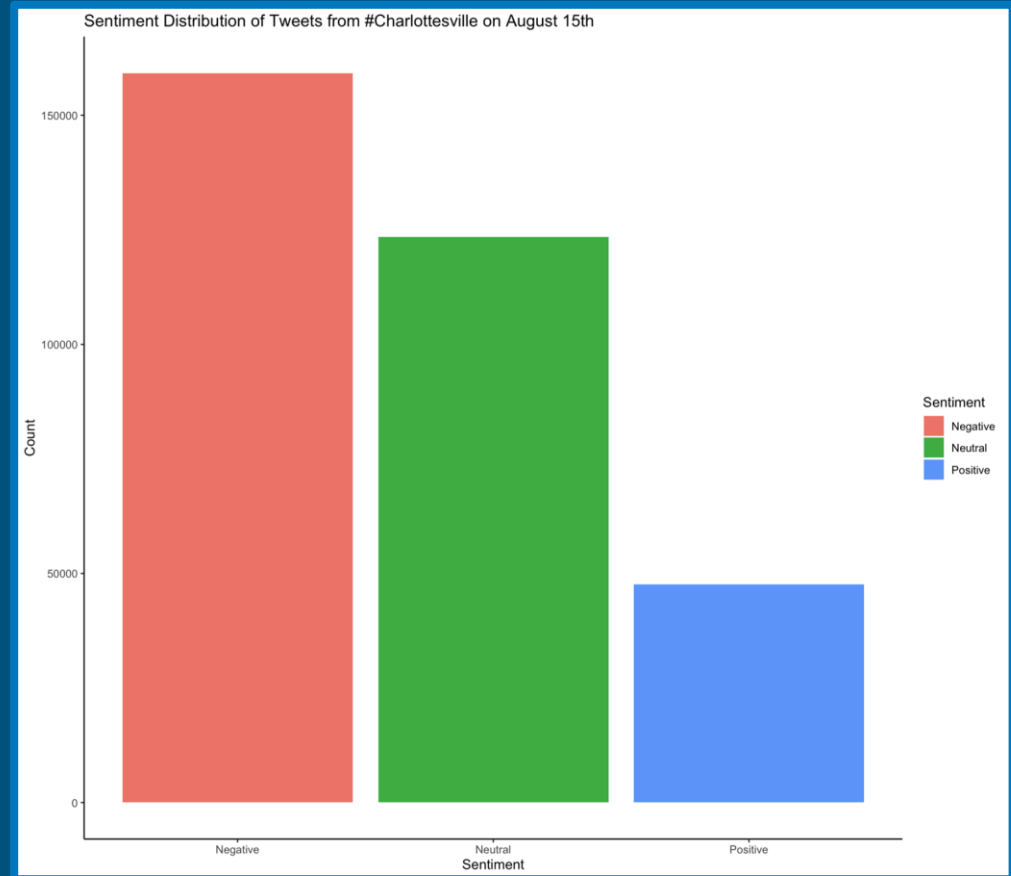
Summary 3: MDS plot of Tweets (Charlottesville, August 2015)

Based on a random sample of 1000 tweets from August 15th

MDS stress value: 0.74924925

This rather high stress value for the MDS indicates that plot may not able to accurately represent the pairwise distances between all the original points in the 2 dimensions.

Explanation

- The MDS plot showcases a sample of 1000 different tweets from August 2015, and how they relate in based on term frequencies and inverse document frequencies, with the MDS plot being based on the transpose of the Term-Frequency Inverse Document Frequency Matrix of the 1000 sample documents and their corresponding terms.

- Proximity of the points (Tweet documents) indicate their similarity. Documents in closer proximity to one another share similar term frequencies and inverse document frequencies, while documents far away from each other indicate different term distributions

- Data point labels correspond to the first 4 digits of the twitter user ID's associated with each tweet posted

Insight

Based on the plot, it appears that a decent portion of the 1000 sample Tweets (Documents) surrounding Charlottesville, August 2015 share similar text content, with cluster 3 containing majority of the Tweet documents, and the tweet documents within cluster 3 being in significantly close proximity to each other, indicating similar text distributions

Numeric Summary (Clusters)

- Cluster 1: 4 Tweets Documents

- Cluster 2: 5 Tweets Documents

- Cluster 3: 982 Tweets Documents

- Cluster 4: 5 Tweets Documents

- Cluster 5: 4 Tweets Documents

MDS Plot of Charlottesville Tweets (Aug 2015)
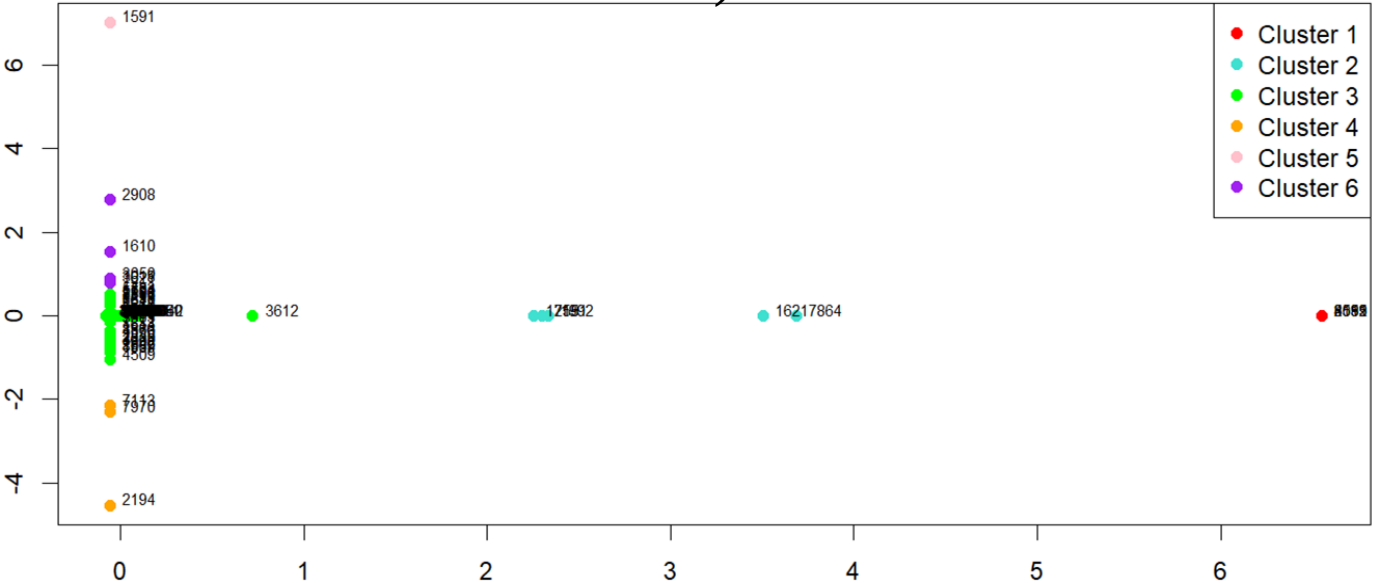
Qualitative Thoughts

Q3: Qualitative thoughts

- Looking at the resulting data visualizations, we gained a clear understanding of public sentiment after this incident, which was mostly negative or neutral.
- Limitations included sheer size of the dataset, as it included thousands of tweet for each observed day. It made it a bit more complicated to create clear data visualizations.
- The sentiment analysis mostly validated expectations that we had for the tweets. We expected the majority of tweets to be negative.

Conclusion

Conclusion

- Future work for this project could include looking at similar events that have taken place in recent years in the United States. We could create new datasets that include tweets written in the following days of those events, and see if they follow the same trend as the #charlottesville tweets. Questions to answer for this further research would be if the public sentiment would be the same for other major events. Are there topics to look that are centered on other issues, like climate change or MeToo? Did events in these categories amass the same amount of public attention?

Thank You for Your Time!