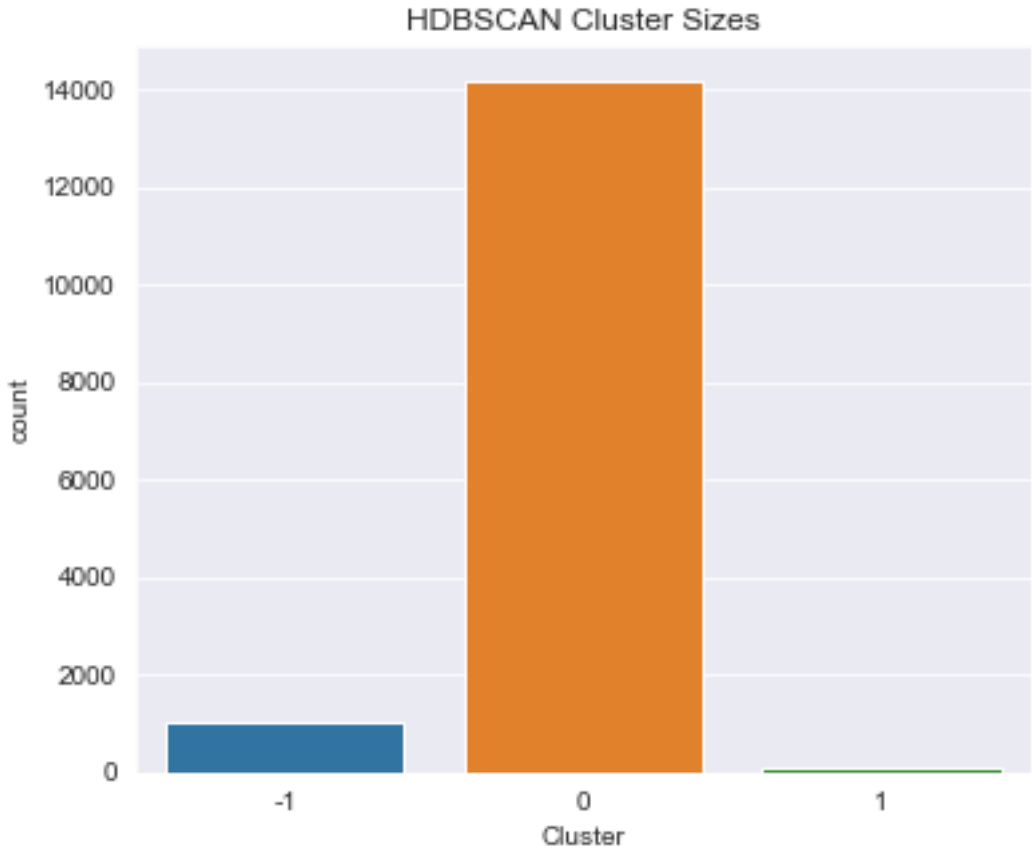
Analysis of Clusters - HDBSCAN

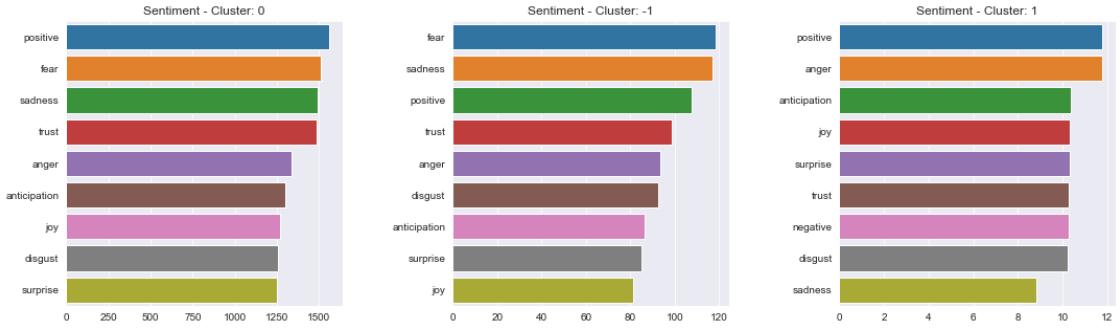
October 20, 2020

1 Analysis of Clusters - HDBSCAN

HDBSCAN was ran with default sklearn parameters and max epsilon 0.1, min cluster size 100. The algorithm used the cosine dissimilarity between document vectors as distance metric. Three clusters were produced, one relatively large, one medium and one small. The medium and small did not seem to grow at the same rate as the large when further data was added, indicating that they likely picked up some characteristics at a point in the scrape which did not continue as more datapoints were scraped. It is unlikely these clusters will be insightful or useful, however we run them through the analysis code for completeness.

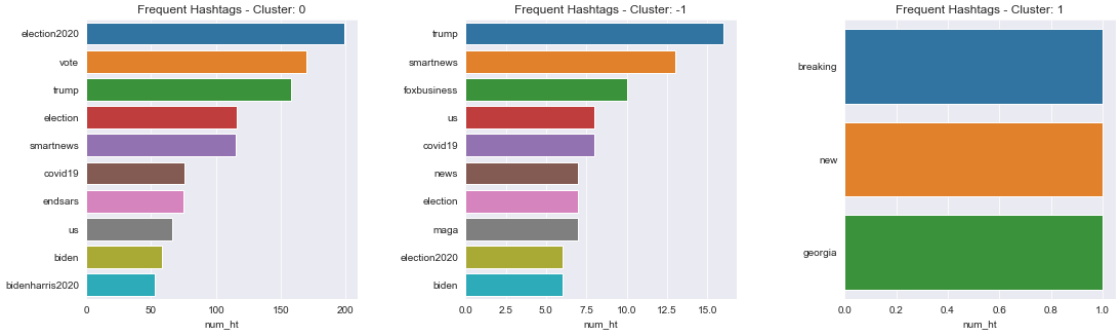


2 Cluster Sentiment



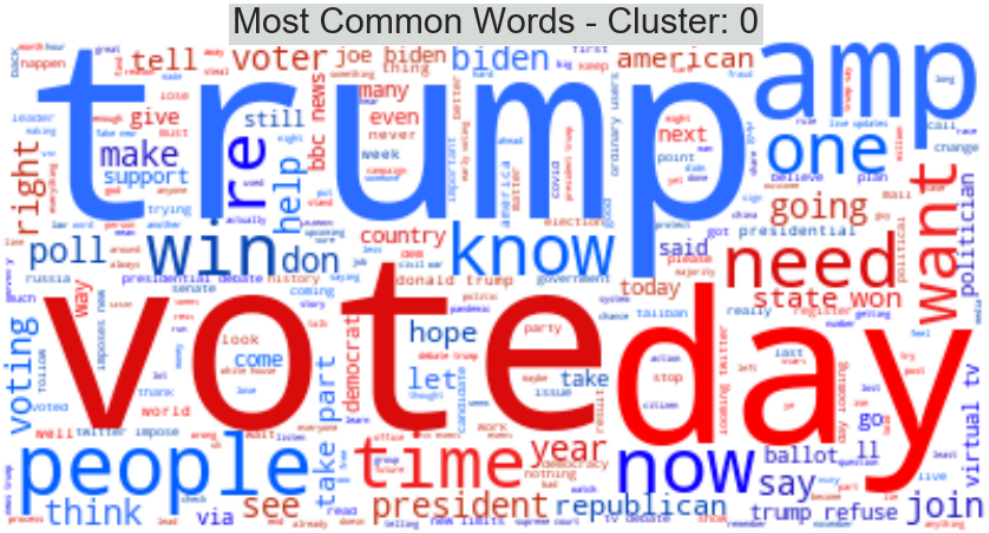
It is interesting to see the more negative emotions slightly further up the hierarchy for cluster -1, but with significant difference in sample sizes, it is not likely we can conclude a significant difference in sentiments from this analysis alone.

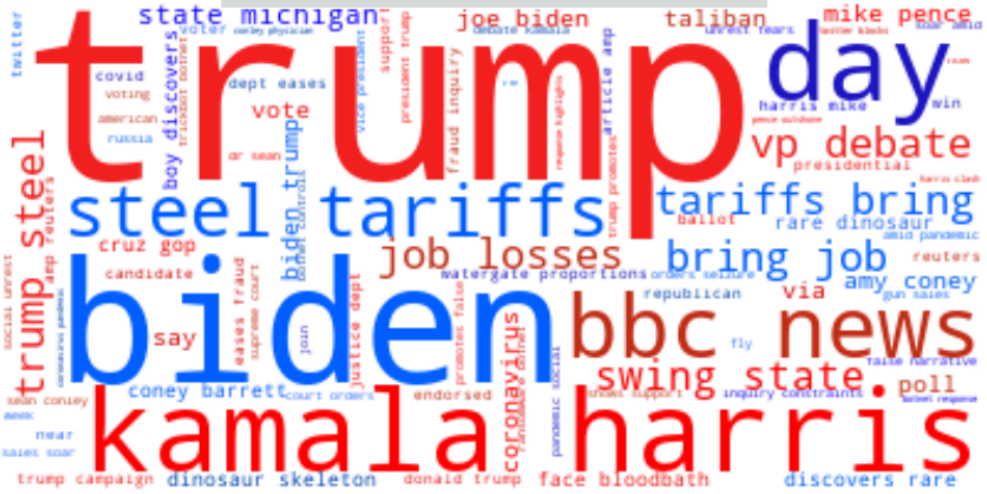
3 Cluster Hashtags



Again, very difficult to draw conclusions from this. Similar hashtags across the 2 main clusters. 'Maga' and 'foxbusiness' being featured in Cluster: -1 could indicate a lean towards pro-Trump. The absense of 'vote' could also indicate this (while 'vote' is fairly arbitrary, it's typically associated with the pro-biden camp). The low volumes in Cluster:1 are to be expected with the low volume of tweets.

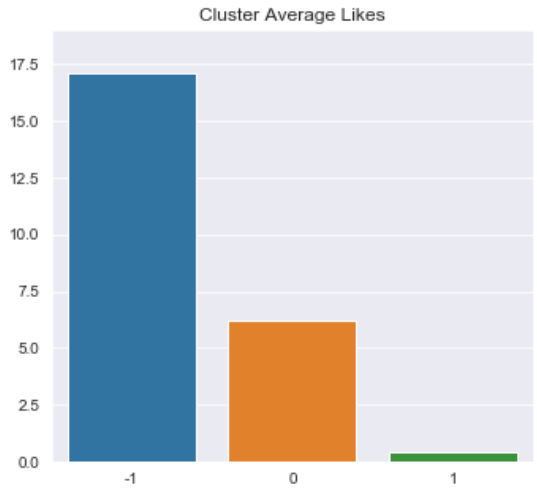
4 Cluster Wordclouds

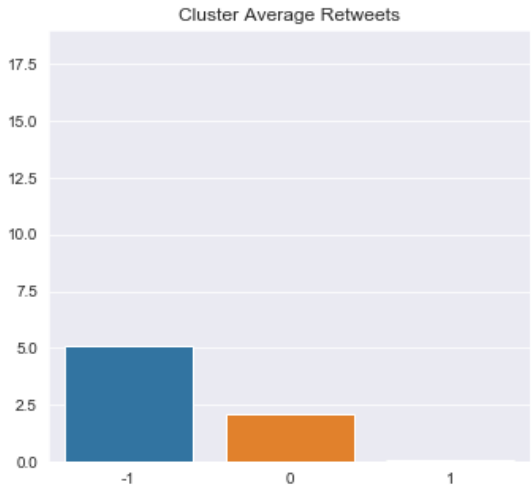




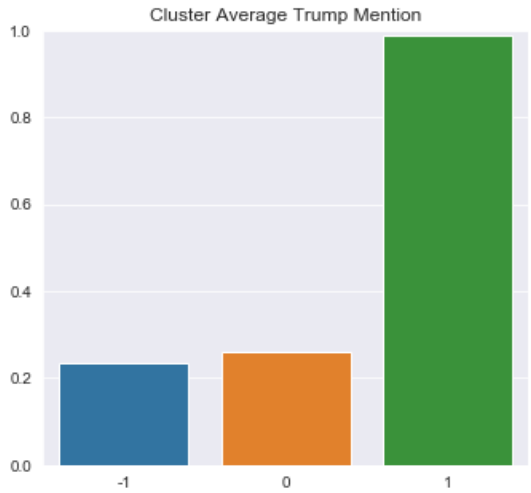
When looking at the words and phrases, we can see a bit of pro-biden and pro-trump in both of the main clusters. References to steel tarrifs seem more common in Cluster:-1 as do references to Biden and Kamala Harris (although this is difficult to definitely conclude with the differing cluster sizes). 'Trump' continues to dominate both, which isn't really a surprise.

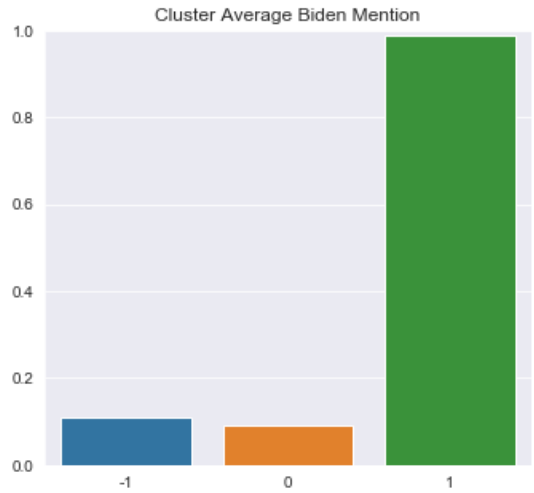
5 Cluster Likes and Retweets





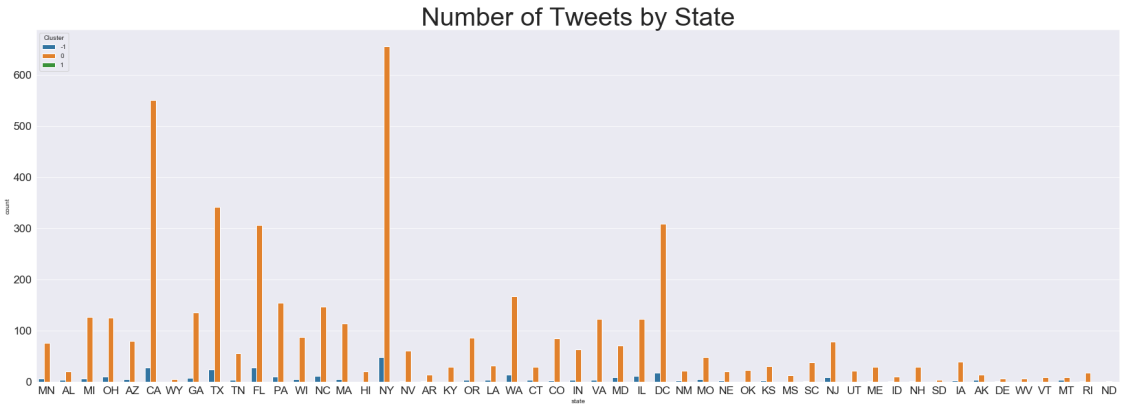
6 Cluster Candidate Mentions





Broadly similar split across -1 and 0. Almost all (if not all) in cluster 1 reference the candidates.

7 Cluster Locations



8 Overall Conclusions

There were noticeable differences but not enough to conclude anything with such a difference in cluster sizes. HDBSCAN clustering did not seem to be very effective on the data.