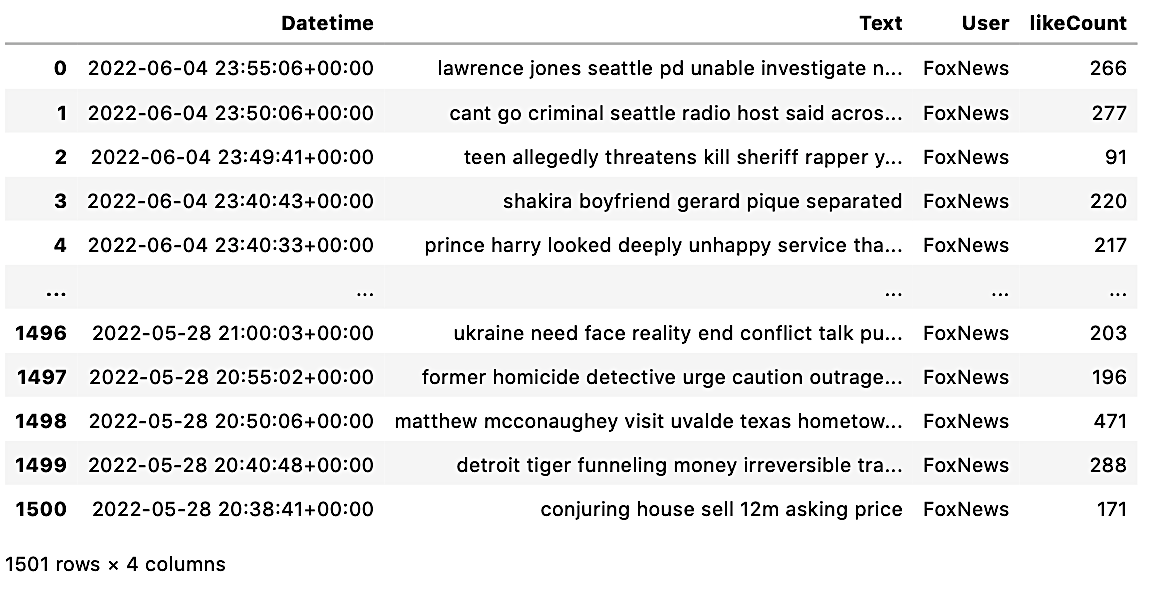
|  |  |
| --- | --- |
| FINAL PROJECT  June 10, 2022 | Abstract  Tweets from three major news outlets were scraped and formatted as a Pandas DataFrame, to determine how their associated political affinities are influencing their news reporting via Twitter.  Katherine Hurtado-da Silva  IST 652 |

Introduction

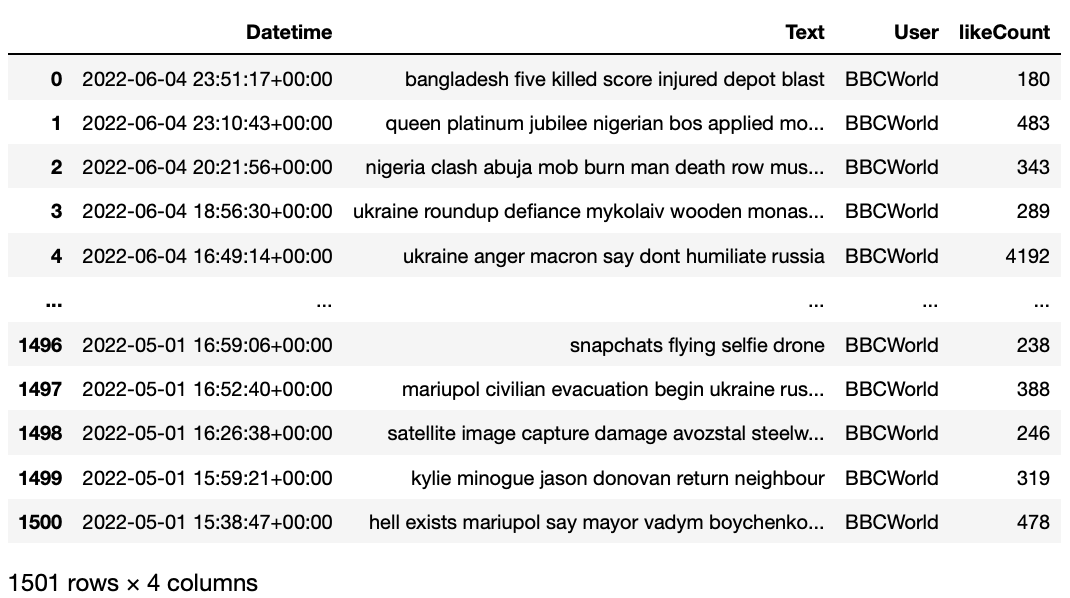
Due to noticeable political affiliations, current news outlets’ credibility has been a contentious topic. AllSides, an online platform that aims to provide “balanced news”, focuses on providing news coverage by several media outlets to enable a reader’s ability to view the differences in how the topic was covered, identify biases by the reporting outlet, and form their own view on the subject. Furthermore, AllSides uses their patented media bias detection and display technology to provide a summative chart of on-line news outlets and the political inclinations in their news reporting. In efforts to determine how political affinities are influencing news reporting, a far-right, centrist, and far-left news outlet was selected from the AllSides Media Bias Chart for further review.

Model: About the Data

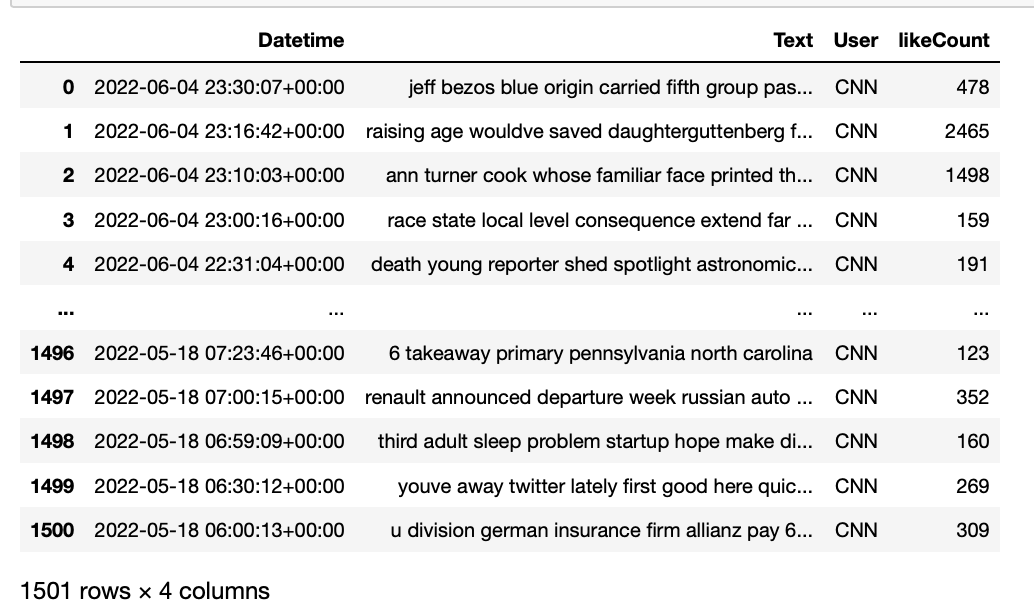
According to AllSides Media Bias Chart, Fox News, BBC, and CNN are considered far-right, centrist, and far-left news outlets, respectively. Therefore, their twitter accounts were selected to compare each outlet’s news topic coverage, sentiment patterns, and audience response through “like counts”. The Snscrape library was used to pull each account’s last 1,500 Tweets posted and its respective number of likes, between 2022-04-01 and 2022-06-05. Data pre-processing steps included changing the tweets to lower-case letters and removing punctuation, hashtags, mentions, special characters, and stop words. Lemmatization was used over stemming because it considers the full context to find the intended root word. Since the meaning of the words in all of the Tweets were important to the sentiment analysis conducted, lemmatization was used to prevent the production of non-meaningful words that often result from stemming. Lastly, WhitespaceTokenizer() was used to discard and separate words by whitespace. None of the pre-processing steps affected the size of the data set. The final data frames resulting from all pre-processing steps can be found in Figure 1, Figure 2, and Figure 3.

**Figure 1. Pre-processed Fox News Tweets Data Frame** 

**Figure 2. Pre-processed BBC News Tweets Data Frame**



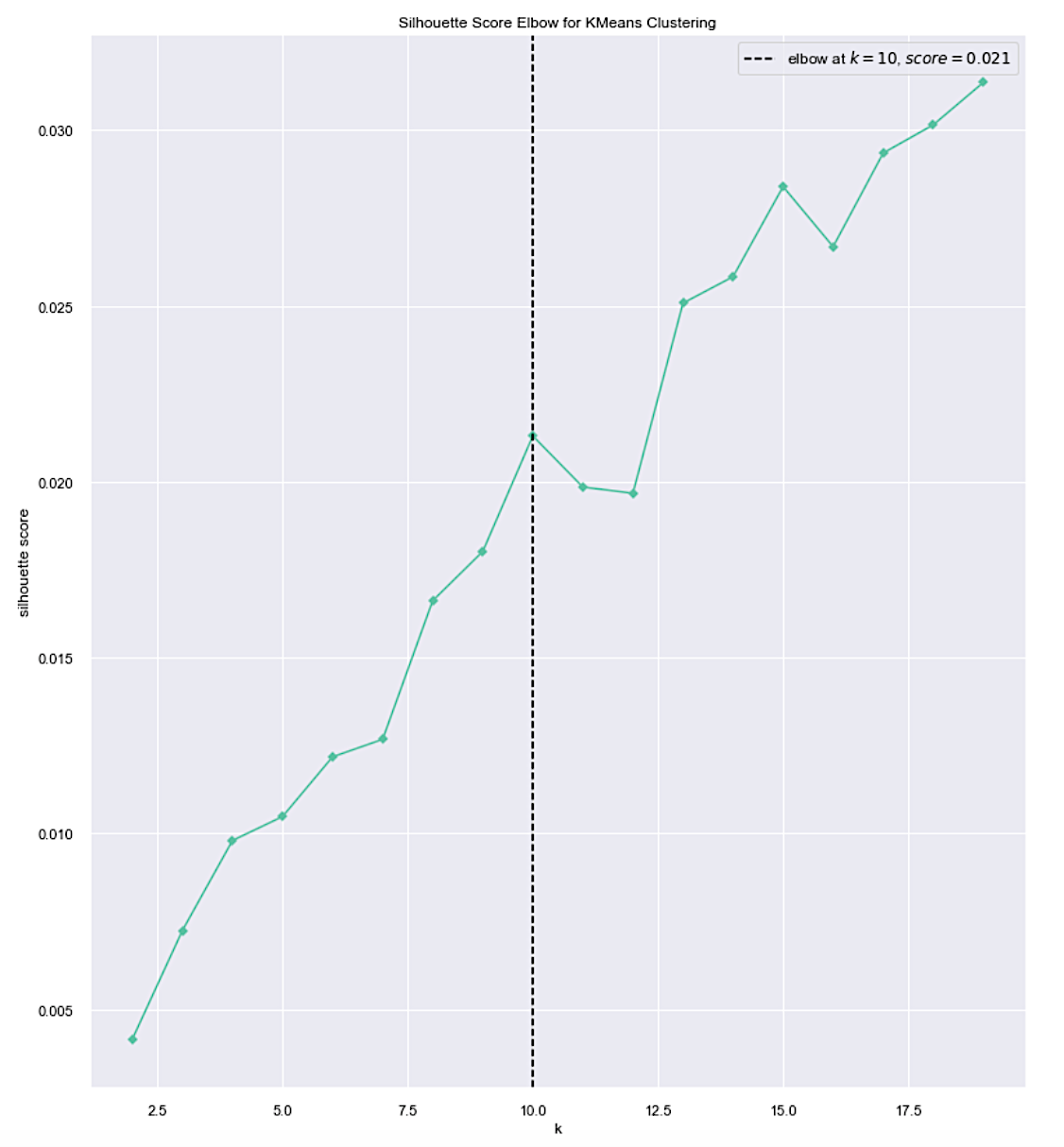
**Figure 3. Pre-processed CNN News Tweets Data Frame**



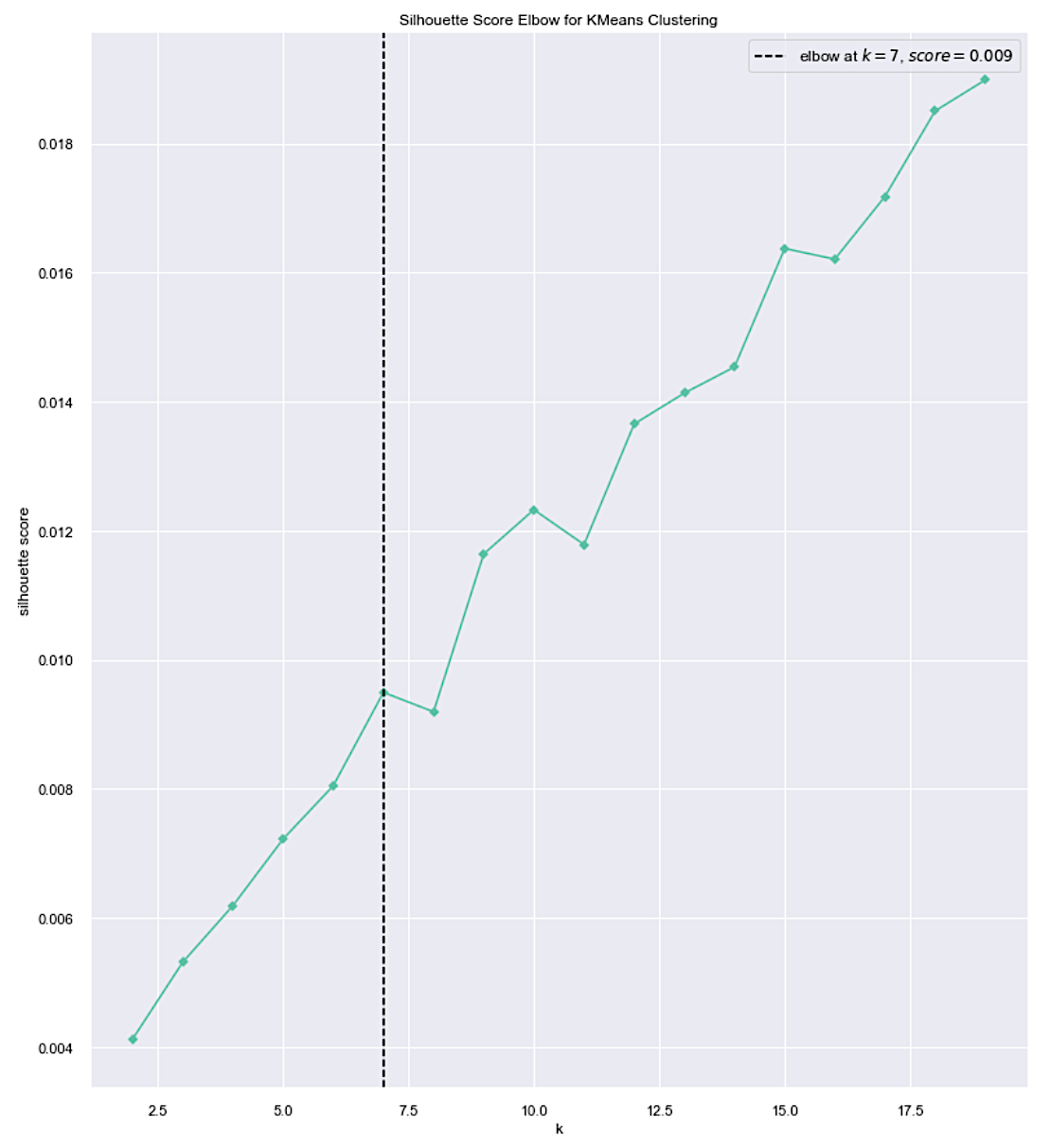
Model: Program Overview

K-means Clustering was used to identify each news outlet's topic coverage, where Elbow & Silhouette plots were generated to determine the optimal number of clusters to use. According to Scikit Learn, a silhouette coefficient is calculated using each sample’s mean intra-cluster distance (a) and the mean nearest-cluster distance that the sample is not a part of (b), where the silhouette coefficient for a sample is (b - a)/max(a, b). The Elbow & Silhouette plots were coded to return the mean silhouette coefficient for cluster sizes (K) ranging from two through twenty, where the “elbow” (visually represented by a dashed line), specified the optimal number of K for the given model. The Tweet portions of each news outlet were also changed to Unicode to ensure it was represented in its universal form and to reduce the chances of data corruption when it was transformed into feature vectors with their associated text weights. The optimal number of K used for each news outlet’s K-means clustering model, can be referenced in Figure 4, Figure 5, and Figure 6.

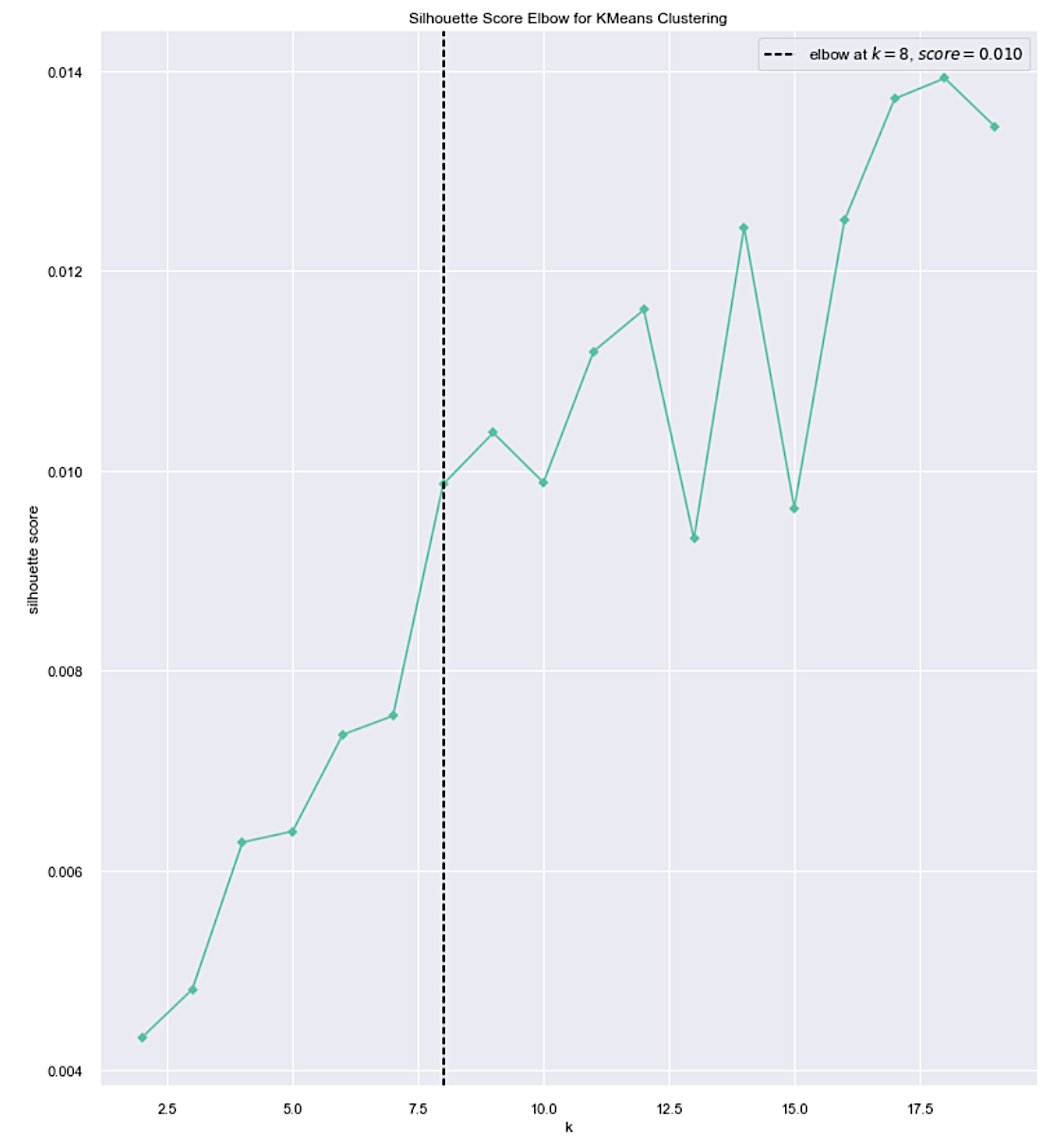
**Figure 4. Fox News Tweets Elbow & Silhouette Plot**

****

**Figure 5. BBC News Tweets Elbow & Silhouette Plot**

****

**Figure 6. CNN News Tweets Elbow & Silhouette Plot**



Sentiment Analysis via TextBlob library was used to find the polarity and subjectivity scores of Tweets for each news outlet as well. Bar plots were generated to view the distribution of the positive, negative, and neutral Tweets by news outlet and compared to determine trends. To generate the bar plots, a range for the scores were outlined and attributed to a label. The bullets below state the ranges that determined whether a Tweet was considered negative, neutral, positive, objective, subjective, or highly subjective.

* if sentiment score >= 0.05 : sentiment label was 'Positive'
* if sentiment score <= -0.05: sentiment label was 'Negative'
* if -0.05 <sentiment score<0.05: sentiment label was 'Neutral'
* if subjectivity score >= 0.6 : subjectivity label was 'Highly Subjective'
* if subjectivity score <= 0.4 : subjectivity label was 'Objective'
* if 0.4 < subjectivity score <0.6: subjectivity label was 'Subjective'

Lastly, box plots of Tweet likes, grouped by sentiment and subjectivity were generated. These box plots were used to determine if Tweets by an identified sentiment or subjectivity level corresponded to higher or lower likes, therefore providing insight on what the audience of the news outlets prefer.

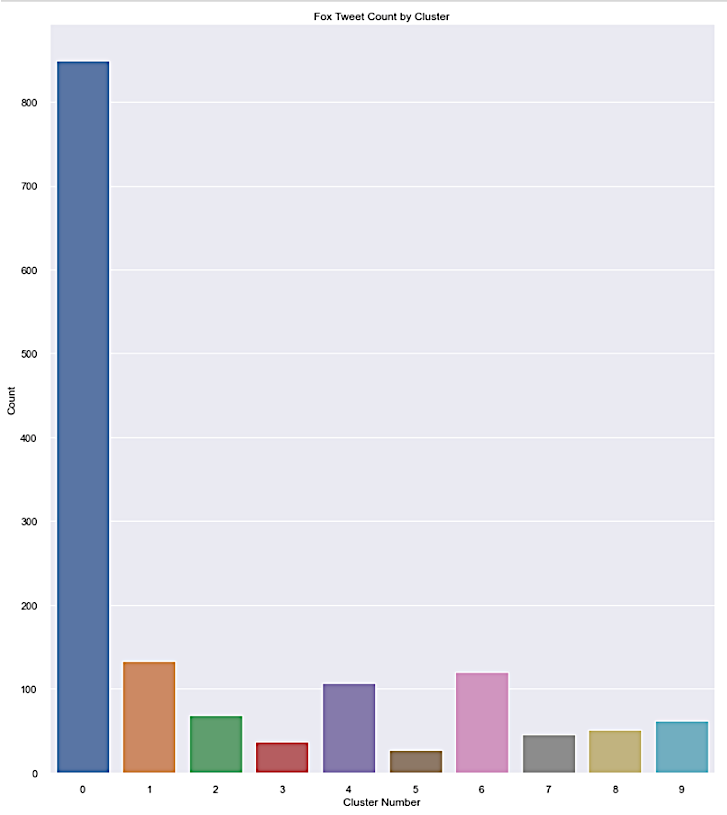
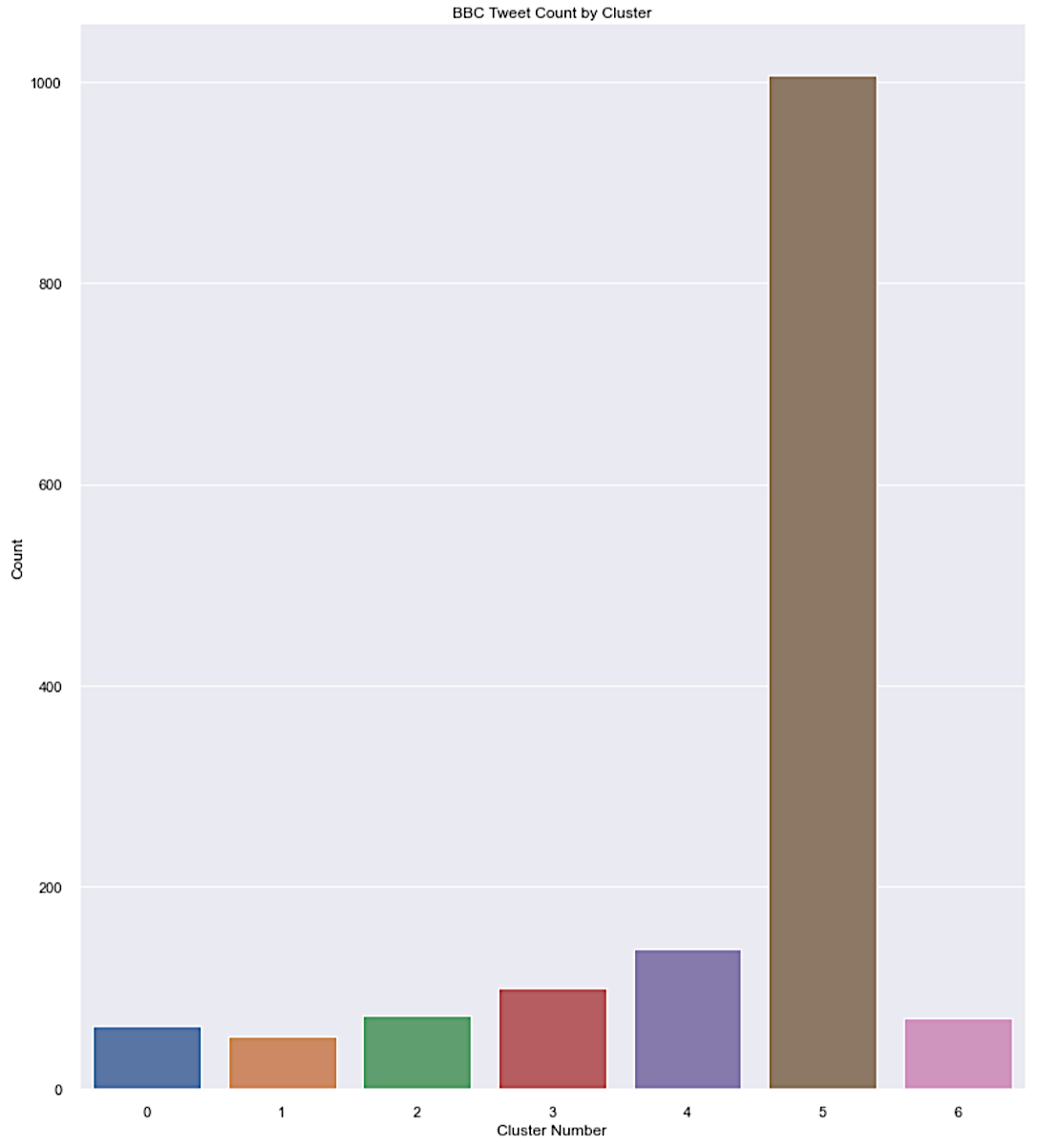
Analysis: Question 1- How did the reported topics for each news outlet differ?

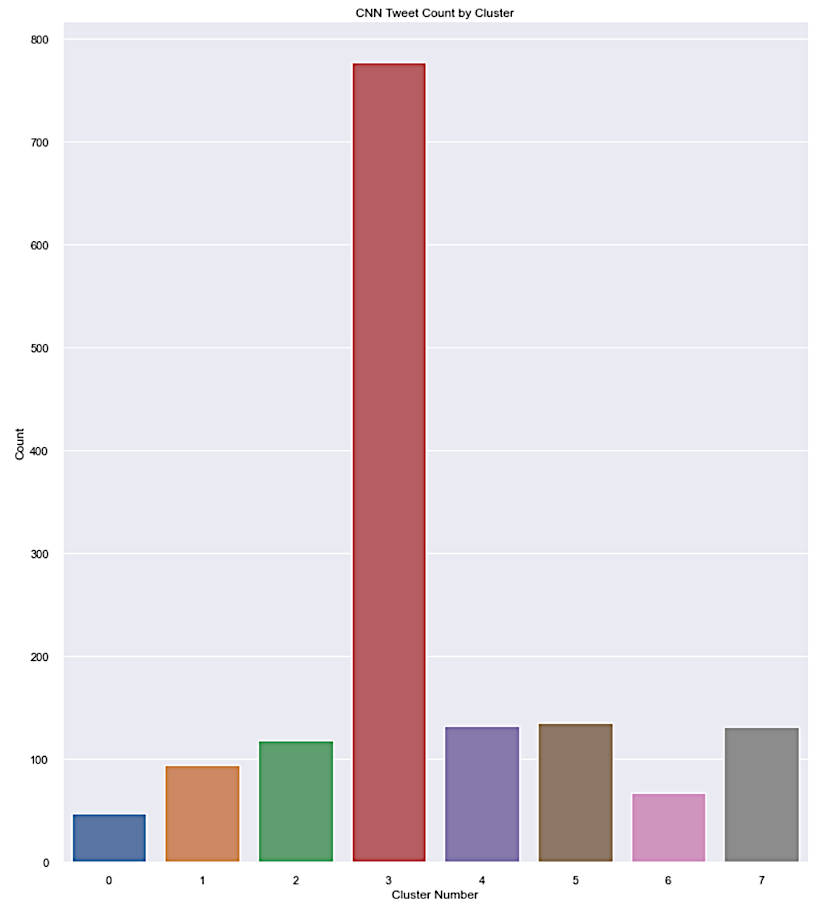
In spite of using the optimal number of clusters, as indicated by the Elbow & Silhouette plots, the resulting clusters did not seem optimal upon further assessment. This was not so much of a surpise since the Sihlouette scores associated with the optimal number of clusters were positive, but close to zero, indicating low fidelity. Figure 7 illustrates the distribution of the tweets in each cluster, further indicating the size of the largest cluster for each news outlet is most likely attributed to topics that would ideally be separated. Figure 8 shows the word clouds for the largest cluster, along with its the centroid features, to help decipher the differences among all three news outlets. Fox’s largest cluster listed the terms “Biden”, “guns”, “Russia”, “house”, “trial”. While reviewing the CSV output of the actual lemmatized Tweets in this cluster, the context of these terms were clearer. The Tweets including the word “gun” was used in reference to gun regulations, “Biden” was mentioned in association to current standings of American infrastructure or gun regulations, and while the “Russia” was mentioned because of the current Ukrainian war, the majority of the frequency was due to the Russia probe, which also explains why the word “trial” was listed as a centroid feature (Durham trial against Sussman). “House” was in reference to house hearings related to gun regulations, or the White House. All in all, the main focus of this cluster was around gun regulation and Russia (with regards to the current war and Trump related probe collectively). “Woman” did not have a singular associated topic, it was used multiple times to reference an individual in unrelated stories ranging from sports, accidents, and general news topics.

BBC’s largest cluster was heavily focused on the Australian election. “Texas” and “shooting” go hand in hand, with the Uvalde school shooting being a major topic covered. “Covid”, “North”, and “Korea”, were all related in the frequent coverage of a Covid breakout in North Korea, where the first reported death to the recent outbreak was considered grounds for concern. Lastly, the Depp defamation trial was also covered in this cluster, but did not make further mentions of Amber Heard or any details beyond the trial decision. Within this large cluster, the main topics reported appear to be on the Australian election, Uvalde school shooting, North Korean Covid outbreak, and the Depp defamation trial.

CNN’s use of “watch”, similar to Fox’s use of “woman”, is not related to any unique topic and used generously through unrelated topics. “Year” was similar to “watch” when it was not being used to report of Queen Elizabeth’s 70-year reign. “Baby” and “formula” also go hand in hand with the baby food shortage being a freqeunt reported issue. “Ukrainian” and “Russian” were mentioned in reference to the current war’s standing. Though not normally included within the same Tweets, each word was used only in the context of the war. Upon further review of the CSV output for this cluster, the use of the words “CNN”, “report”, and its connection to the word “watch”, became clearer. The Tweeting style of CNN includes references to themselves reporting within the news Tweets, resulting in a high frequency of these words in spite of not being related to the actual news. Lastly, “president”, though used to report stories for President Zelensky, Vladimir Putin and other leaders, had the highest frequency of use associated with Biden and varying events such as his visit to South Korea, efforts towards baby formula shortage, and overall politics. Ultimately, the spread of topics in this cluster focused predominantly on the baby formula shortage, President Biden’s steps to address the issue along with his visit to Asia, and the Ukrainian invasion.

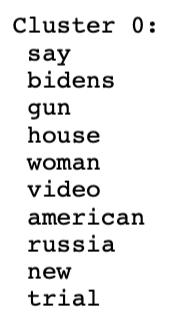
**Figure 7. Tweet Counts in Clusters by News Outlet**

**** ****

****

**Figure 8 Word Cloud & Centroid Features for Largest Cluster by News Outlet**

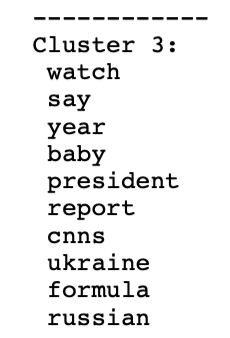
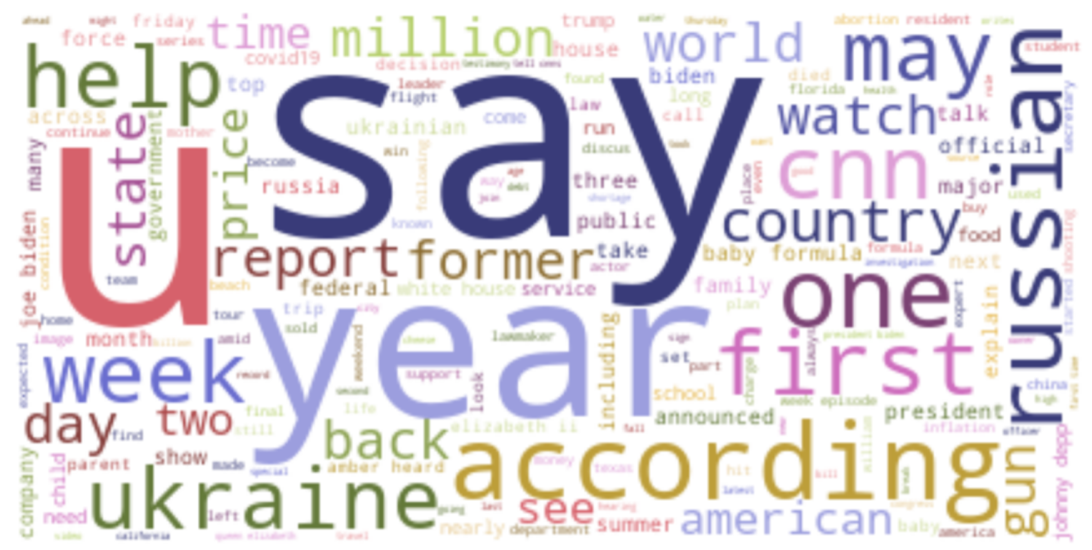
***Fox News Word Cloud & Centroid Features for Largest Cluster***

****** 

***BBC News Word Cloud & Centroid Features for Largest Cluster***

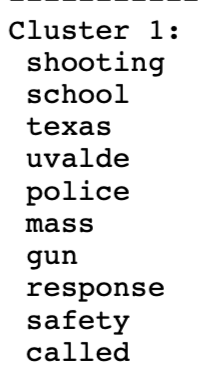


***CNN News Word Cloud & Centroid Features for Largest Cluster***

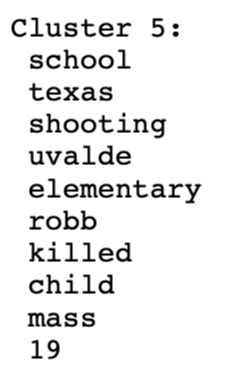
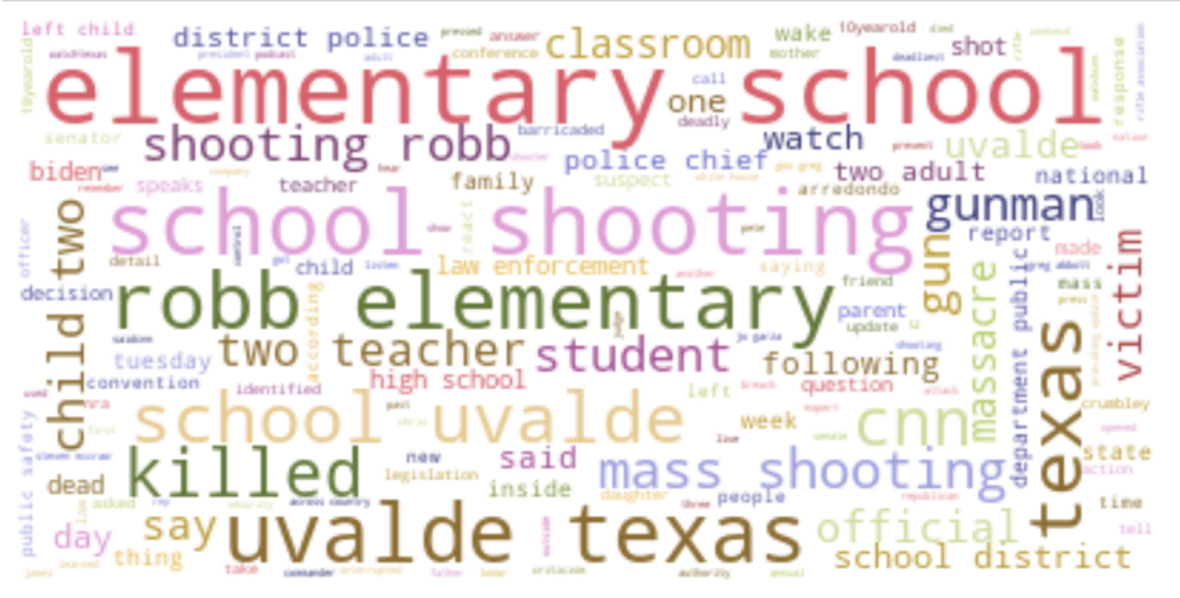


The Uvalde school shooting was a leading story in BBC, but this was not the case with Fox or CNN. With gun regulations taking center stage in Fox’s largest cluster and CNN having neither gun regulations nor the school shooting appearing at all, a closer look was taken to determine how much of a focus this story was for Fox and CNN news. Upon further review, this topic was uniquely clustered for both news outlets. Figure 9 and Figure 10 depict the word cloud and associated centroid features for the Uvalde school shooting.

**Figure 9. Fox News Word Cloud & Centroid Features for Uvalde School Shooting**

**** 

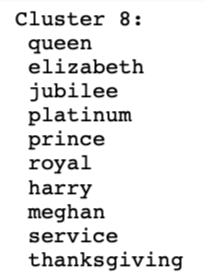
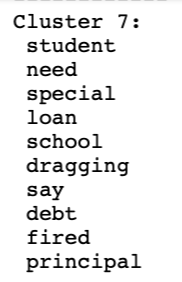
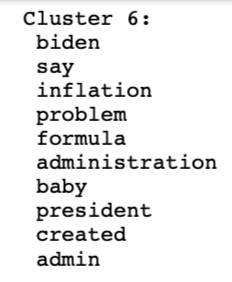
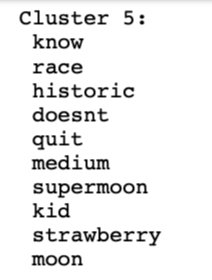
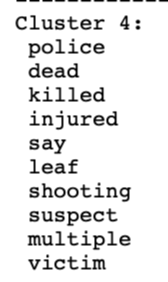
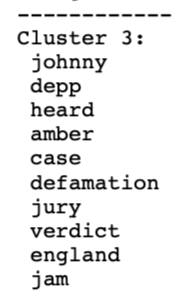
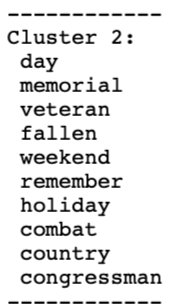
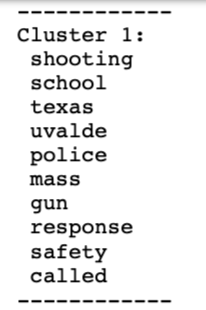
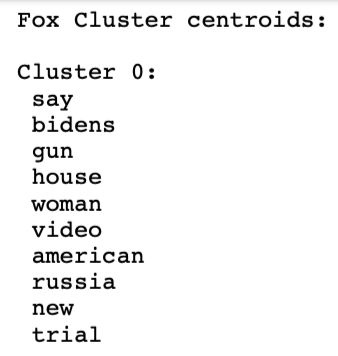
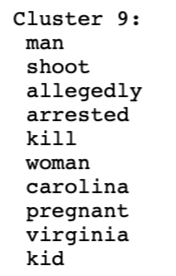
**Figure 10. CNN News Word Cloud & Centroid Features for Uvalde School Shooting**

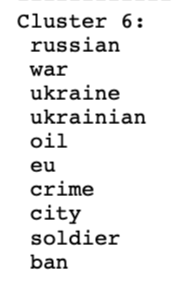
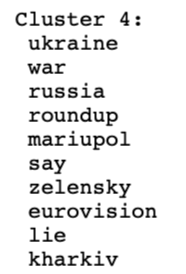
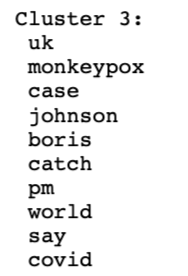
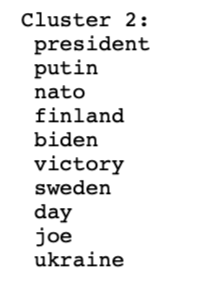
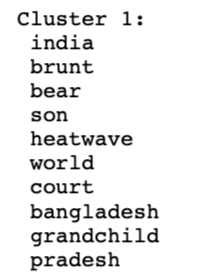
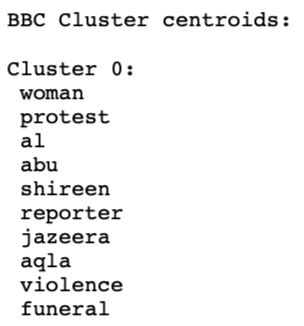


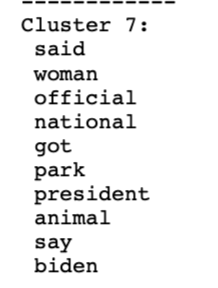
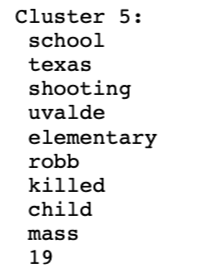
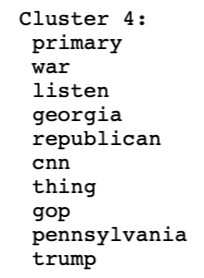
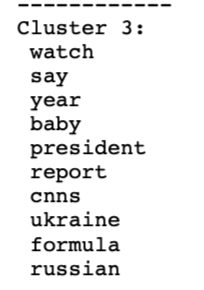
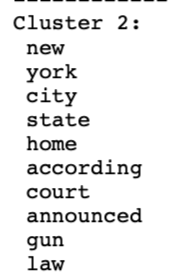
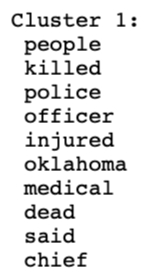
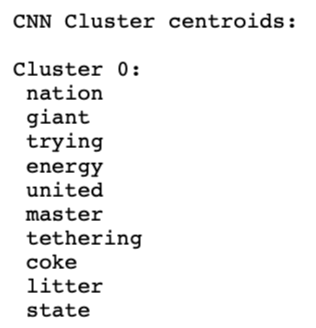
With the assumption all tweets in these clusters are solely about the Uvalde school shooting, it appeared as though Fox and CNN had, at best, 134 Tweets vice 136 of their tweets dedicated to this tragedy. However, when diving into the CSV outputs and manually reviewing the tweets in the cluster, Fox actually had 53 while CNN had 87, bringing their respective data set percentage coverage of the school shooting to 4% and 6%. However, the greatest difference was in the information reported covering the school shooting. Both clusters align with the first four centroid features “school”, “shooting”, “Texas”, and “Uvalde”, but vary with the details pertaining to the aftermath of the event. Fox’s frequent use of “police”, “mass”, “gun”, “response”, “safety”, and “called”, centered around the investigation being conducted for the police response to the shooter. The Tweets’ focus is entirely on charges related to potential gunmen (unrelated to Uvalde), police response to Uvalde shooting, and accusations on the root of mass shootings that range from “Christian Nationalism” to “patriarchy masculinity”. CNN’s Tweets frequently informs details that pertain to the victims of the school shooting. Centroid features “elementary”, “Robb”, “killed”, “child”, and “19”, provided the gut-wrenching context of the school shooting’s victims. It is evident that Fox news’s coverage withheld important details about the Uvalde school shooting, providing a narrative that appears to redirect concern to how the situation was handled by the police, as opposed to who was at the receiving end of the violence. While CNN also reported on the investigation of the police’s handling of the school shooting, the emphasis was on informing and reiterating the victims were children. After re-visiting the Uvalde-related Tweets for BBC news in cluster 5, it was concluded that Fox new was the only outlet that omitted any mention of the victims being children, or that the target was an elementary/primary school.

Figure 11 lists all the of centroid features for each news outlet. Aside from what has already been discussed, the centroid features show Fox has individual clusters dedicated to Memorial Day, the Johnny Depp defamation trial, the Biden administration and associated failures, Queen Elizabeth’s Jubilee, student loans, and overall crime. On the other hand, BBC appears to report on global issues, with clusters dedicated to women under Taliban rule, social and economic issues in India, NATO news as it relates to the Ukrainian Invasion and Finland, disease concerns such as monkey pox and covid, and two clusters dedicated to unique issues surrounding the Ukrainian invasion, one with Mariupol and the other with oil as the focus. Although Depp’s defamation trial is listed, it was not frequent enough to warrant its own cluster, and does not appear to get as detailed as Fox news did. The Queen’s jubilee was also not reported frequently enough to be listed as a centroid feature for any of BBC’s topic clusters. CNN’s clusters differed the most in its coverage of the primaries, New York gun regulations, the Oklahoma shooting, and Trevor Reed’s release from Russia, which did not appear in Fox or BBC news. It is also worth noting that inflation was not emphasized in BBC or CNN news, as it was in Fox news. Furthermore, Monkey Pox and Covid do not appear to be significant in Fox or CNN reporting, as it was for BBC news.

**Figure 11. Centroid Features by News Outlet**

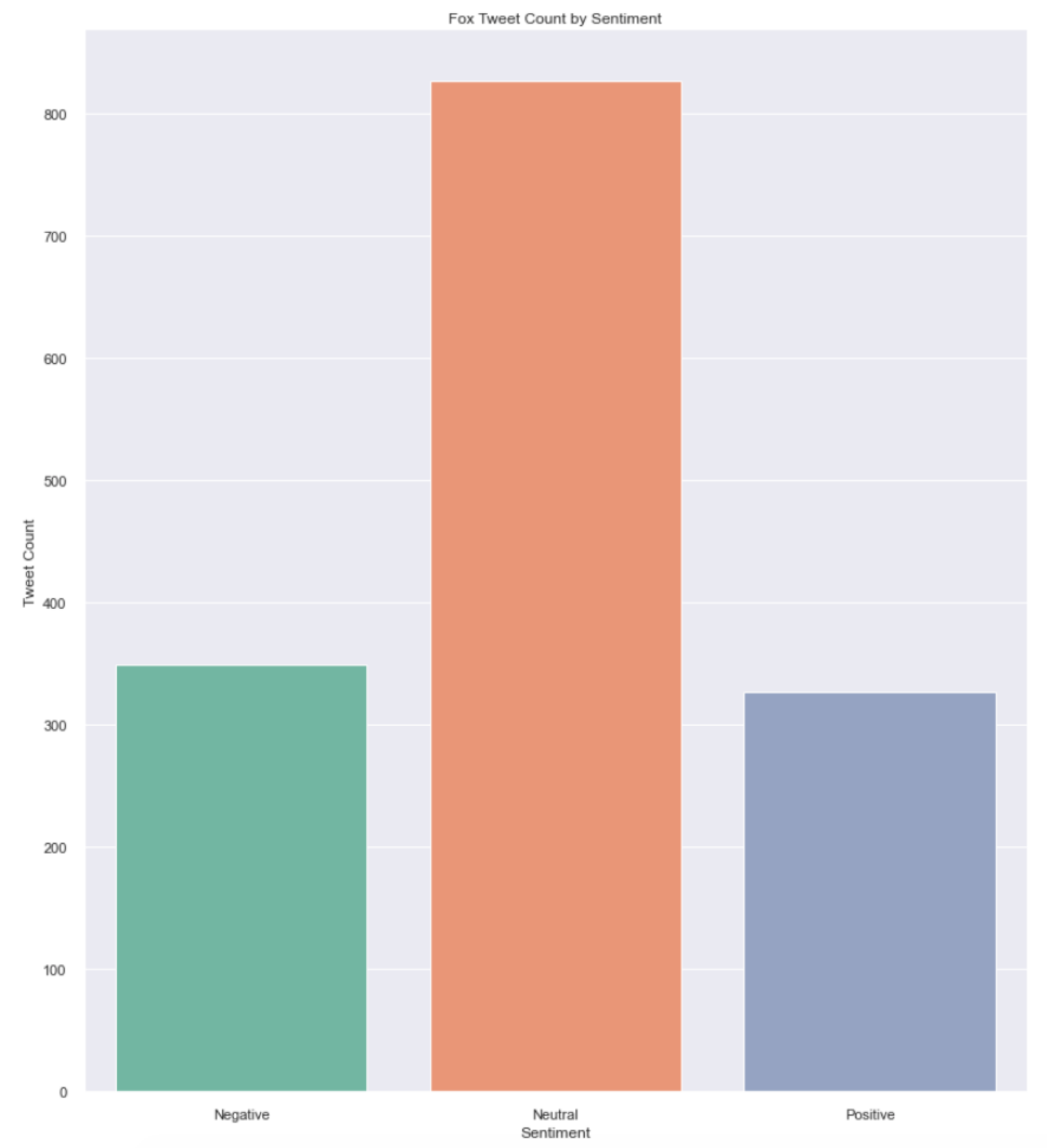
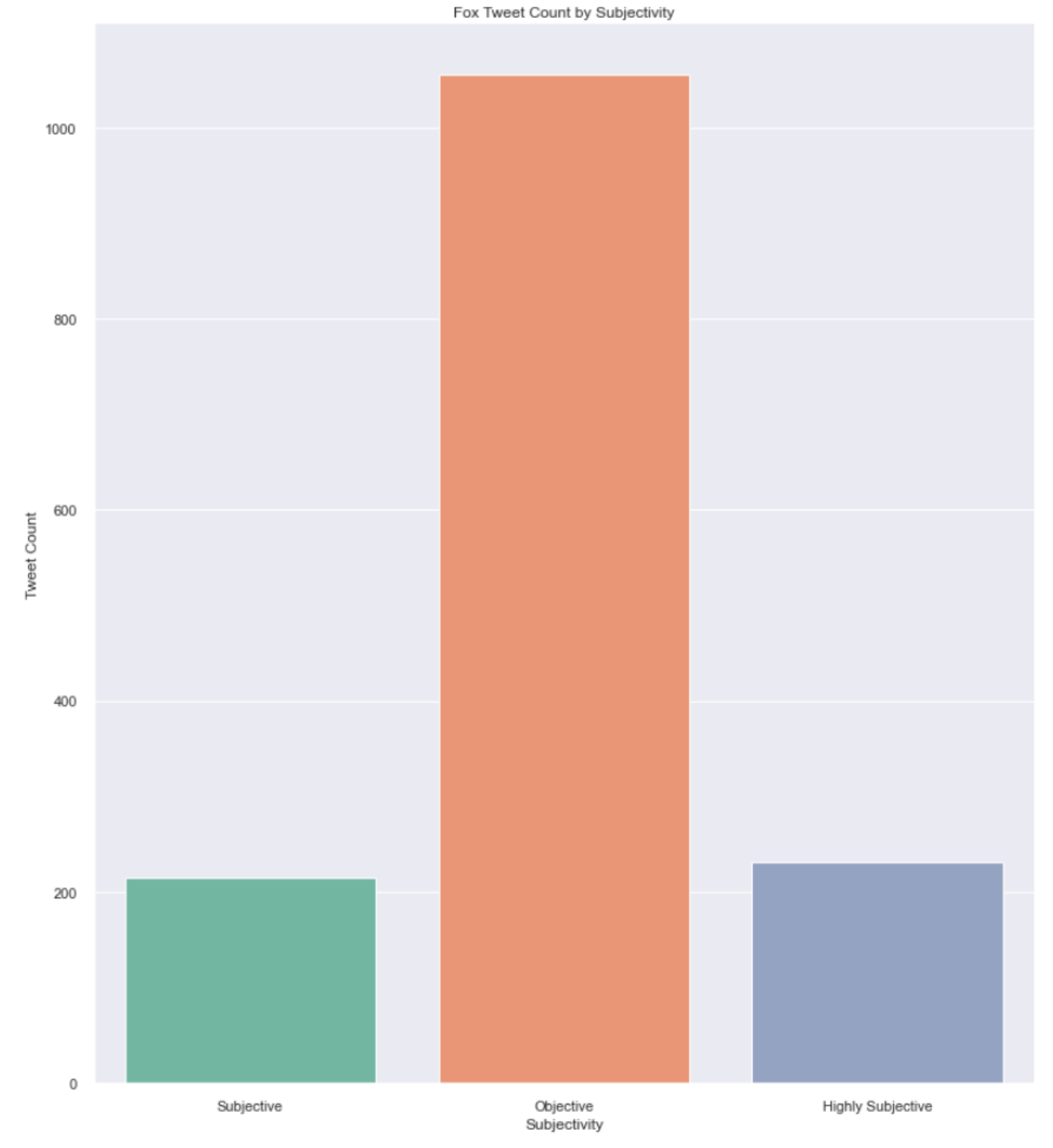




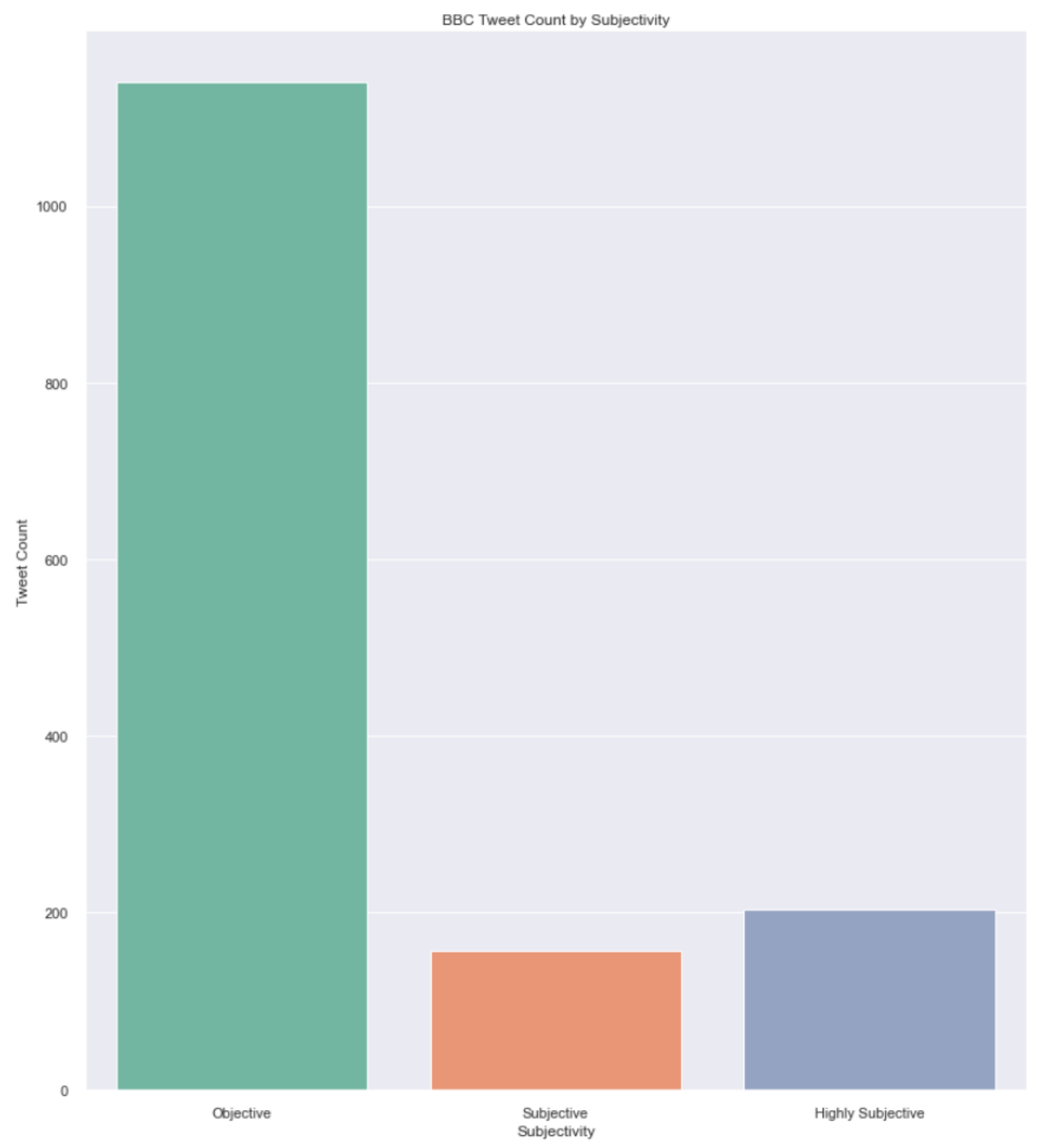
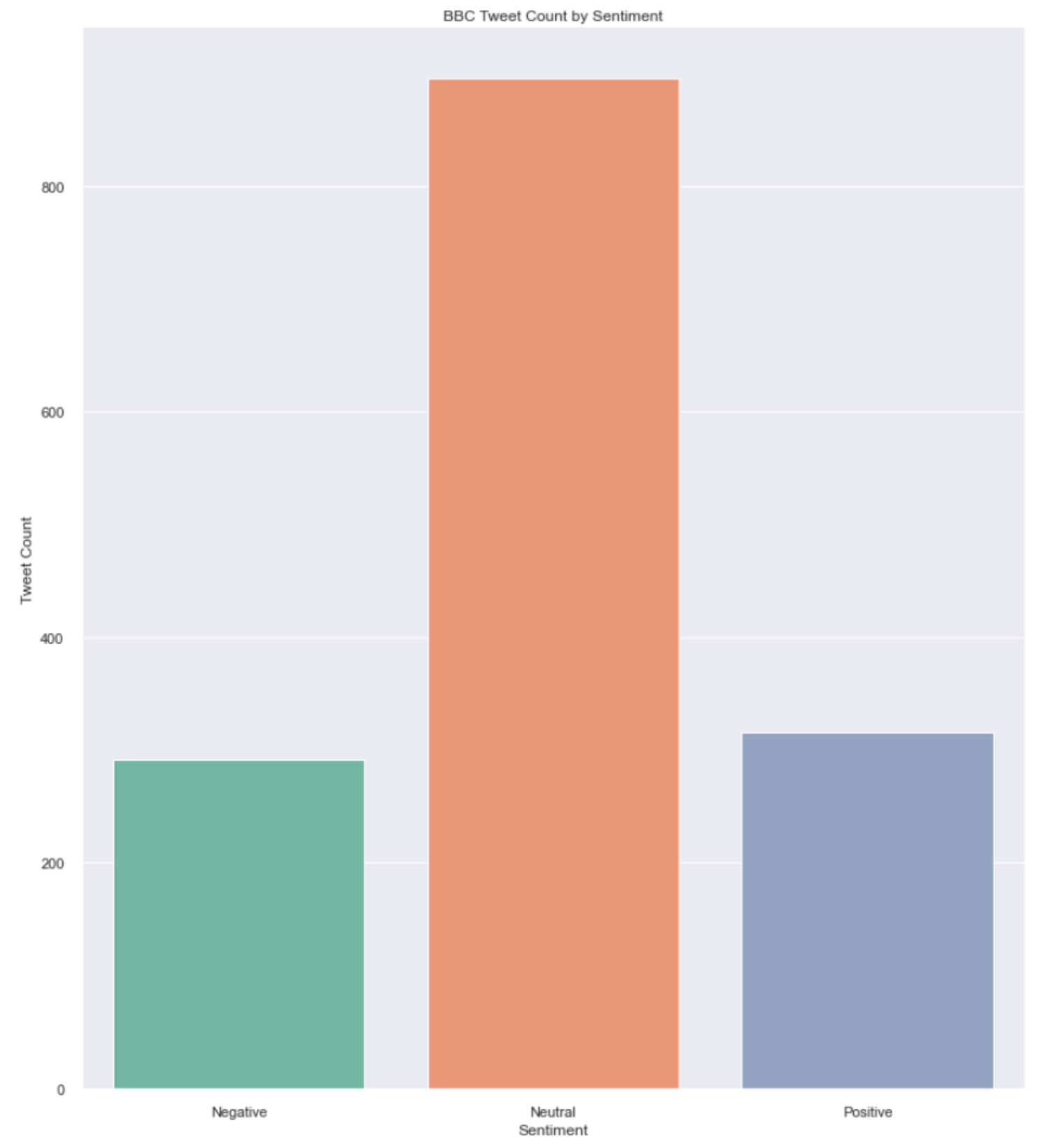
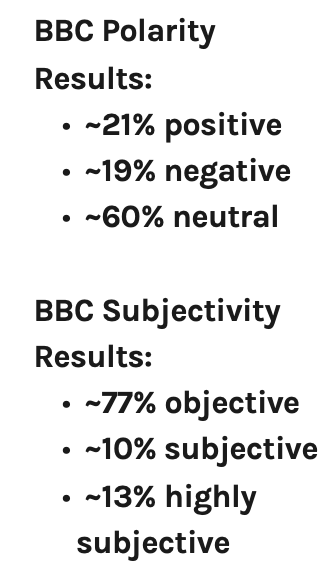
Analysis: Question 2- How did negative, positive, and neutral tweets vary according to each news outlet?

Sentiment results on the data collected for Fox, BBC, and CNN indicate that BBC, the 'centrist' news outlet, has the highest percent of objective and neutral Tweets. The findings also indicate the 'far-left' outlet, CNN, has highest rates of subjective, positive, and negative Tweets. Compared to 'centrist' statistics with BBC, the 'far-right' Fox Tweets do not vary in objectiveness or sentiment significantly. BBC Tweets had the highest percent of Tweets labeled as 'neutral'. Fox Tweets came in second with 55%. Both news outlets had more than twice as many neutral Tweets as CNN. BBC Tweets had the highest percent of Tweets labeled as 'objective. Fox Tweets had 70%, while CNN had 60%. CNN's subjective Tweets, collectively amounts to 75% more than those of BBC and about 30% more than Fox. CNN, which is associated with a 'far-left' political affinity, appears to report the most subjective and non-neutral news via Twitter. BBC, as expected, is the most neutral and objective news source. Figure 12, 13, and 14 illustrate the distribution and percentage of Tweets by sentiment and subjectivity.

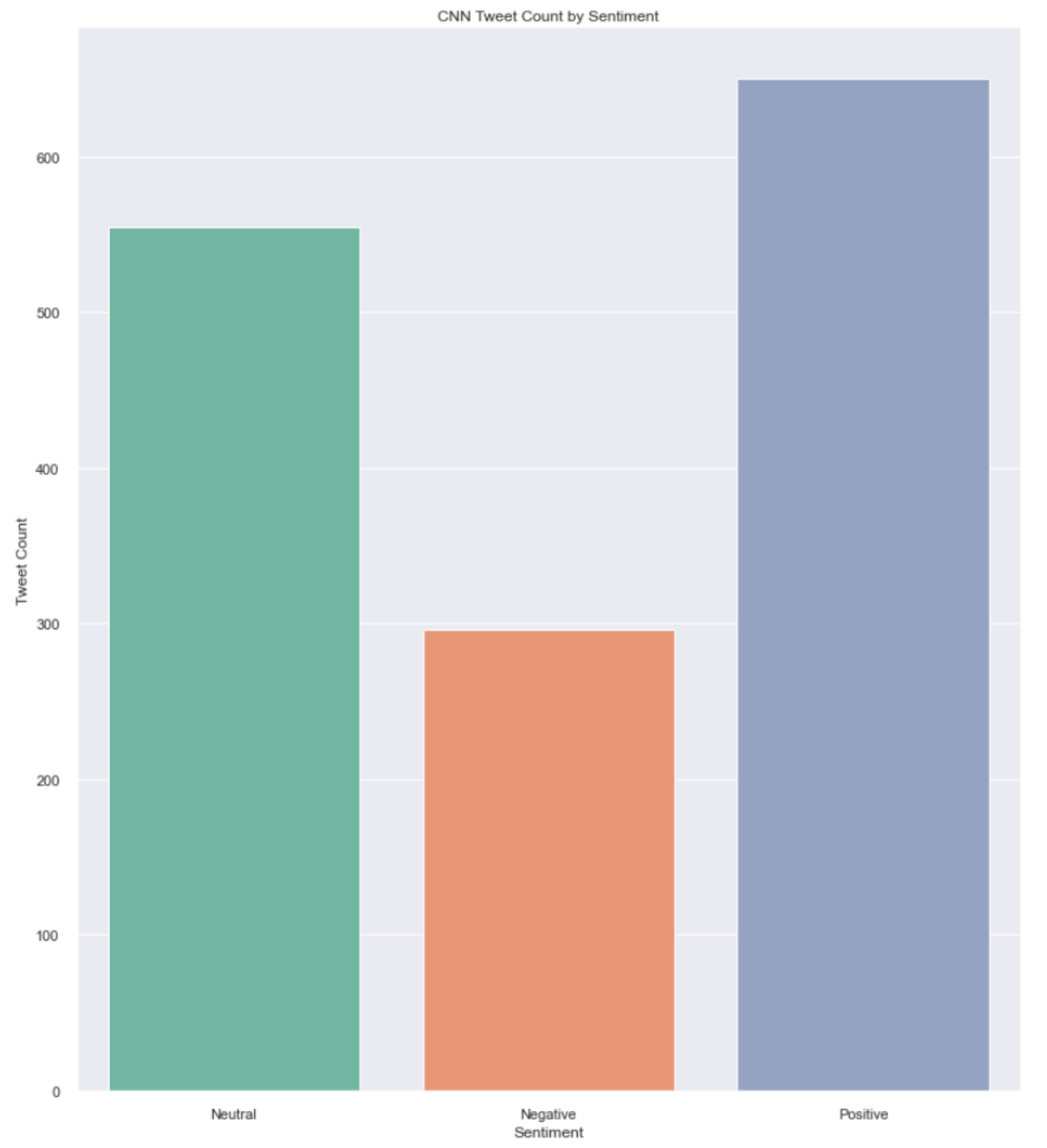
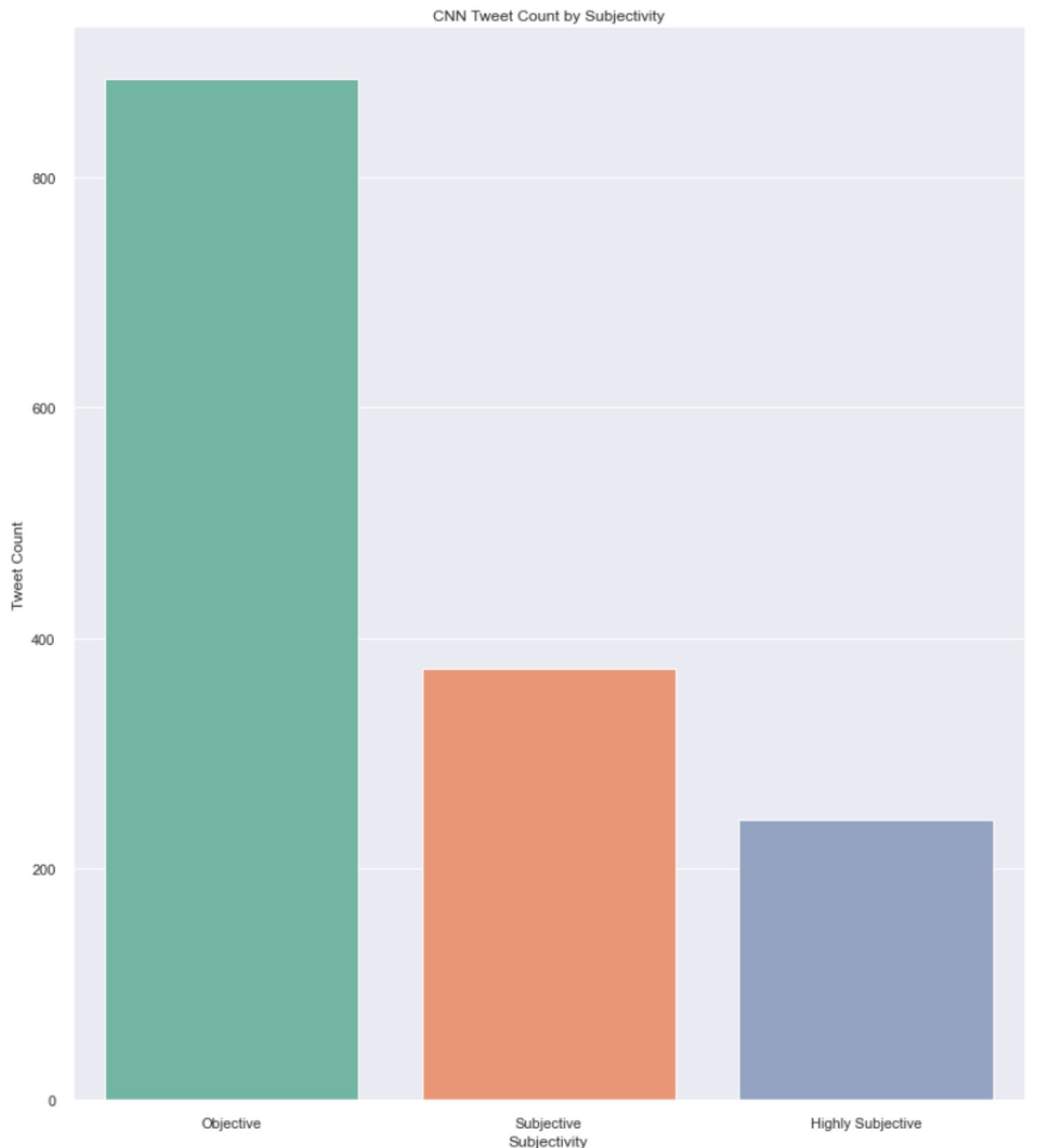
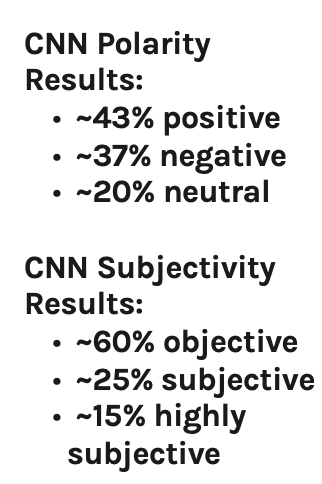
**Figure 12. Fox News Tweets Sentiment & Subjectivity**

**Figure 13. BBC News Tweets Sentiment & Subjectivity**

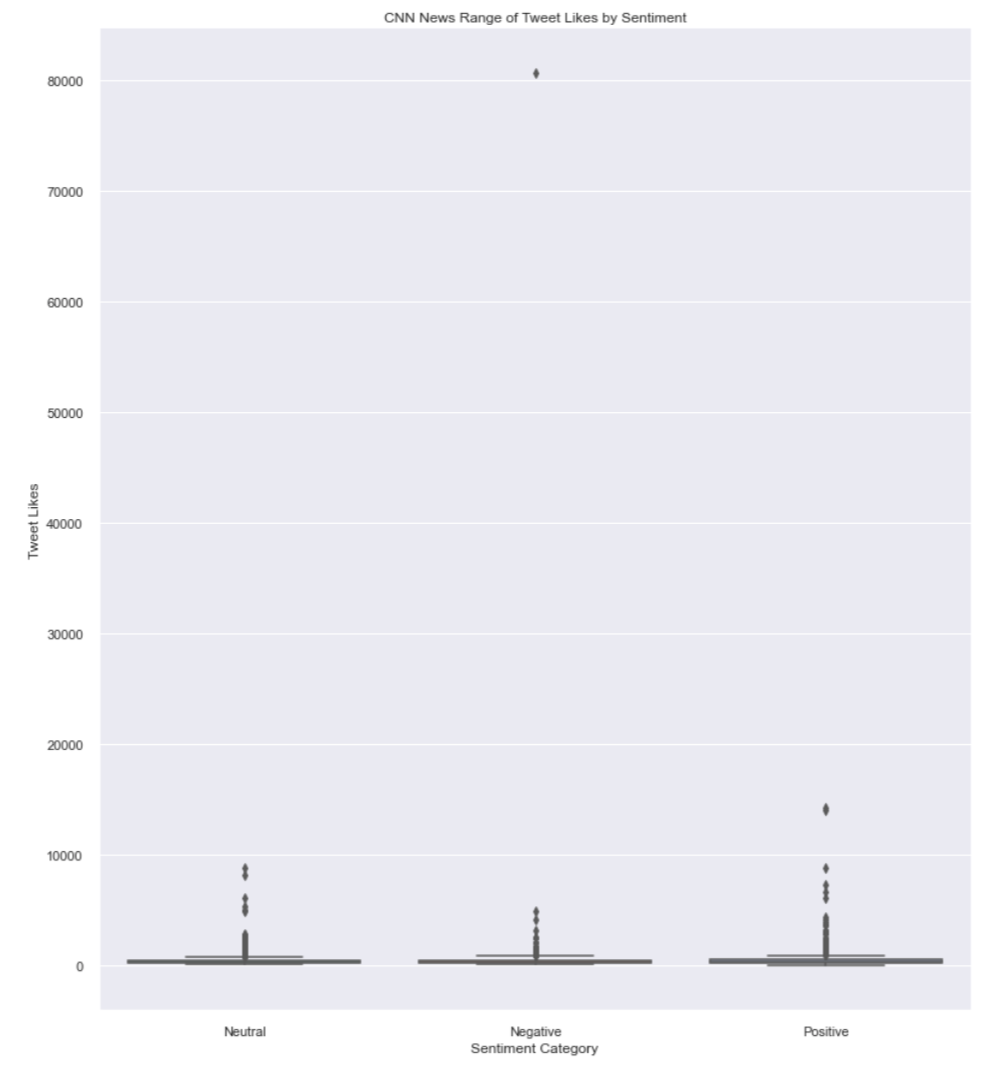
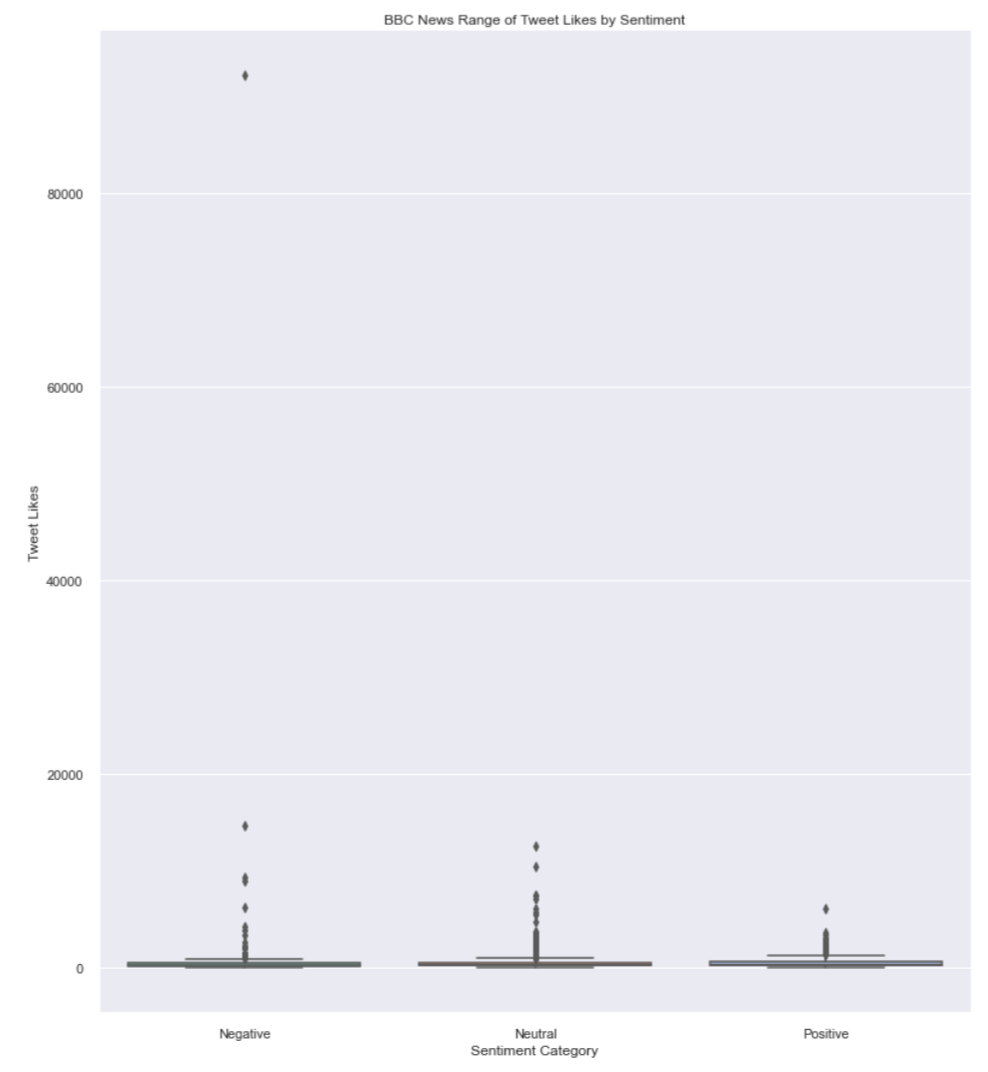
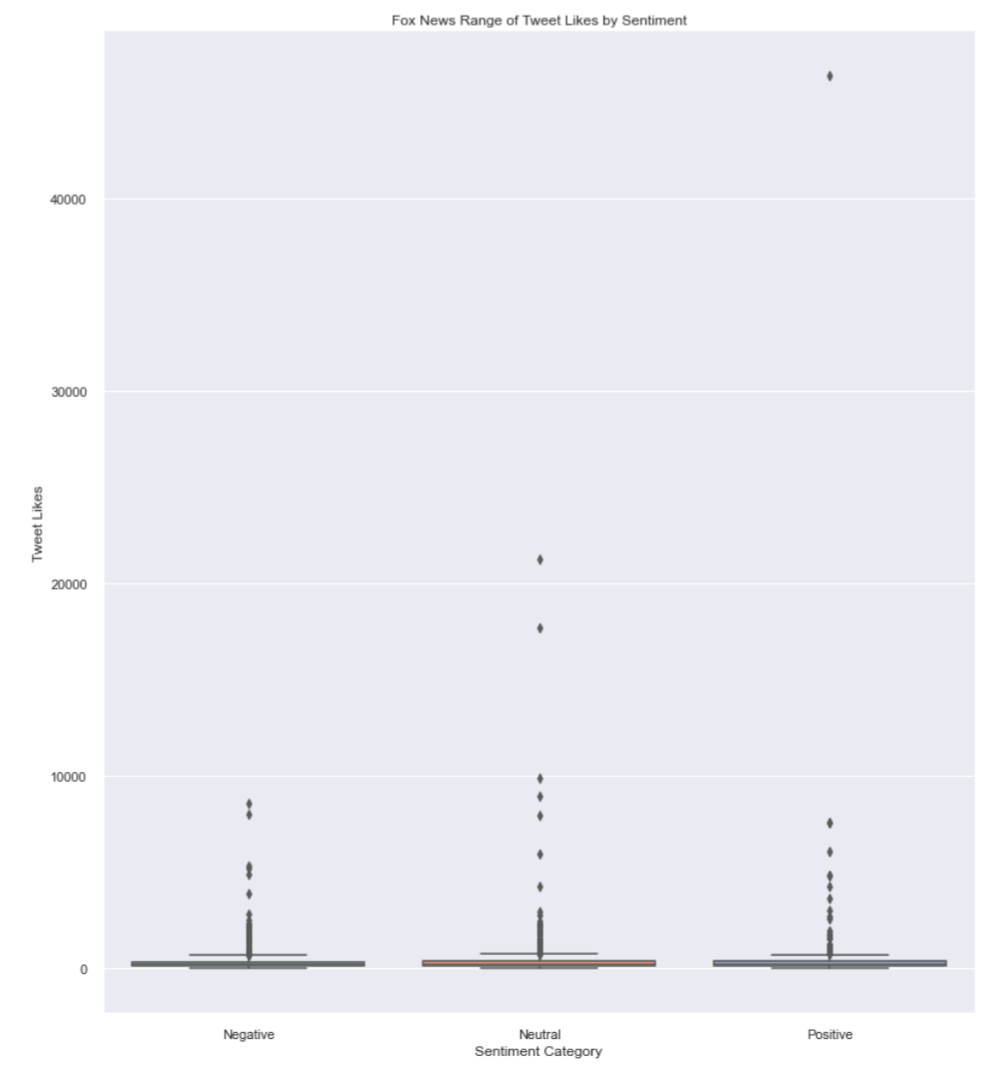
**Figure 14. CNN News Tweets Sentiment & Subjectivity**

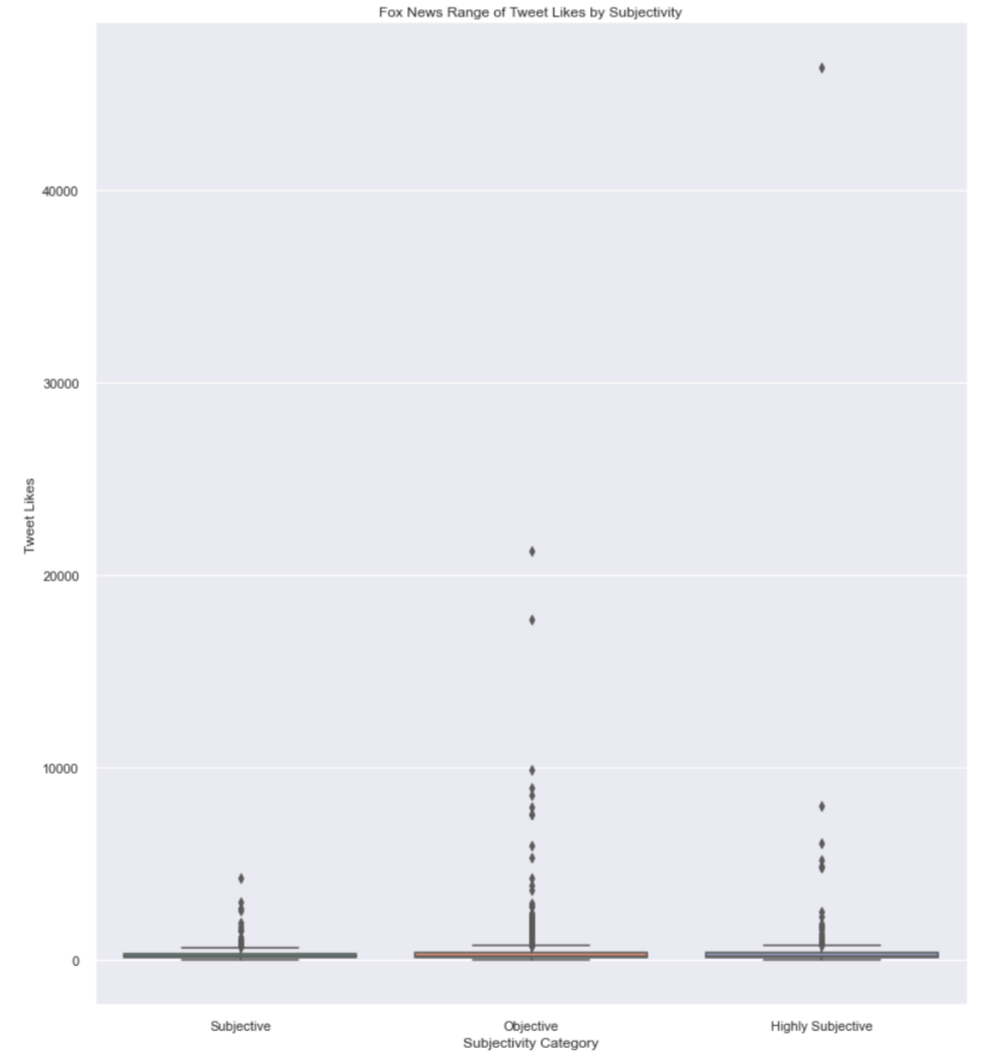
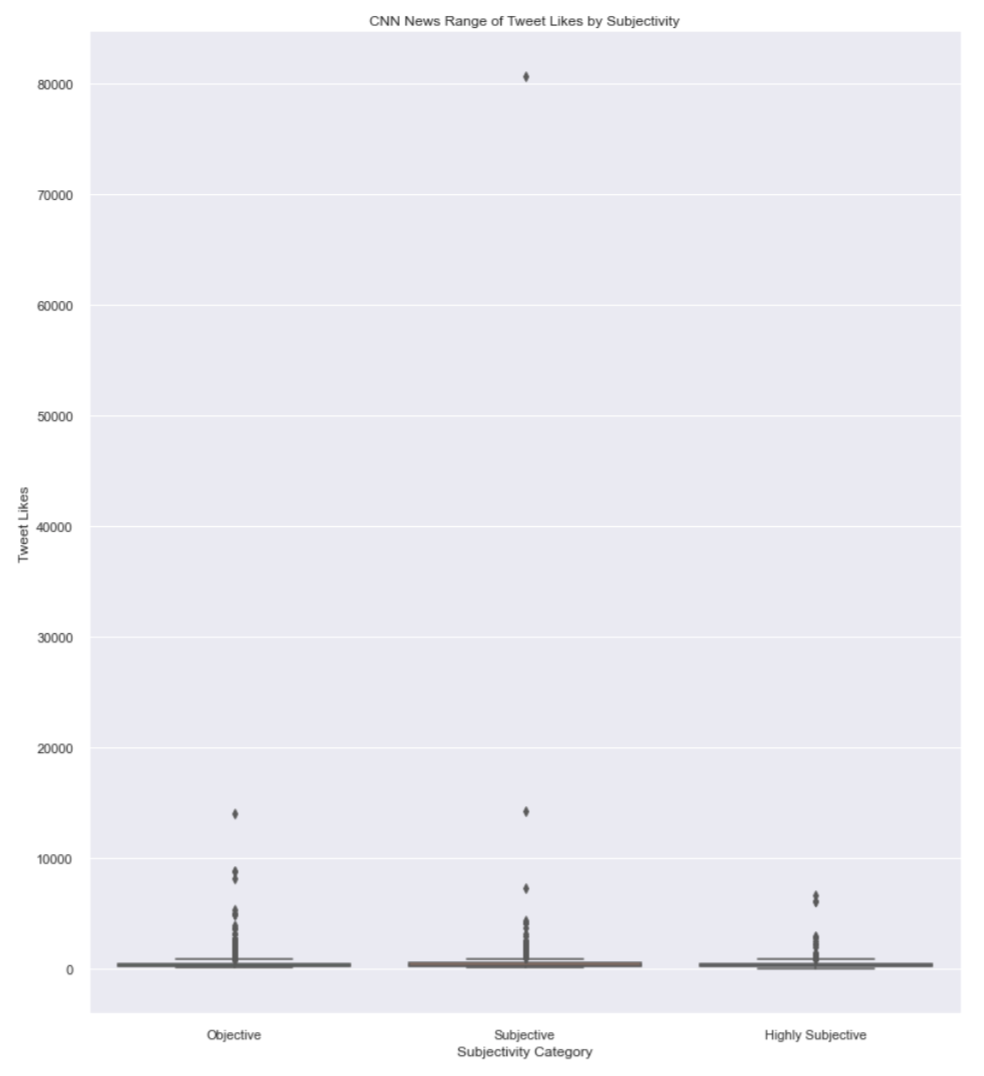
Model & Analysis: Question 3- What do the "like" counts for tweets indicate about each news outlet's audience?

The box plots provided in Figure 15 and Figure 16 indicate that there are no significant fluctuations in Tweet likes by neither of the news outlet’s audience. However, subtle differences do exist in the range of like counts when outliers are considered. Fox News has Tweet likes outlier counts that extend to 9,000 for “negative”, a little over 20,000 for “neutral”, and a little less than 50,000 for “positive” Tweets. With the exception of a single “negative” Tweet by BBC, that reached a little less than 90,000 likes, the number of outlier Tweet likes for Fox completely overshadows those of BBC and CNN. news has a higher range of Tweet likes than Fox and CNN news. Figure 16 also shows the outlier Tweet likes are the highest for Fox when the Tweets are “highly subjective”, a stark contrast to BBC and CNN’s lowest outlier counts for the respective category of subjectivity. Together, both figures show Fox news has a significant following over BBC and CNN news, where outliers on the high extremes of the distribution are attained mostly through “positive” and/or “highly subjective” Tweets. BBC and CNN news, do not have as many outliers as Fox, but do show a trend that their outliers on the high end were attributed to negative/objective and negative/subjective Tweets, respectively.

**Figure 15. Sentiment Tweet Like Range by News Outlet**



**Figure 16. Subjectivity Tweet Like Range by News Outlet**

**** **** 

Analysis: Essential Question- How are political affinities influencing news reporting via Twitter?

Based on the clusters found in the analysis, there is evidence to support Fox news informed the public on topics that directly appealed to far-right audiences, such as the Russia probe, gun-regulations, and inflation, a topic that reflects poorly on the current administration. Other topics covered could be considered superficial compared to those covered by BBC, the centrist news outlet covering global news and at significantly higher rates of objectiveness than both Fox and CNN. It is due to the variance in topic coverage that a direct comparsion between all three news outlets was no achievable. In spite of how close Fox news’ sentiment scores were to BBC, their news topics were not comparable to BBC or CNN. The Uvalde school shooting was the only cluster that could be compared across all three news outlets and unveilved details pertaining to the event were completely omitted, the focus was on details that would deter gun regulation support. Though accurately reported, CNN’s overall subjectivity score was the lowest of the three news outlets, meaning their news reporting needs to focus on improving their opinion-free reporting. Generally speaking, the following steps should be explored to enhance the findings on each news outlet’s reporting practices:

* Use a query in the Snscrape code that collects tweets during specific and non-continuous 24 hour intervals to vary.
* Assess score ranges selected for polarity and subjectivity labels, is the “neutral” range too big?
* Look into what makes Fox News 'far-right', are opinion-based “reporting” being shared on a different account?
* Optimized number of clusters needs to be re-evaluated with potential pre-processing steps that include bigrams.
* Group tweets by matching it to dictionaries of terms associated with specific topics like gun violence, elections, economics, etc., followed by sentiment analysis of each topic to compare how it is reported on accurately.

Conclusion

Although improvements to the model are needed to make more accurate conclusions, it is evident that political affinities are affecting the types of topics being reported on and the level of detail, where reporting objectivity decreases with far-left and far-right political affinities. Furthuremore, the audiences appear to like the Tweets within the same range, with no significant fluctuations in negative, positive, neutral, objective, subjective, or highly subjective Tweets being posted. However, the outliers that extend beyond the upper quartiles of all of these labeled Tweets do indicate Fox has a larger audience, that prefer highly subjective Tweets. BBC and CNN’s outlier like counts do not reach the levels of Fox news, indicating their respective audiences are much smaller. These audiences also perfer objective to subjective Tweets, with highly subjective Tweets having the smallest range in likes in the lower bounds. Given these findings, it is important to be aware of political biases in news reporting to ensure due diligence is done to search, find, and understand all perspectives before forming a stance on an issue.

Sources

<https://www.statology.org/seaborn-boxplot-multiple-columns/>

<https://www.statology.org/seaborn-grouped-bar-plot/>

<https://www.statology.org/seaborn-figure-size/>

<https://seaborn.pydata.org/generated/seaborn.scatterplot.html>

<https://www.statology.org/correlation-in-python/>

<https://www.allsides.com/unbiased-balanced-news>

https://www.analyticsvidhya.com/blog/2021/01/sentiment-analysis-vader-or-textblob/

https://medium.com/swlh/tweets-classification-and-clustering-in-python-b107be1ba7c7

https://vitalflux.com/elbow-method-silhouette-score-which-better/

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html>