Gender & Racial Bias on YouTube

MSBA 327: Text Analytics

Anthony Philpott, Tanapat Klomjit, Sweta Kumari

Golden Gate University

Table of contents

Abstract **4**

Introduction **5**

History about the Racial and Gender Bias in New Media 6

Problem Statement **7**

Data Collection **7**

Methodology & Statistical Approaches **7**

Text Representation (Pre-Processing) 8

Text Classification and Clustering 8

Sentiment Analysis 9

Text Translation & Speech Synthesis 9

Software & Tools **9**

Text Representation (Pre-Processing) **11**

Exploratory Data Analysis **12**

Analysis and Model Result **13**

Text Clustering 13

Sentiment Analysis 26

Text Translation & Speech Synthesis (Google Trans & Google Text-to-Speech) 30

Conclusion **32**

Limitations and Future improvements **32**

References **34**

Appendix **36**

Abstract

News industries and social media inform users of current events that affect and influence our daily lives. This paper aims to analyze potential gender and racial bias within MSNBC and Fox News on their Youtube video. It is important to understand if there is explicit bias in these two news brands that go beyond anecdotal evidence. Objectivity in news allows the population to make sound decisions based on facts alone. Using natural language processing this paper covers text representation, classification, clustering, sentiment analysis, and speech synthesis. These processes were completed using the three data analysis tools of SPSS, Python, and Tableau. The results showed Fox News was skewed higher compared to MSNBC for posting videos related to the keywords of 'blm', 'women', 'woman', black lives matter'. This paper also discovered MSNBC and Fox News chose to cover or not to cover certain important topics that relate to the images of political figures who happen to be women. Both news organizations the sentiment for the video title and description were skewed neutral and the transcript was skewed positively. Lastly, this paper covers the importance of inclusion for all free from any potential language barrier an individual may encounter. Speech synthesis allows for the democracy of the news through text to speech.

Keywords: **text representation, classification, clustering, sentiment analysis, speech synthesis**

Introduction

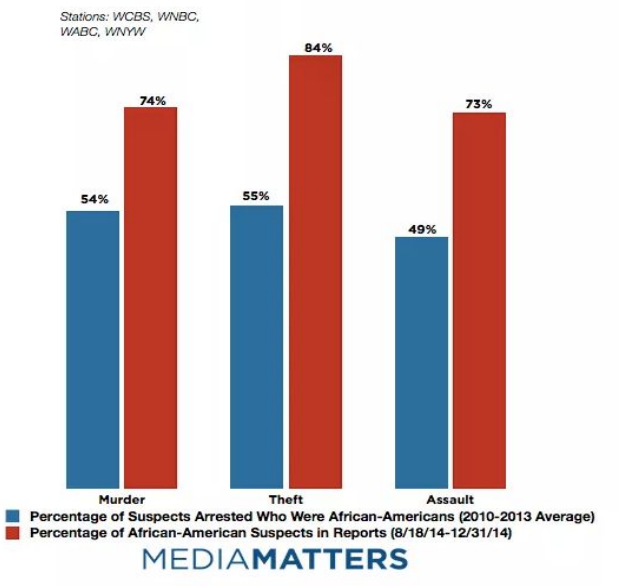
Race and gender bias issues have always been a profound impact on society. The race is sometimes regarded as fluid, old, and overhauled by social categories. It affects society in many ways in many areas. For instance, the study has proven that when Black and White job applicants sent out similar CVs to employers, Black candidates were half as likely to be called in for interviews as White job applicants with equal qualifications (Cherry, 2020). Such a discernment act is likely the result of biases toward racial groups. Though some people consider that race is restricted by biology, it is now extensively admitted that this classification system was in fact conceived for social and political reasons. It is supporting in increasing the inequities and injustice for all ethnicities and religions. Race remains an essential and influential determinant majorly in the context of conservation (Frontiers, 2019).

The other example is biasness in Fox News. Fox News already paid $10 Million to settle racial, gender bias suits. According to NPR, there are several allegations of racial discrimination against the network, along with several gender bias and retaliation claims (Folkenflik, 2018). In addition, these topics are debatable and sensitive as they seem to gain people's attention when seeing inequalities in these areas. Although we recognize that race and gender topics can be complex due to various backgrounds and experiences, it’s worth investigating how the media impact our society and can contribute to the way people understand messages. According to the eMarketer in 2019, YouTube dominates other media platforms regarding its digital video consumption with almost 90% usage from US digital viewers. (Moshin, 2020). In this project, YouTube is a primary new media for text analytics due to its popularity in the current Digital Age.

History about the Racial and Gender Bias in New Media

Several factors promote the racial and gender bias. News channels are a great resource of information with focus on delivering news to the general public. News channels have influenced racism in our society in a significant manner. It targets racial and gender bias, and therefore it is wise to explore how the news channel contributes to these biases. Many researchers analyzed the "identifier" word patterns that are used by news such as "black" and "white". The research proposes that on average, “black” is used 3 times, more in news than "white". The over usage of the word “black” becomes a racial bias because it can condition the mind to associate the word with a negative connotation (Kulaszewicz, 2015). Here is an instance that shows how news channels can negatively impact the racial bias:

“According to New York City Police Department statistics, African Americans were suspects in 54 percent of murders, 55 percent of thefts, and 49 percent of assaults. But the suspects in the stations' coverage were black in 74 percent of murder stories, 84 percent of theft stories, and 73 percent of assault stories” (Harris, 2015, para2). The negative images shown in the media affect the black people in society.



Problem Statement

The scope of this project is to analyze different TV news on their YouTube channels (Fox News, CNN, CBS News & MSNBC), and conduct statistical analysis based on video title, transcript and descriptions. The analysis aims to examine how various news channels portray any gender and racial bias. The goal of the analysis is to analyze if there is explicit bias in news brands.We accomplish this goal by conducting analysis such as text representation, classification, clustering, sentiment analysis, and speech synthesis. Our goal also includes to convert the news transcript in language such as Hindi, Thai to avoid language barrier. This allows users to listen to the content of the news without knowing the specific language.

Data Collection

The source of our data is derived from YouTube with a focus on the Fox News, CNN, CBS News & MSNBC subscription pages. The data will consist of 806 videos from news brands that pertain to race and gender topics. These videos will be ordered by relevance and will be between the dates of June 2019 and June 2020. The qualitative features of interest that pertain to text analysis are the video title, video description, and the video transcript. The quantitative features will consist of video metrics of the videos such as view count, video plays, etc.

Software & Tools

1. Python Programming Language: The Python programming language is used to complete the task. This includes data manipulation, analyzing the text, visualizations, etc.
2. Few Python Libraries/modules/client:

NLTK (Natural Language Toolkit): is a leading platform for building Python programs to work with human language data (NLTK, 2020).

pandas: is a library that is a ” data analysis and manipulation tool”(Pandas,2020).

NumPy: is a library that provides ”scientific computing” (Numpy,2020).

pprint: module provides a capability to “pretty-print” arbitrary Python data structures in a form which can be used as input to the interpreter. (pprint, 2020)

nltk.sentiment.vader: VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to polarity (positive/negative) of emotion (Hutto & Gilbert, 2014).

YouTubeDataAPI: is a Python client for the YouTube Data API. The youtube-data-api package is a wrapper to simplify GET requests and JSON response parsing from the API (YouTube Data API, 2020).

YouTubeTranscriptApi: is a python API which allows us to get the transcripts/subtitles for a YouTube video. It also works for automatically generated subtitles, supports translating subtitles (youtube-transcript-api 0.3.1, 2020).

Googletrans: is a free and unlimited python library using the Google Translate Ajax API to detect and translate text. This method is reliable because it uses the same server as translate.google.com.

Google Text to Speech API (GTTS): uses to convert text to speech. GGTS supports several languages including English, Hindi, French, Thai and many more.

3. **IBM SPSS Modeler Text Analytics**: offers powerful text analytic capabilities, which use advanced linguistic technologies and Natural Language Processing (NLP) to process a large variety of text data, extract and organize the key concepts (IBM, n.d.).

4. **Tableau**: is used as our main data visualization tool aftering export statistical results in .csv file format.

Methodology & Statistical Approaches

The objective of the project is to obtain insights from videos posted on Youtube by some of the most prominent American news channels. We use Python and IBM SPSS Modeler to accomplish the goal of the analysis. We start with scraping the data using YouTube API. With Python, We connect to the YouTube API in order to pull both the qualitative and quantitative features. With the help of Python's module, libraries, packages, we retrieve the YouTube videos attributes along with the transcript of each video. The data is later used to perform pre-processing, exploratory data analysis, manipulation, sentiment analysis and speech synthesis. Python library Pandas and NLTK corpus package remove stop-words, transform cases, punctuations, and perform lemmatization and stemming. IBM SPSS Modeler is used to perform cluster analysis in order to get insights about the videos and how news channels are biased towards race and gender.

Text Representation (Pre-Processing)

Text Representation is the process where we retrieve all the information about the YouTube videos. We use google api python client, youtube-data-api, youtube_transcript_api to connect and retrieve information about the YouTube videos. Later, we retrive the transcript of the videos to get the text. Our aim is to retrieve video title, transcript, and description for our analysis.

Text Classification and Clustering

We perform text classification and clustering to get insight about the race and gender-based topic. We use IBM SPSS and its text mining capabilities to explore the bias. The analysis is performed on the YouTube video attributes such as the video captions, title, and description. We focus on clusters that only pertain to gender and race.

Sentiment Analysis

We provide an analysis of the text from each of the news brands to determine the positive, negative, and neutral of the text. We use the Vader module and polarity score to get the score of positive, negative, compound and neutral. The sentiment is analyzed for the video transcript, title, and description.

Text Translation & Speech Synthesis

Translation is necessary for the spread of information between different cultures. In this project, we perform a text translation from the original video title in English to other languages

like Hindi and Thai using Googletrans API. After that, we convert text input to speech by machine using Google Text to Speech API (GTTS).

Text Representation (Pre-Processing)

Before proceeding with text analysis, we installed the google api python client, youtube-data-api, youtube_transcript_api to connect and retrieve information about the YouTube videos (See Appendix 1). After this, we imported the necessary libraries for pre-processing of the data (See Appendix 2). API Key is used to connect to YouTubeAPI (See Appendix 3). YouTubeDataAPI is used to search the video and retrieve information about the videos. A search result contains information about a YouTube video, channel, or playlist that matches the search parameters specified in an API request (YoutTube Data API, n.d.). We retrieved information for FOX news channels (See the Appendix 4 for FOX channel for reference). Same process has been repeated for other three channels (CNN, MSNBC, CBS_NEWS).

We retrieved below attributes from video:

`video_id, channel_title, channel_id, video_publish_date, video_title, video_description, video_category, video_thumbnail, collection_date`.

After retrieving the data in JSON file format, all files turned into a dataframe and combined to get one dataframe for our analysis (See Appendix 5). The result dataframe has 806 rows with 9 columns and contains information about the all four channels. Our scope of analysis is to get the sentiment score for video transcript as well. To retrieve the information about the transcript of each video, YouTubeTranscriptApi.get_transcript method is used (See Appendix 6). It gave the transcript of all videos along with video id. To get all attributes in one dataframe,

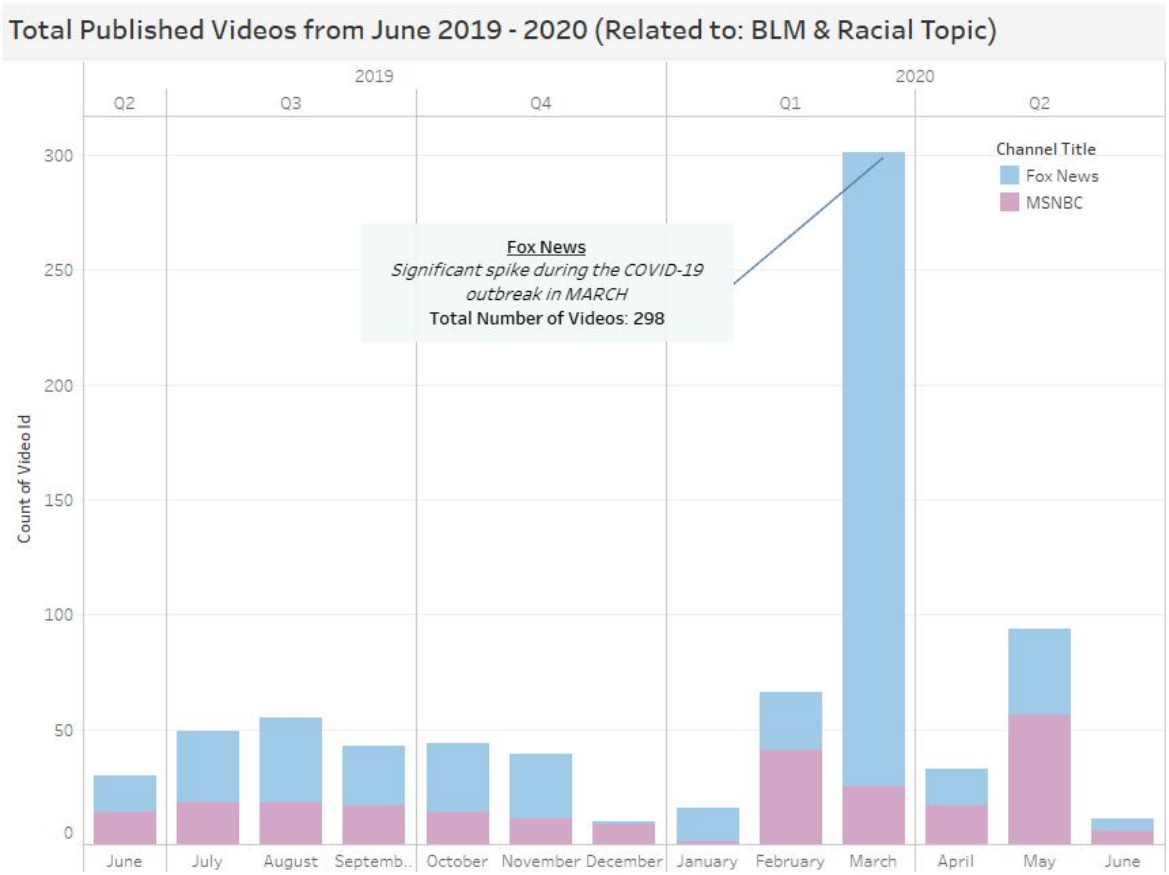
Inner join was used to join the dataframe “final_df “and “result” (See Appendix 7). This data frame is used for sentiment analysis.

Exploratory Data Analysis

Total count of Records per channel (See Appendix 8)

```
Fox Records:  544
Msnbc Records:  247
Cnn Records:  15
Cbs Records:  0
```

In this EDA section, our goal is to conduct an initial investigation over the original dataset seeing what the data can tell us via statistical analysis. As shown in the result Appendix 8, we retrieved different numbers of published videos because some channels like CNN and CBS did not enable its transcript option. This finding helps us to rescope our direction to focus on Fox and Msnbc news. Thus, our primary dataset for text analysis consists of 791 videos.



The analysis begins with observing the overall trend regarding the number of videos over the period from June 2019 to June 2020. The bar chart represents the comparison between Fox news and MSNBC videos. As a result, Fox news clearly posted more videos than MSNBC related to those keywords `['blm','women','woman','black lives matter']` in the parameter. Fox published about 298 videos related to racial content (almost 614.7% more than its monthly average of published videos from 44.9) Interestingly, the massive spike of published videos happened in March 2020, in which the COVID-19 outbreak happened regarding national emergency, banned on international travel and Statewide Stay-at-Home Order (AJMC staff, 2020). In addition, we also speculated another surge of published YouTube videos in May. Two of selected keywords are `blm` and `black lives matter`, this spike might happen due to higher media traffic for Mr. Floyd's incident on May 25, 2020. Consequently, we expected this increasing trend of more published videos over this content due to more civil right movements around the globe.

Analysis and Model Result

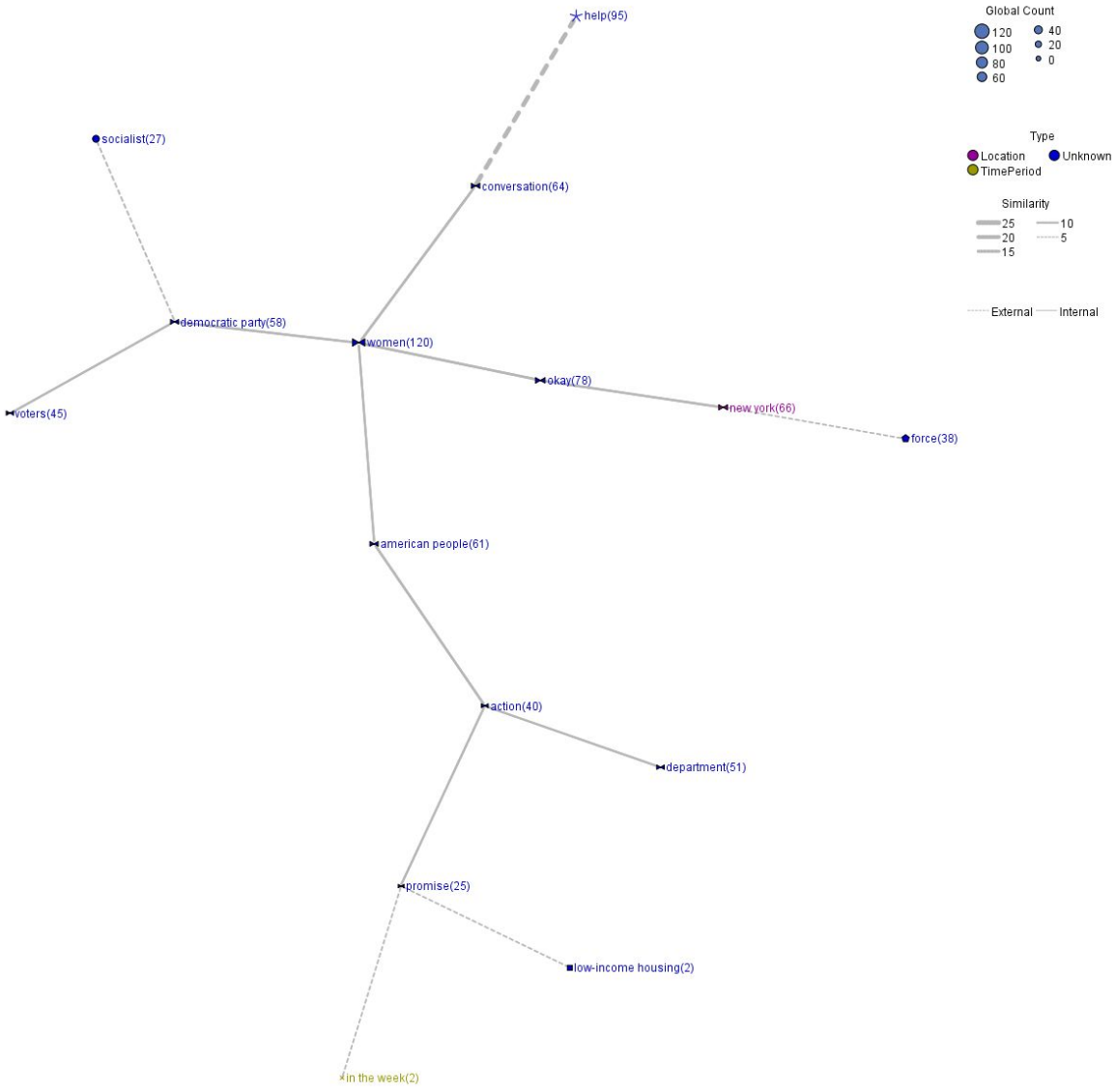
Text Clustering

To aid in providing more structure to our news project the decision to perform text clustering was decided. We wanted to explore the topics of race and gender-based on the clusters that were created. The environment in which we perform this analysis was done in SPSS and utilizing its text mining capabilities. Each dataset consisted of separate data from MSNBC and Fox News and was set with a maximum of 100 clusters. There was also a maximum of 10 concepts, 20 internal links, and 20 external links. Lastly, the minimum concepts in the cluster

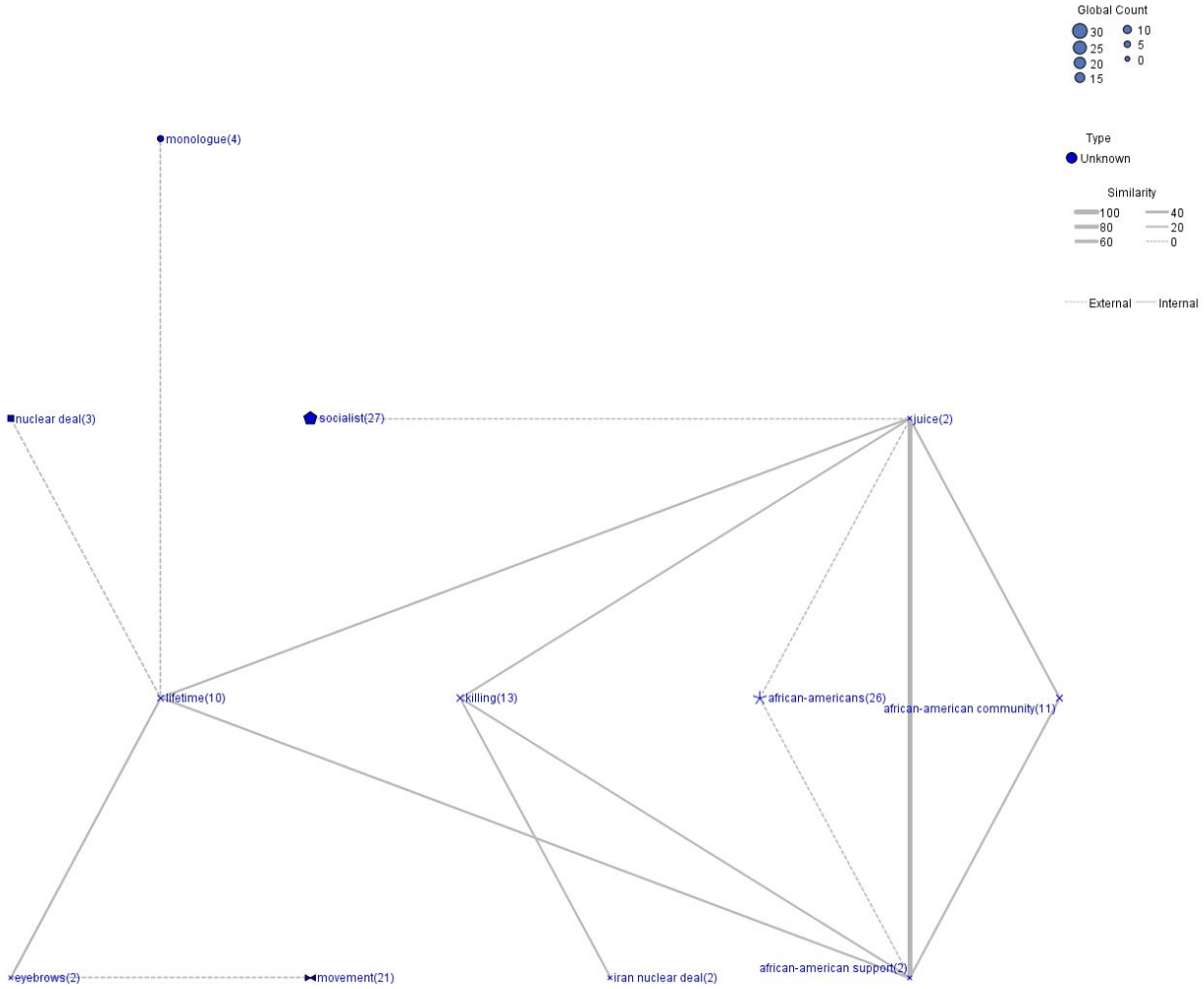
were set to 3, and the minimum link value set to 2. The decision was to focus only on clusters that only pertain to gender and race. The text clustering analysis performed on the video captions, title, and description.

Focusing on the video captions for Fox News had clusters of women, Amy Klobuchar, African-American support, Terry Reed, Molly Line, and Speaker Nancy Pelosi. Women were represented in 10 concepts, 9 internal and 20 external links. When examining the women in the concept web there is a strong internal similarity to conversation, democratic party, okay, and the American people. There are also three external links that stood out which were socialist, help, and force. The cluster Amy Klobuchar yielded 10 concepts, 9 internal and 18 external concepts. Strong internal concepts tied to her were good ideas, middle class, racism. Harmful misinformation was tied to her externally. African American support was represented in 7 concepts, 9 internal and 18 external links. The strong internal links were killing, African-American community, lifetime and the external links were socialist and movement. Terry Reed who accused Joe Biden of sexual assault was represented in 8 concepts, 7 internal and 20 external links. The strong internal link between heart, claim and external link of lives, love, and guy. Molly Line who is a Fox News Correspondent was represented in 4 concepts, 4 internal and 13 external links. The strong internal links are priest, big picture, and external links of mother. Speaker Nancy Pelosi was represented in 9 concepts, 19 internal and 18 external links. The strong internal concepts are secretary of defense, sources, demand, briefing and the strong external links consist of articles of impeachment, Iraq, and methods. The visual representation of women and african-american support are below:

Women


help(95)
socialist(27)
conversation(64)
democratic party(58)
women(120)
okay(78)
new york(66)
voters(45)
force(38)
american people(61)
action(40)
department(51)
promise(25)
low-income housing(2)
in the week(2)
Global Count
120
100
80
60
40
20
0
Type
Location
Unknown
TimePeriod
Similarity
25
20
15
10
5
External
Internal

African-American Support

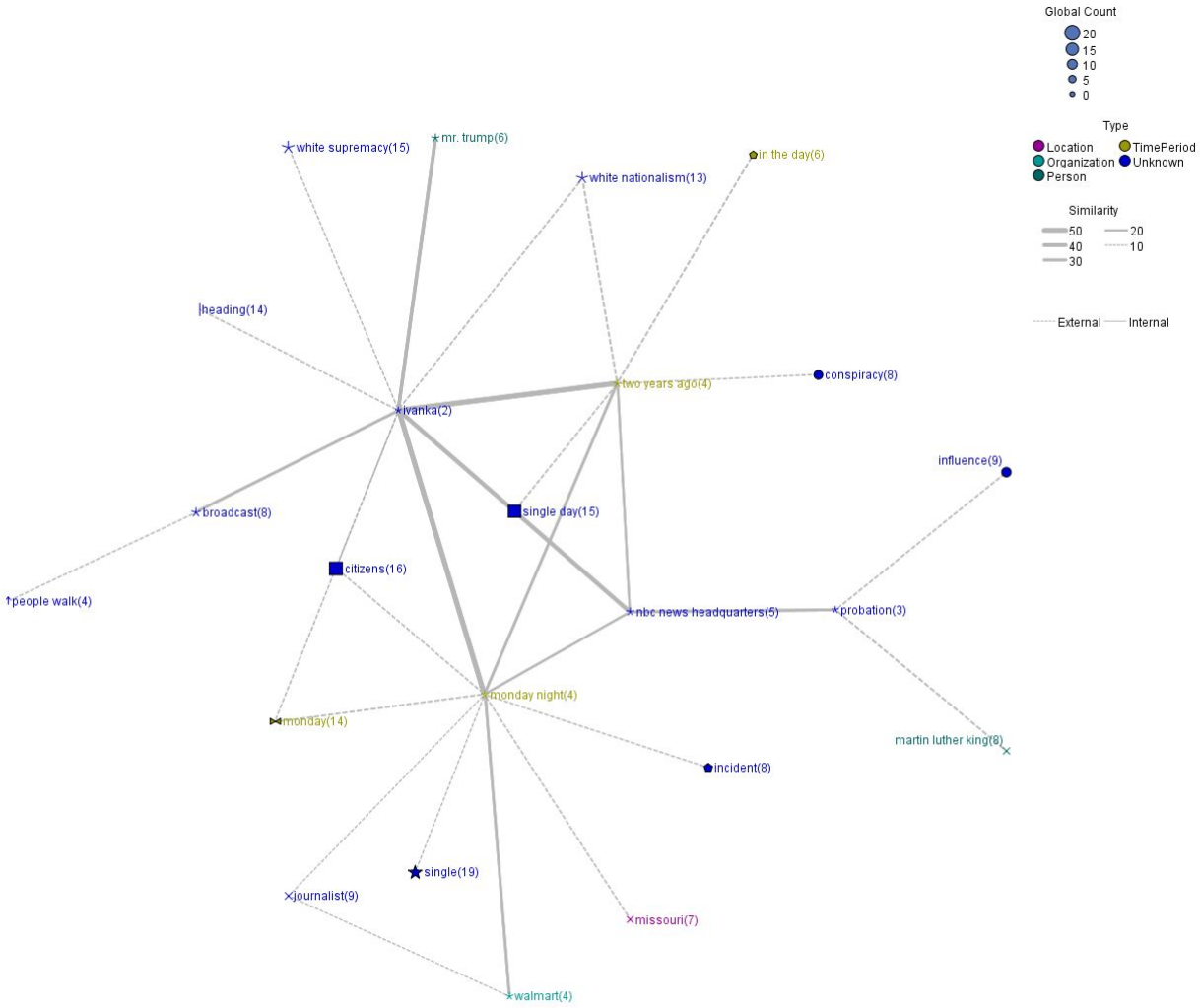


Focusing on video captions for MSNBC the clusters we have are African-American voters, agenda of white, black man, black people, congresswomen of color, Harvey Weinstein, illegal immigration, Ivanka, ladies, and racism. African-American voters were represented in 8

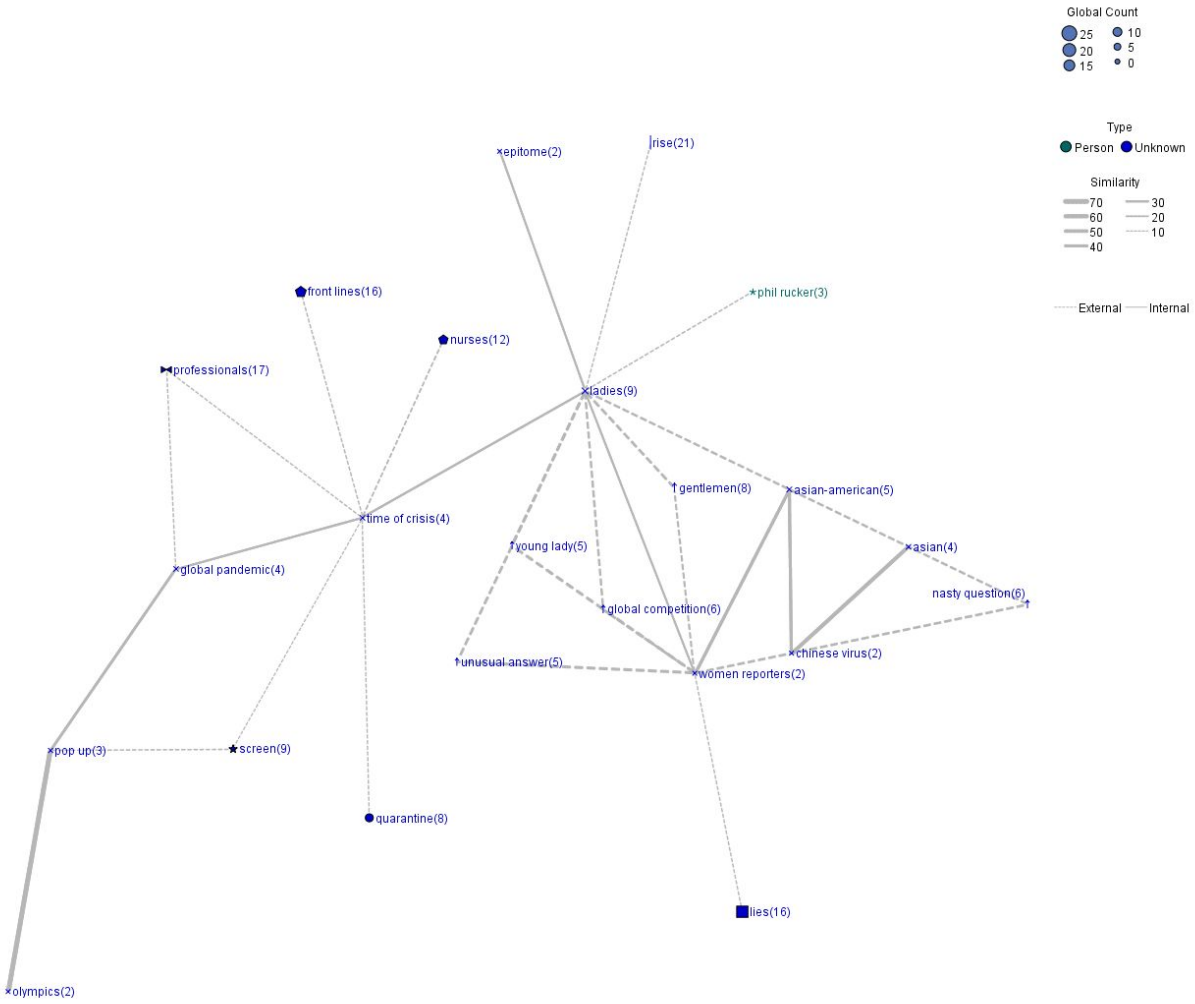
concepts, 7 internal, and 19 external links. When examining the African-American voters in the concept web the two internal similarities were achilles heel and pledged delegates. When viewing the external links the concepts of democratic nominee, results, and defeat Donald and African American community were observed. Agenda of white were represented in 9 concepts, 9 internal, and 20 external links. The strong internal links were members of congress, agreement, distraction, bait, people saw and the external links were security, key, denial, and women of color. Black men were represented in 10 concepts, 10 internal, and 18 external links. When examining the black men in the concept web the internal similarities were black, murder, video, rights, and young man. When viewing the external links knee, social media, attack, language were concepts. Black people were represented in 9 concepts, 8 internal, and 14 external links. The strong internal links were feeling, riots, slavery, and the external links were conviction, black lives tears, senator, and young people. Congresswomen of color were represented in 10 concepts, 12 internal, and 20 external links. The strong internal links were brown people, birtherism, pushback, and the external links were Charlottesville, race-baiting, space, influence, and presidential campaigns. Harvey Weinstein was represented in 10 concepts, 15 internal, and 20 external links. The strong internal links were rape, greater risk, sexual assault, and the external links were assault, survivors, privilege. Illegal immigration was represented in 10 concepts, 11 internal, and 20 external links. The strong internal links were strong economy, invaders, sites, invasion, and the external links were immigration, mass shooting, American life, profit and connection. Ivanka was represented in 8 concepts, 10 internal, and 20 external links. The strong internal links were Mr trump, Walmart and the strong external links were white nationalism, white supremacy, heading, and single. Ladies were represented in 10 concepts, 9

internal, and 20 external links. The strong internal links were time of crisis, women reporters, and the external links were Asian-American, Asian, nasty questions, unusual answers, front lines, Chinese virus, professionals. Racism was represented in 10 concepts, 9 internal, and 20 external links. The strong internal links were republican party, racist, Mexicans, immigrants, party and the external links were attack, women, election, votes, republicanism, democrats, and democratic party. The visual representation of Ivanka and ladies are below:

Ivanka



Ladies



Comparing Fox News and MSNBC video caption transcriptions a few key insights were discovered. Fox News mentions Amy Klobuchar who was running for president of the United States in the democratic party. The cluster racism was linked to her which could be attributed to

her previous history on voter suppression and the disenfranchisement of voters of color as a senator in Minnesota. No mention of Amy Klobuchar in the top 100 clusters on the MSNBC dataset. One can infer that the reasoning was to discredit Amy Klobuchar due to her running against President Trump. MSNBC's decision not to include her in concepts could be viewed as concealing harsh truths.

Terry Reed is another example of the big difference between the coverage of Fox News and MSNBC as there wasn't wany mention of the sexual assault accuser to Joe Biden. Again the reasoning could be similar to the Amy Klobuchar situation or could be the refusal to publish a story that could be viewed as inaccurate. Regardless of the case, the news is supposed to be impartial to political viewpoints and report events that are accurate.

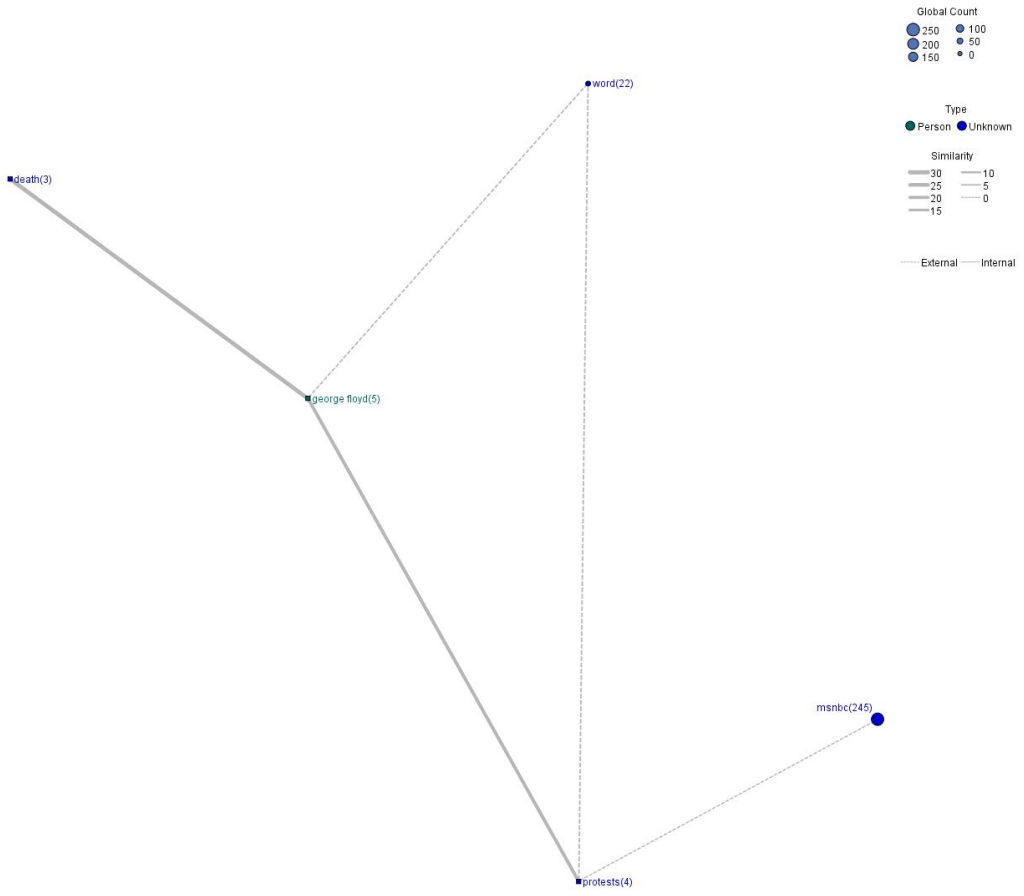
Ivanka was a cluster in MSNBC that did not appear in Fox News and was associated with concept links of white nationalism and white supremacy. Being President's daughter and who contributes to government affairs the association to these terms could be due to her silence on the issue.

Lastly, Ladies stood out during the analysis which appeared on MSNBC but not Fox News and was associated with the concepts of women reporters, Asian-American, and nasty questions and the Chinese virus. These associations could be due to President Trump's attack on women reports and especially those for are Asians. The attacks extend to commenting on asking nasty questions and the balming of COVID-19 on the Asian community. The depiction of this not in Fox News potential to decrease the negative connotation between these words and the president of the United States.

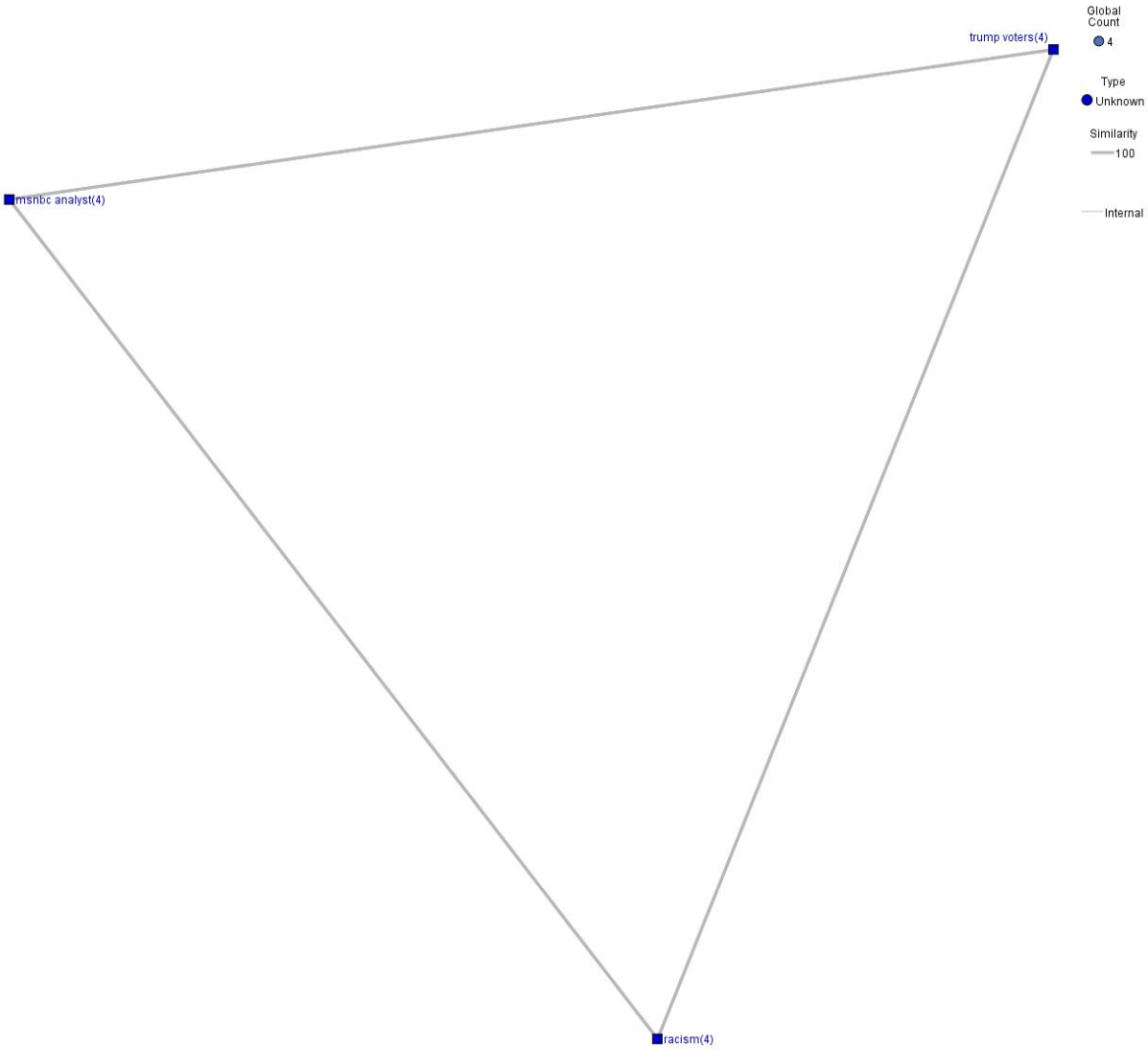
Next was the performance of the text clustering on the video title for Fox News and MSNBC. Fox News provided clusters of MSNBC Analyst and division. MSNBC Analyst had 3 concepts, 3 internal, and 0 external links. The internal links were racism and trump voters. Division had 3 concepts, 3 internal, and 0 external links which were RNC chairwoman, Nancy Pelosi.

MSNBC provided 1 cluster of importance which was George Floyd. George Floyd was represented in 3 concepts, 2 internal, and 3 external links. The strong internal links were protest, death and the external links were word.

George Floyd



MSNBC Analyst

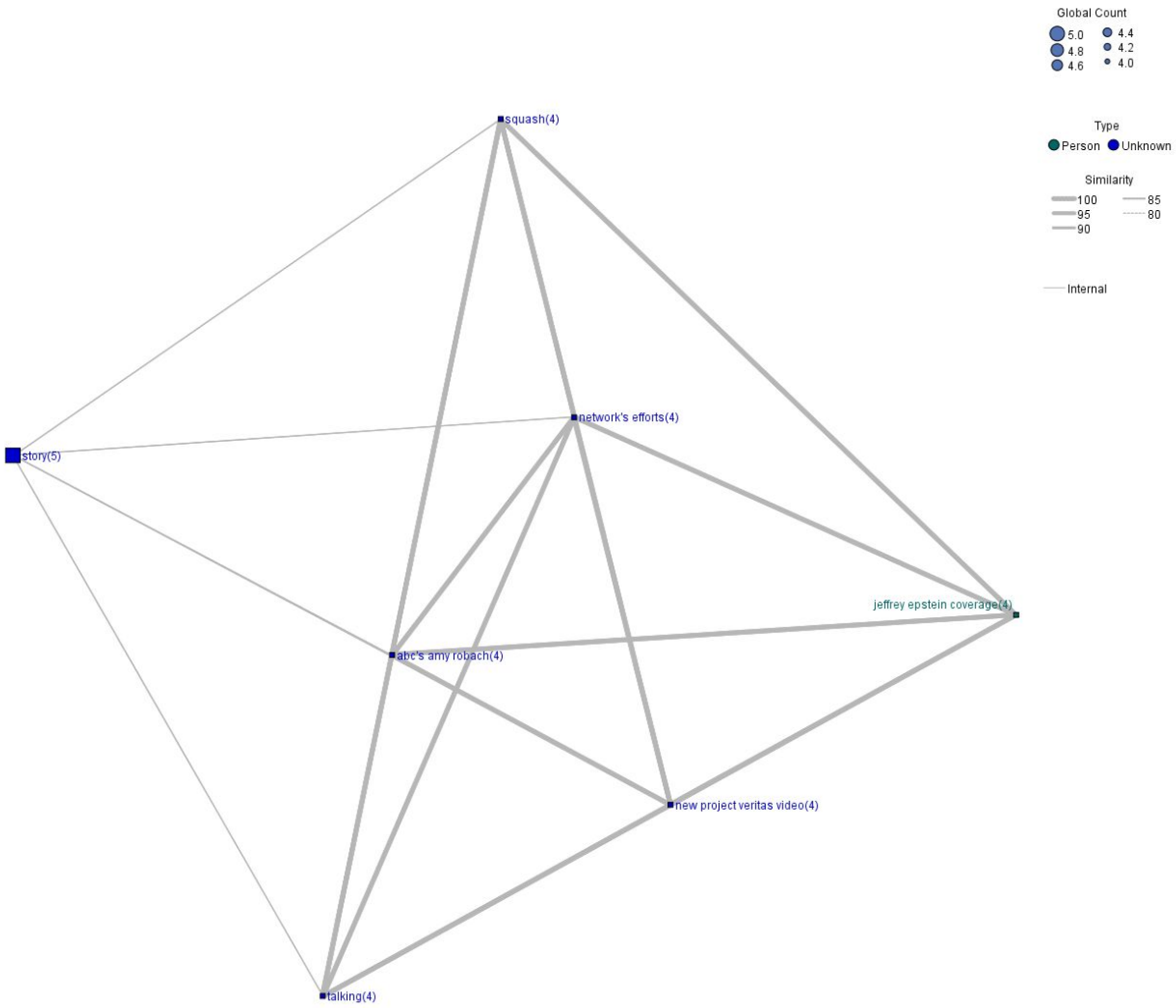


The difference between the clusters in the video title category skewed more abrasive from Fox News than from MSNBC. The lack of overall clusters makes it harder to determine similarity or dissimilarity between news brand titles.

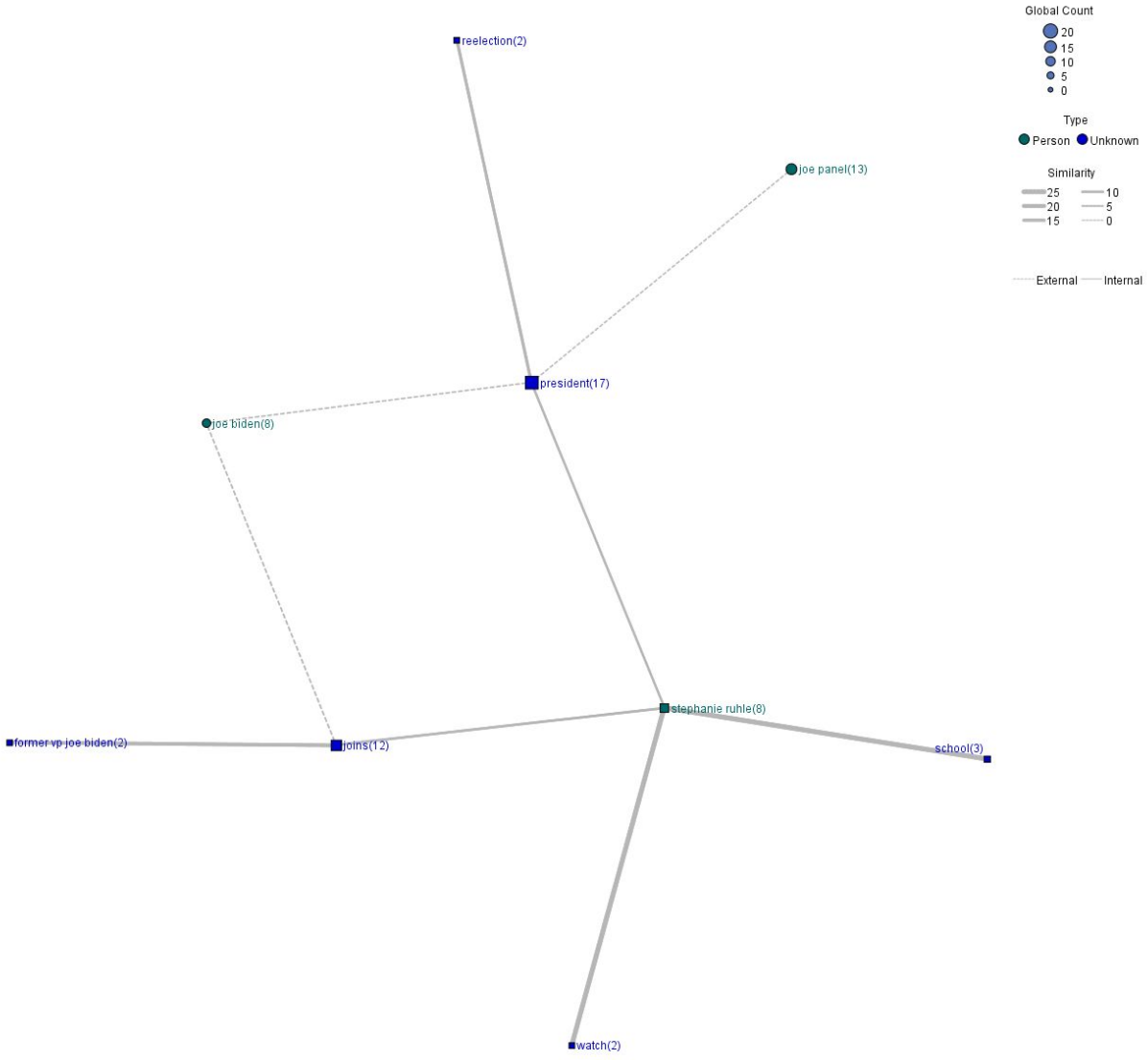
The last portion of the cluster analysis consisted of the video description of the two news entities. Running clusters analysis on Fox New's video description yielded #ingragamangle and ABC's Amy Robach for relevant clusters. #ingragamangle which refers to Laura Ingraham who is the host of Fox News Ingraham Angle has 5 concepts, 7 internal, and 7 external links. The internal links were toxic masculinity mindset, coronavirus, lives, fox business. The external links were president trump, Fox news channel. ABC's Amy Robachhad 7 concepts, 20 internal, and 0 external links. The internal links were network efforts, story, Jeffery Epstein coverage, and squash.

MSNBC's video descriptions cluster results yielded Stephanie Ruhle who is a MSNBC anchor which has 7 concepts, 6 internal, and 3 external links. The internal links were president, joins, former VP Joe Biden. The external links were Joe Biden and Joe panel which refers to the MSNBC Morning Joe show.

ABC's Amy Robach



Stephanie Ruhle



When it comes to video descriptions between the two brands there were not a lot of clusters. Though Fox News did lead when it comes to having more concepts that relate to race or gender. Toxic masculinity mindset ABC's Amy Robach stood out the most out of the clusters.

ABC's Amy Robach pertains to the television anchor who was caught on a hot mic saying ABC killed her report on Jeffery Epsitn and the Virginia Roberts story.

Sentiment Analysis

After retrieving the attributes of YouTube videos (discussed in section "Text Representation (Pre-Processing))", Three attributes video transcript, video title, video description are used to analyze the sentiment towards race and gender based topics. We use the Vader (Valence Aware Dictionary for Sentiment Reasoning) module to analyze the sentiment and get the score for positive, negative, neutral and compound of the emotion. "The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive)" (Sentiment Trading, 2020). It is helpful in structuring the dataset. We computed the sentiment score for all three attributes: video transcript, video title and video description (See Appendix 9).

We also determine the polarity that takes into account the amount of positive or negative terms in a given sentence. Polarity is determined based on below condition:

- If the compound score is greater than 0.2, it is considered to be positive sentiment and labelled as 1.
- If the compound score is less than -0.2, it is considered to be negative sentiment and labelled as -1.
- In other cases, it is considered to be neutral sentiment and labelled as 0.

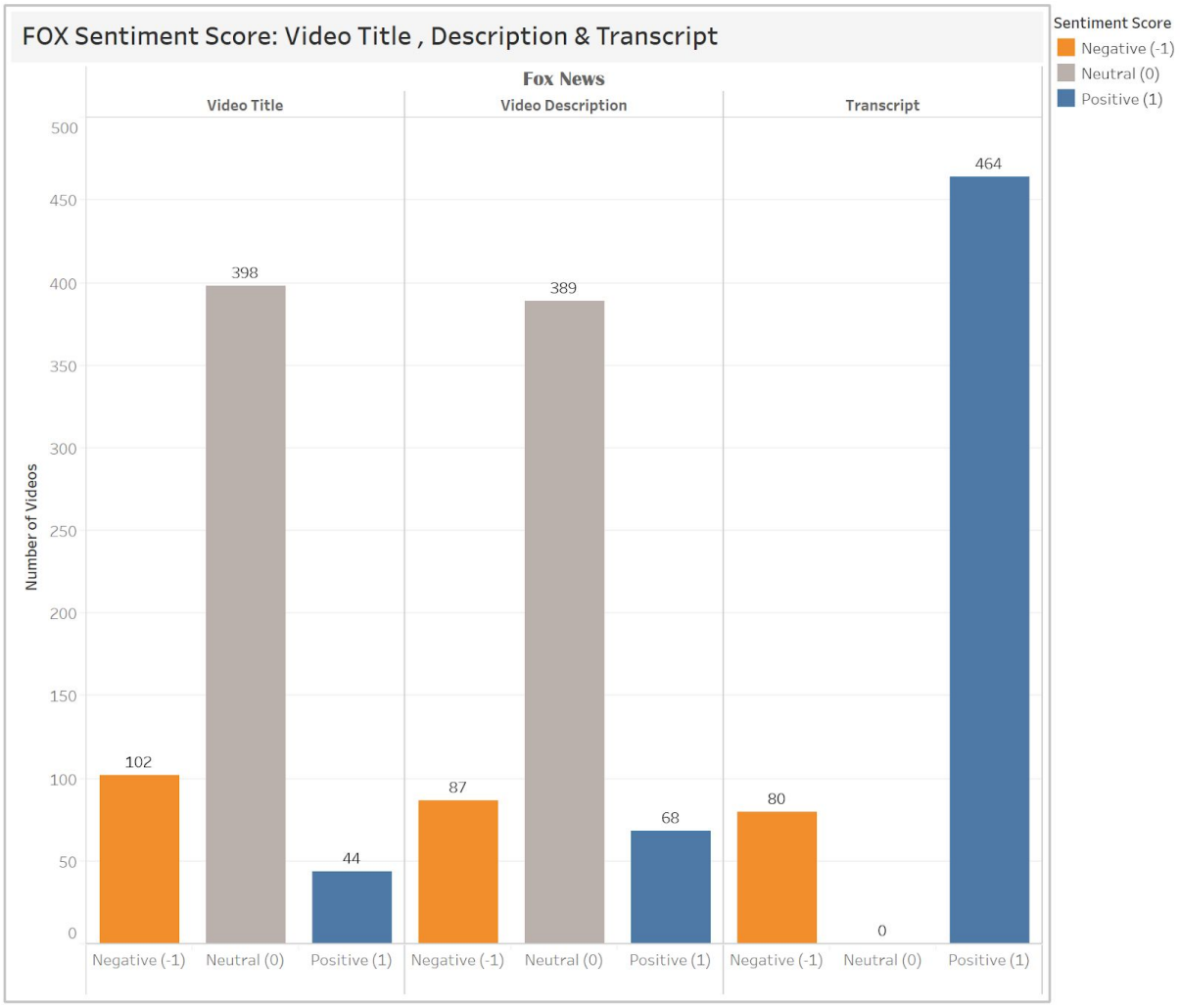
The polarity was determined for all three attributes: video transcript, video title and video description. We have shown the polarity by trans_label, title_label, description_label for video transcript, video title and video description respectively (See Appendix 10, 11, 12).

We started merging all the dataframes: the dataframe consists of all attributes and the dataframes where we got all sentiment scores. In other words, each dataframe that was created for sentiment score was merged to get the final dataframe where we drop all the negative, positive, neutral, compound scores and keep the polarity for simplification and analysis.

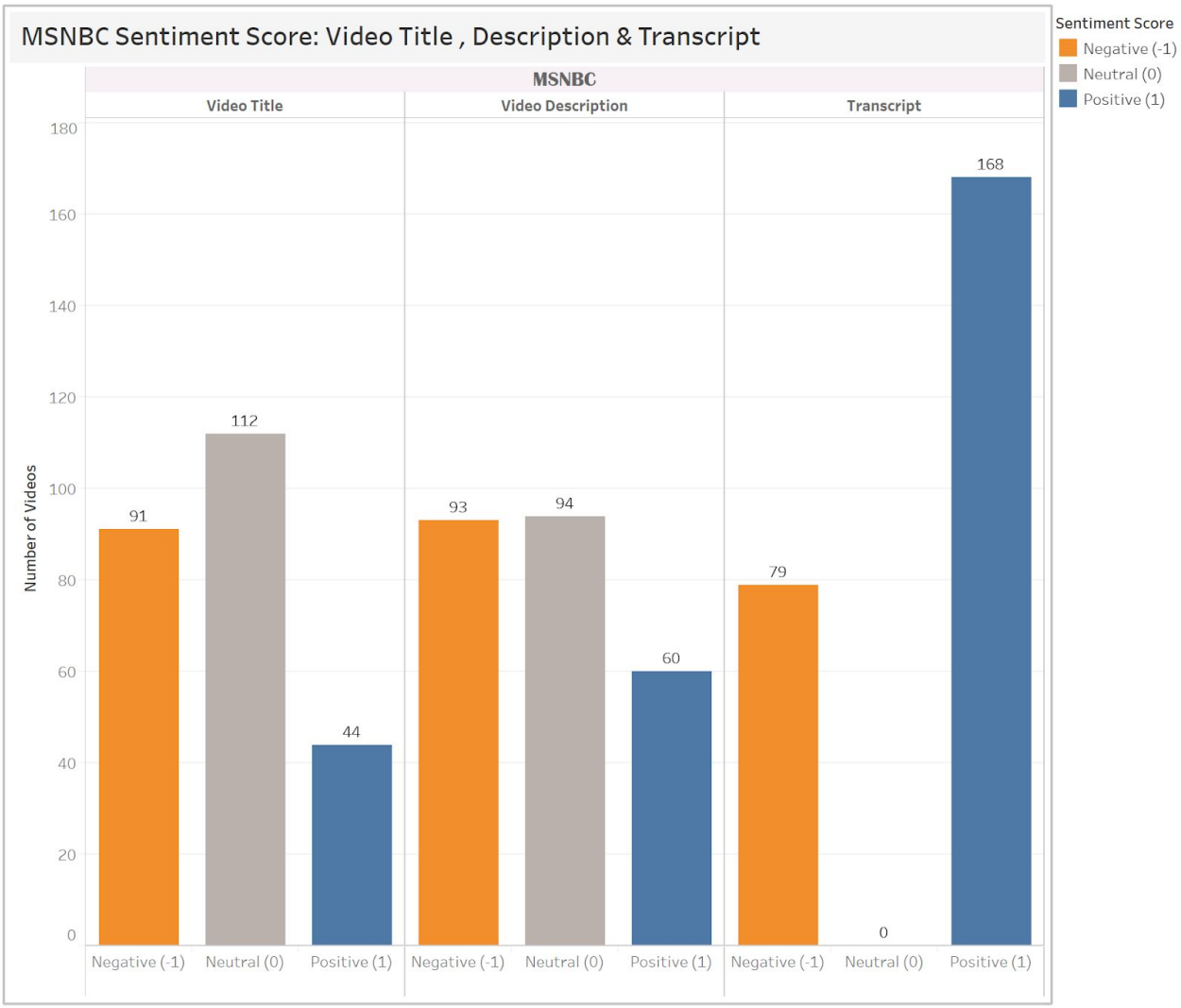
To analyze the sentiment score of each news channel, we counted polarity for each value (1, -1, 0) that gives count for the positive, negative and neutral sentiment. Later, we calculated the percentage of each sentiment. This step was done for all three attributes: video transcript, video title, video description (See Appendix 13, 14, 15, 16, 17, 18). The model result is discussed in the next section that gives overall sentiment analysis for the news channel.

Results of Sentiment Analysis

We will then evaluate various aspects of sentiment analysis based on the sentiment score for three key attributes: video title, video description, and transcript.



The bar graph above shows FOX sentiment scores ranging from negative tone (-1), neutral tone (0), and positive tone (1) in each category. This sentiment score on Fox channel indicates that both video title and video description are mostly neutral. This reflects that most of the video title and description from FOX are for informative purposes with less strong and emotional words. On the other hand, there is almost 131.18% more negative sentiment than positive in the video title. However, when analyzing the transcript of FOX videos, the result shows that it contains overly positive sentiment scores.



In comparison with MSNBC channel, the bar chart above displays the sentiment score for its published YouTube videos. The result clearly shows the different proportion of sentiment score for MSNBC when compared to FOX channel. Beginning with video title and description, both negative and neutral sentiment are relatively on the same level for MSNBC, whereas FOX dominates with neutral sentiment. In fact, this implies a higher total percentage of negative sentiment for these categories. Based on the sentiment score, it seems like MSNBC expressed dissents and more negative sentiment in the video titles and description toward gender and racial

matters. Based on the video transcript, the sentiment for the actual content of the video from MSNBC channel is leaning toward a positive sense.

Text Translation & Speech Synthesis (Google Trans & Google Text-to-Speech)

The growth of the translation industry has been increasing in the current information age. Therefore, translation is essential for effective communication between different languages and cultures. Due to an increasing demand for in non-English languages, some companies can achieve international reach. In this section, we demonstrated how to translate an original text of a video title in English to another language, and then generate speech with appropriate accent.

```
video_title_txt
```

```
'Reid: Black Or Brown People Acutely Feel The Danger Of Donald Trump | MSNBC'
```

Firstly, the goal is to translate a sample of the video title above to both Thai and Hindi.

```
from googletrans import Translator
translator = Translator()
#translate to Thai (TH)
thai_trans = translator.translate(test, dest='th')

#extract translated text
mytext_th = thai_trans.text
print("Orginal Video Title (EN) ", video_title_txt)
print("Translated Video Title (TH) ",mytext_th)
```

```
Orginal Video Title (EN)  Reid: Black Or Brown People Acutely Feel The Danger Of Donald Trump | MSNBC
Translated Video Title (TH)  Reid: คนผิวดำหรือน้ำตาลรู้สึกถึงอันตรายของ Donald Trump อย่างรุนแรง | MSNBC
```

The result above shows before and after the translation using translator() function from Googletrans library. The translation is relatively accurate.

```
hindi_trans = translator.translate(test, dest='hi')
mytext_hi = hindi_trans.text
print("Orginal Video Title (EN) ", video_title_txt
print("Translated Video Title (HI) ",mytext_hi)
```

```
Orginal Video Title (EN)  Reid: Black Or Brown People Acutely Feel The Danger Of Donald Trump | MSNBC
Translated Video Title (HI)  रीड: काले या भूरे रंग के लोगों ने डोनल्ड ट्रम्प के खतरे को महसूस किया एमएसएनबीसी
```

In addition, we perform the same step translating from English to Hindi as we set the new destination language as `hi`.

```
# Import the required module for text
# to speech conversion
from gtts import gTTS

#for accent
language = 'th'

#convert text to voice
myobj = gTTS(text=mytext_th, lang=language, slow=False)

#export to mp3 voice file
myobj.save("MSBA327_Voice_TH.mp3")
```

```
#for accent
language = 'hi'

#convert text to voice
myobj = gTTS(text=mytext_hi, lang=language, slow=False)

#export to mp3 voice file
myobj.save("MSBA327_Voice_HI.mp3")
```

After retrieving translated text, we generated speech from the machine using Google Text-to-Speech API (GTTS). As shown in Python scripts above, we selected appropriate accents based on the translated language as our text input. Finally, we exported the machine-generated speeches both in Thai and Hindi in mp3 format.

- Machine-generated speech in English: MSBA327_Voice_EN
- Machine-generated speech in Thai: MSBA327_Voice_TH
- Machine-generated speech in Hindi: MSBA327_Voice_HI

Conclusion

In this paper, we focused on gender and racial bias in Fox News and MSNBC. The purpose of this is to understand from a scientific understanding if there is a bias and if it was being depicted in their Youtube videos. We used Python and Youtube API to build a dataset so that we could text representation, classification, clustering, sentiment analysis, and speech synthesis.

The results depicted Fox News were skewed higher compared to MSNBC for posting videos related to the keywords of 'blm', 'women', 'woman', black lives matter'. This paper also discovered MSNBC and Fox News chose to cover or not to cover certain important topics that relate to the images of political figures who happen to be women. Sentiment analysis helps identify differences in how channels treat controversial racial and gender topics. Both news organizations the sentiment for the video title and description were skewed neutral and the transcript was skewed positively. Lastly, this paper covered the importance of inclusion for all free from any potential language barrier an individual may encounter. Future analysis of this should continue in order to make sure the news industries are acting in a fair and unbiased manner.

Limitations and Future improvements

Although we have conducted extensive text analysis on video content on Youtube from various channels, our analysis does have some limitations. Due to a freemium Google account, every time we call Google API, it has a stop-limit for data usage and the amount of queries we can acquire daily. In addition, since we needed to analyze videos with transcript enabled, we ran

into small sample sizes for some TV channels, such as CNN and CBS. This limitation might happen due to their data policy on published videos.

In the future, it would be interesting to extend the study to other news channels, for example: Bloomberg, ABC News, or other local channels. With a Google premium account, we will gain the ability to retrieve more data with more parameters such as user comments for certain videos on YouTube.

References

AJMC staff (2020, July 3). *A Timeline of COVID-19 Developments in the First Half of 2020*. Retrieved from https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

Cherry. (2020). *How Does Implicit Bias Influence Behavior?* Retrieved from https://www.verywellmind.com/implicit-bias-overview-4178401

Diversity. (n.d.). *What is Gender Bias ?* Retrieved from https://www.diversity.com/page/What-is-Gender-Bias

Folkenflik, D. (2018). *Fox News Pays $10 Million To Settle Racial, Gender Bias Suits.* Retrieved from https://www.npr.org/sections/thetwo-way/2018/05/16/611504340/fox-news-pays-10-million-to-settle-racial-gender-bias-suits

Frontiers (2019). *Race and Gender Bias in the Research Community on African Lions.* Retrieved from https://www.frontiersin.org/articles/10.3389/fevo.2019.00024/full

Harris, J.D. (2015). *NYC media coverage of black suspects is way out of proportion with black arrest rates.* Retrieved from https://www.vox.com/2015/3/26/8296091/media-bias-race-crime

Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

IBM. (n.d.) *IBM SPSS Modeler Text Analytics 18.2.1 User's Guide.* Retrieved from https://elearning.ggu.edu/mod/resource/view.php?id=1611569

Kulaszewicz, K.E. (2015). *Racism and the Media: A Textual Analysis.* Retrieved from

https://sophia.stkate.edu/cgi/viewcontent.cgi?article=1478&context=msw_papers

Mohsin, M. (2020). 10 Youtube Stats Every Marketer Should Know in 2020 [Infographic]. Retrieved from https://www.oberlo.com/blog/youtube-statistics

NumPy. (2020). *NumPy*. Retrieved from https://numpy.org/doc/stable/user/whatisnumpy.html

Pandas.(2020). *Pandas*. Retrieved from https://pandas.pydata.org/

Pprint. (2020). *pprint*. Retrieved from https://docs.python.org/2/library/pprint.html

Sentiment Trading. (2020). *VADER Sentiment Analysis in Algorithmic Trading.* Retrieved from https://blog.quantinsti.com/vader-sentiment/#:~:text=Compound%20VADER%20scores%20for%20analyzing,1%20(most%20extreme%20positive).

YouTube Data API. (2020). *YouTube Data API.* Retrieved from https://pypi.org/project/youtube-data-api/

YouTube Data API. (n.d.) *YouTube Data API.* Retrieved from https://developers.google.com/youtube/v3/docs/search

youtube-transcript-api 0.3.1. (2020). *youtube-transcript-api 0.3.1.* Retrieved from https://pypi.org/project/youtube-transcript-api/

Appendix

Appendix 1: Installation of clients and api.

```
#Install google api python client, youtube-data-api, youtube_transcript_api
pip install --upgrade google-api-python-client
pip install youtube-data-api
pip install youtube_transcript_api
```

Appendix 2: Import necessary libraries

```
[10] import os
     from pprint import pprint
     import numpy as np
     import pandas as pd
     from youtube_api import YouTubeDataAPI
     from youtube_transcript_api import YouTubeTranscriptApi
     import nltk
     nltk.download('punkt')
     from nltk.tokenize import sent_tokenize
     from nltk.tokenize import word_tokenize
     import nltk
     nltk.download('vader_lexicon')
     from nltk.probability import FreqDist
     from nltk.corpus import stopwords
     nltk.download('stopwords')
     from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
     import matplotlib.pyplot as plt
```

Appendix 3: API Key

```
[11] # Key to connect to YoutubeAPI
     api_key =('AIzaSyCvPvXeOJZZE4LuhHazUTHUm8UaOlwZahM')
     yt = YouTubeDataAPI(api_key)
```

Appendix 4: Retrieve Video attributes using YouTube Data API.

```
[12] #Fox
     y_terms= ['blm','women','woman','black lives matter']
     fox_searches = yt.search(q=y_terms, channel_id='UCXIJgqnII2ZOINSWNOGFThA',
                              max_results=1000, order_by = 'relevance',
                              published_after= 1577935688.0,
                              published_before= 1591064888.0,
                              videoCaption ='closedCaption')
     print(fox_searches[0])
```

```
{'video_id': 'gZY0UsddDss', 'channel_title': 'Fox News', 'channel_id': 'UCX
```

Appendix 5: Convert JSON file into dataframe and merge all the dataframe into one dataframe.

```
[17] #turn JSON into DF
     fox_search = pd.DataFrame(fox_searches)
     cnn_search = pd.DataFrame(cnn_searches)
     msnbc_search = pd.DataFrame(msnbc_searches)
     cbs_search = pd.DataFrame(cbs_searches)

     #Concatenating
     df_search = [fox_search,cnn_search, msnbc_search, cbs_search]
     result = pd.concat(df_search)
     result
```

Appendix 6: Retrieve video transcript along with video id

```
new_result=[]
text_result=''
for video_id in result['video_id']:
  trans_results = []
  try:
      trans = YouTubeTranscriptApi.get_transcript(video_id)
      for script in trans:
        for key, value in script.items():
          if key=='text':
            trans_results.append(value)
          else:
            continue

      print(video_id)
      print(trans_results)

      text_result=','.join(trans_results)
      new_result.append([video_id,text_result])

  except Exception:
        continue

final = np.array(new_result)
final_df = pd.DataFrame(final, columns=['video_id','new_result'])
```

Appendix 7: Merge the dataframe to get one final dataframe consists of all attributes

```
[21] new_merge = pd.merge(final_df,result, on ='video_id')
```

Appendix 8: Total count of Records per channel

```
[112] #Total count of records per channel: indicating video with transcript
      print("Fox Records: ", len(fox_append))
      print("Msnbc Records: ", len(msnbc_append))
      print("Cnn Records: ", len(cnn_append))
      print("Cbs Records: ", len(cbs_append))
```

```
Fox Records:  544
Msnbc Records:  247
Cnn Records:  15
Cbs Records:  0
```

Appendix 9: Analyze sentiment score (new_result is replaced by attributes: transcript, title and description to get the score of all three attributes).

```
#new_result sentiment
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA

sia = SIA()
sent_results = []

for line in trans_list:
    pol_score = sia.polarity_scores(line)
    pol_score['new_result'] = line
    sent_results.append(pol_score)

pprint(sent_results[:5], width=100)
```

Appendix 10: All sentiment scores of Video transcript

```
#determine polarity
trans_score['trans_label'] = 0
trans_score.loc[trans_score['compound'] > 0.2, 'trans_label'] = 1
trans_score.loc[trans_score['compound'] < -0.2, 'trans_label'] = -1
trans_score.head()
```

	neg	neu	pos	compound	new_result	trans_label
0	0.156	0.803	0.041	-0.9995	>> Laura: LAST NIGHT I SHOWED,YOU HOW THE LEFT...	-1
1	0.095	0.797	0.108	0.7201	ACTIVELY MONITORING THE,INVESTIGATION IN THIS ...	1
2	0.089	0.841	0.069	-0.9792	HAVE A REPEAT OF LAST NIGHT.,WE WILL SEND IT B...	-1

Appendix 11: All sentiment scores of video title

```
#determine polarity title
title_score['title_label'] = 0
title_score.loc[title_score['compound'] > 0.2, 'title_label'] = 1
title_score.loc[title_score['compound'] < -0.2, 'title_label'] = -1
title_score.head()
```

	neg	neu	pos	compound	video_title	title_label
0	0.250	0.750	0.000	-0.4588	Candace Owens: Victimhood has become a mental ...	-1
1	0.328	0.672	0.000	-0.5994	Trey Gowdy on George Floyd's death: 'I...	-1
2	0.000	0.635	0.365	0.3182	Tucker: Media fan racial flames	1

Appendix 12: All sentiment scores of video description

```
#determine polarity description
description_score['description_label'] = 0
description_score.loc[description_score['compound'] > 0.2, 'description_label'] = 1
description_score.loc[description_score['compound'] < -0.2, 'description_label'] = -1
description_score.head()
```

	neg	neu	pos	compound	video_description	description_label
0	0.000	1.000	0.000	0.0000	Blexit movement founder Candace Owens and form...	0
1	0.160	0.840	0.000	-0.6360	DOJ launches probe into death of George Floyd;...	-1
2	0.182	0.714	0.104	-0.2942	Things are falling apart in Minneapolis and, a...	-1

Appendix 13: Count of positive, negative, and neutral sentiment score for video transcript

```
#value count transcript
merging_trans=final.groupby('channel_title')
merging_trans.trans_label.value_counts()
```

```
channel_title  trans_label
CNN             1              4
Fox News        1            344
               -1             31
MSNBC           1             97
               -1             49
Name: trans_label, dtype: int64
```

Appendix 14: Percentage of positive, negative, and neutral sentiment for video transcript

```
#value count transcript
merging_trans=final.groupby('channel_title')
merging_trans.trans_label.value_counts(normalize=True) * 100
```

```
channel_title  trans_label
CNN             1            100.000000
Fox News        1             91.733333
               -1              8.266667
MSNBC           1             66.438356
               -1             33.561644
Name: trans_label, dtype: float64
```

Appendix 15: Count of positive, negative, and neutral sentiment score for video title

```
#value count title
merging_title=final.groupby('channel_title')
merging_title.title_label.value_counts()
```

```
channel_title  title_label
CNN             0              3
               -1              1
Fox News        0            303
               -1             49
                1             23
MSNBC          -1             58
                0             57
                1             31
Name: title_label, dtype: int64
```

Appendix 16: Percentage of positive, negative, and neutral sentiment for video title

```
#value count title
merging_title=final.groupby('channel_title')
merging_title.title_label.value_counts(normalize=True) * 100
```

```
channel_title  title_label
CNN             0            75.000000
               -1            25.000000
Fox News        0            80.800000
               -1            13.066667
                1             6.133333
MSNBC          -1            39.726027
                0            39.041096
                1            21.232877
Name: title_label, dtype: float64
```

Appendix 17: Count of positive, negative, and neutral sentiment score for video description

```
#value count description
merging_descrip=final.groupby('channel_title')
merging_descrip.description_label.value_counts()
```

```
channel_title  description_label
CNN            -1                    3
                0                    1
Fox News        0                  311
               -1                   34
                1                   30
MSNBC          -1                   57
                0                   48
                1                   41
Name: description_label, dtype: int64
```

Appendix 18: Percentage of positive, negative, and neutral sentiment for video description

```
#value count description
merging_descrip=final.groupby('channel_title')
merging_descrip.description_label.value_counts(normalize=True) * 100
```

```
channel_title  description_label
CNN            -1                   75.000000
                0                   25.000000
Fox News        0                   82.933333
               -1                    9.066667
                1                    8.000000
MSNBC          -1                   39.041096
                0                   32.876712
                1                   28.082192
Name: description_label, dtype: float64
```