COMM 281/CS 206: Exploring Computational Journalism
Cable News Archive Analysis

Geraldine Moriba
Charlie Jarvis
Flor Coelho
Theodora Boulouta

Trust in traditional news outlets is at an all-time low, and the reputations of those institutions are under scrutiny. Cable news channels - CNN, FOX News, and MSNBC – have been criticized for operating through the prism of bias. If this is the case, are cable news programmers contributing to the rise in polarization in the U.S.? The goal of this particular project was to conduct an exploration of whether an artificial intelligence (AI) tool could be used to effectively analyze cable television news for patterns and trends in content, bias and coverage.

The Financial Times has developed a bot that automatically flags whether their articles quote too many men. The BBC has started measuring how many female experts get on the air. These are initiatives aiming to balance gender representation in the news by training journalists to consciously include more women in their stories. What are other ways to automate reporting that is less biased? What other biases should be analyzed? Innovators throughout the news industry are currently exploring ways to use artificial intelligence to improve newsgathering, news production and the distribution processes. Can AI also be used to monitor and analyze the biases in news coverage?

Our goal was to determine whether it was possible to create a tool with the power to do these three things: identify coverage patterns, analyze bias in a way that is both efficient on large data sources, and make it scalable across a variety of topics and parameters. To answer these challenges we started by developing an algorithm to identify instances in which cable news channels address a specific coverage using video clips and audio transcriptions.

Phase I: Generate data to compare topic coverage across three cable news channels

The first stage was to identify a topic and then develop an AI tool to identify coverage of that topic on cable news.

We created a data set using cable news video (Fox News, MSNBC and CNN) and corresponding transcripts from TV News Archive for the four months leading up to the 2016 and 2018 elections.

To identify our first research topic we used polling studies of voters affiliated with both American political parties during the 2016 and 2018 national elections from the PEW Research Center, a nonpartisan fact tank. Immigration was a topic that dominated headlines in both elections.

> Top Election Topics 2016: economy, terrorism, foreign policy, health care, gun policy, **immigration**, social security, education, supreme court appointments, and treatment of racial and ethnic minorities.
>
> Top Election Topics 2018: supreme court appointments, health care, economy, gun policy, Medicare, social security, taxes, **immigration**, treatment of racial and ethnic minorities, and environment.

Our next task was writing a topic modelling algorithm. Our topic modeling algorithm can be broken up into two parts. The first is topic lexicon generation and the second is topic identification.

A topic lexicon represents a list of words frequently co-occurring with the keyword of interest, in this case “immigration.” We can classify transcript segments as being about immigration when enough these “signal words” appear. For example, seeing the words “reform” and “illegals” mentioned in close proximity can indicate that immigration is being discussed, even if the word “immigration” is never explicitly mentioned.

To generate a lexicon for a given topic we lemmatized our transcripts and then tokenized the contents, in order to look at words individually. Lemmatization refers to the process of removing inflectional endings and returning the base form of a word, which is known as the lemma. This facilitated counting the occurrence of “reform” and “reforms”, for example, as identical[1]. We then identified the most frequently co-occurring words with our topic word in the lemmatized transcripts and inserted them in to a ranked lexicon[2].

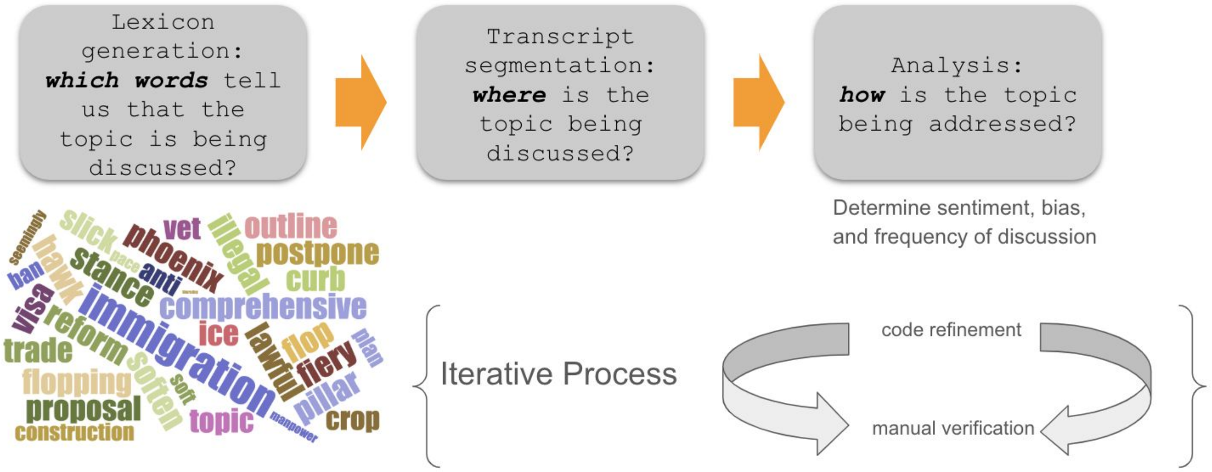
The second part of topic identification is a scoring technique that allowed us to take an excerpt of text and assign it a score based on the number of words in the lexicon that appeared in the segment. The appearance of a word that was ranked more highly in our lexicon had a greater score contribution than a word that appeared at a lower rank in the lexicon. Using this scoring technique, we broke the transcript into small overlapping segments. Next, we iterated through the segments in the transcript and assigned each segment a score. If the score exceeded a certain threshold, the segment classified as talking about our topic. Combining overlapping

[1] We used the nlp library ‘spacy’ https://spacy.io/ for lemmatization

[2] Words were classified as co-occurring with a target word based on a “mutual information” score. A co-occurrence was defined as a word occurring within a certain number of words of an instance of the topic word. The mutual information score was computed as such: (number of co-occurrences of word with topic word)/(total number of occurrences of word in transcript* total number of occurrences of topic word in transcript). This allowed for frequently occuring words such as “and” to be filtered out of our lexicon

segments, we effectively a comprehensive list of transcript excerpts of various lengths that addressed our target topic.

This was an iterative process of code refinement and manual verification. This was especially true when choosing the points that each lexicon word should contribute, and the corresponding score threshold for classifying an excerpt as “about immigration.” Manual readings of transcripts were compared to computer-generated outputs, until we identified values that allowed us to precisely identify relevant transcript portions.



Lexicon generation: ***which words*** tell us that the topic is being discussed?

Transcript segmentation: ***where*** is the topic being discussed?

Analysis: ***how*** is the topic being addressed?

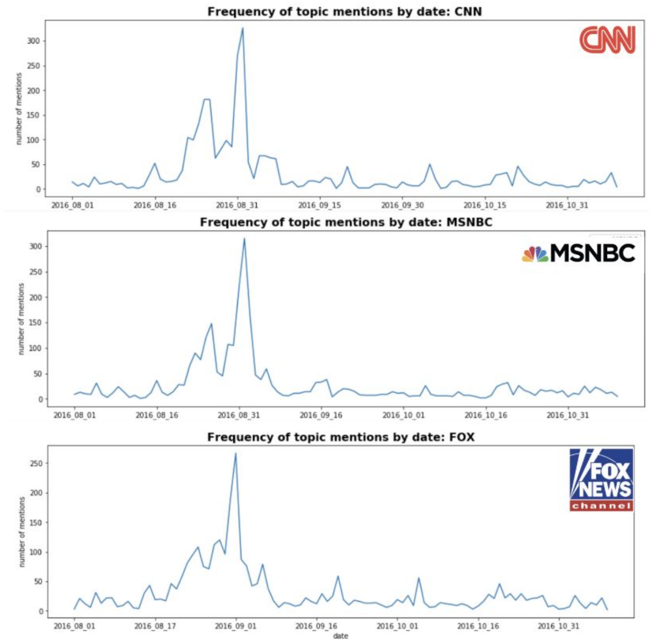
Determine sentiment, bias, and frequency of discussion

Iterative Process

code refinement

manual verification

To test our methodology on a larger data scale we started by looking at when immigration coverage spiked in 2016 on the three cable news networks.

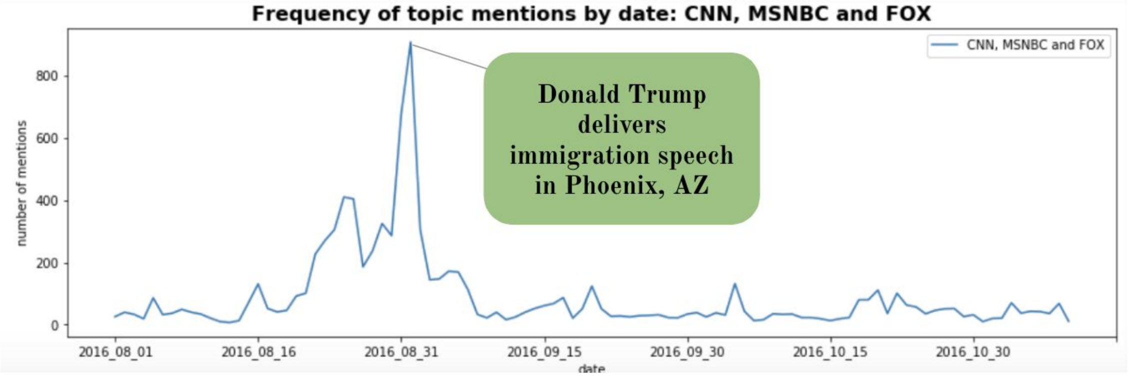


Top Immigration Coverage Dates (2016)

FOX:	CNN:	MSNBC:
2016_09_01 : 267	2016_09_01 : 326	2016_09_01 : 315
2016_08_31 : 189	2016_08_31 : 268	2016_08_31 : 218
2016_08_29 : 120	2016_08_25 : 181	2016_09_02 : 165
2016_08_28 : 112	2016_08_26 : 181	2016_08_26 : 148
2016_08_25 : 108	2016_08_24 : 133	2016_08_25 : 121

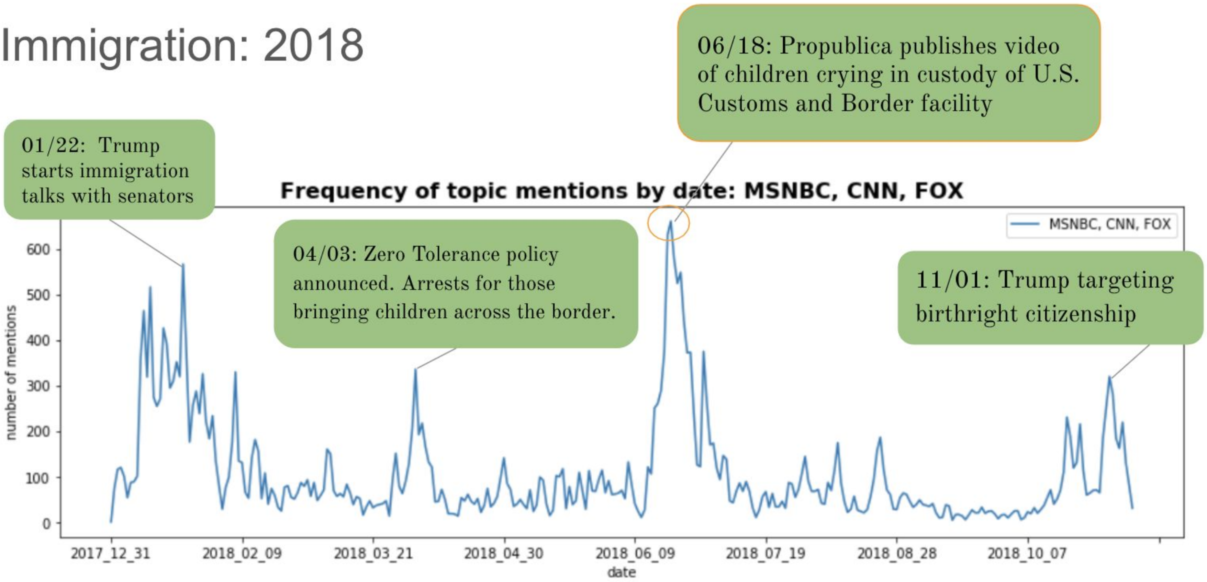
Were there any notable differences between the coverage of immigration as a news topic in 2016? No, the pattern of coverage is relatively consistent between each cable news channel. More importantly for the purposes of this study, the computational findings match the events of the day.

Immigration: 2016



What about 2018? When did immigration receive the most coverage? Here the results also demonstrate that the computational findings match the events of the day.

Immigration: 2018



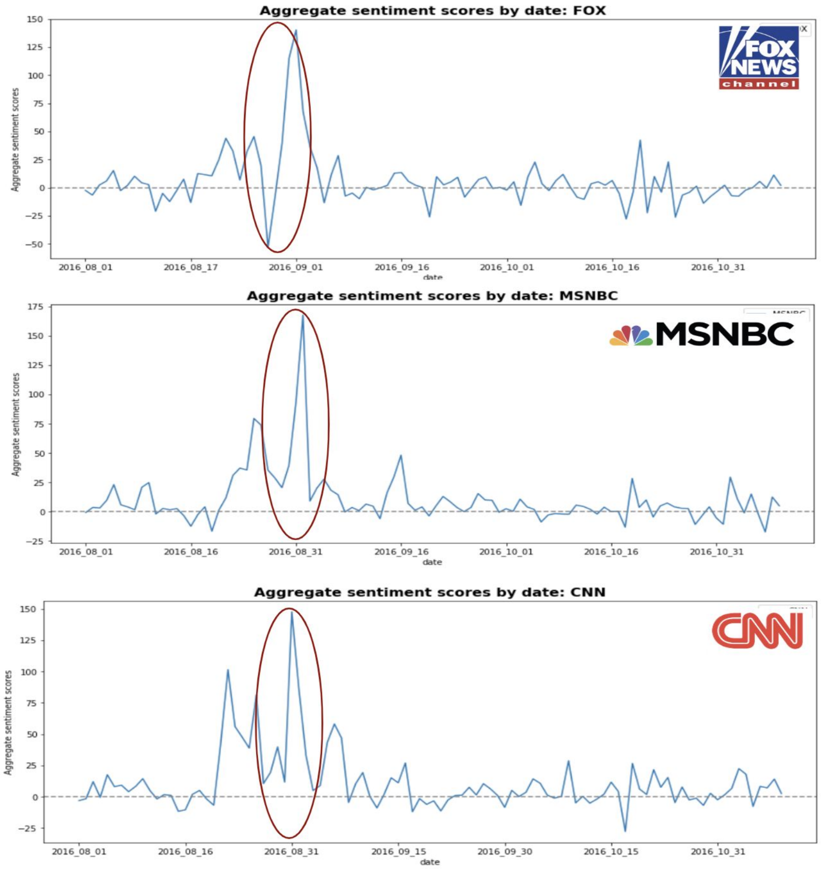
The results of the coverage patterns and their precise agreement with coinciding historical events demonstrates that this methodology works. To test it further we ran the same test on other election topics. Those results were also accurate.

Phase II: Analyze and compare coverage of immigration as a topic

In Phase II, our goal was to look at news coverage in a manner that was more nuanced than simply computing the frequency of coverage of a particular topic. Here, we sought to identify *how* the topic was being covered. In other words, we examined the way that language was used to convey emotion and editorial bias.

Sentiment Analysis

This technique was used to measure the quantity and degree of positive or negative the words used during discussions of the topic, in this case, immigration. Sentiment analysis relies upon determining the emotional tone of each word, and assigning a corresponding value, where positive language receives positive scores and negative language receives a negative score. Sentiment analysis is designed to provide an understanding of the attitudes and emotions expressed.



Sentiment scores peaked on August 31 and September 1, 2016.

The peaks occurred the day before and the day of presidential candidate Donald Trump campaign speech on immigration. As a partisan cable news network that is demonstrably right-leaning, it is predictable that Fox would have positive sentiment scores. The results for MSNBC and CNN were more surprising. These two cable news networks had positive scores as well, even though their reporting on the topic had a negative tone on these dates.

Here is one example of a transcript chunk from that two-day period.

> “UNTIL TRUMP ENDS THE EARLY TRANSITION OF AMERICA FROM THE **GREATEST** NATION IN HISTORY INTO SOME PATHETIC THIRD RATER ALSO-RAN, MULTICULTURAL MESS, UNTIL BLEEDING HAS STOPPED, THERE IS NOTHING TRUMP CAN DO THAT WON'T BE **FORGIVEN**. EXCEPT CHANGE HIS IMMIGRATION POLICIES. AT LEAST ONE PERSON FOUND THE TIMING OF YOUR BOOK RELEASE SO **FUNNY**. I'VE WATCHED THIS 20 TIMES. YOU HAVE TO WATCH WITH IT ME NOW.
>
> WHO KNEW THAT IT WOULD BE DONALD TRUMP TO CONVERT THE GOP BASE TO **SUPPORTING** AMNESTY ON THE SAME WEEK ANN COULTER'S BOOK --
>
> HE EITHER FEELS SORRY FOR YOU OR HE'S **LAUGHING** AT YOU.”
>
> **Sentiment score: 1.1568**
>
> MSNBC 08/31/2016

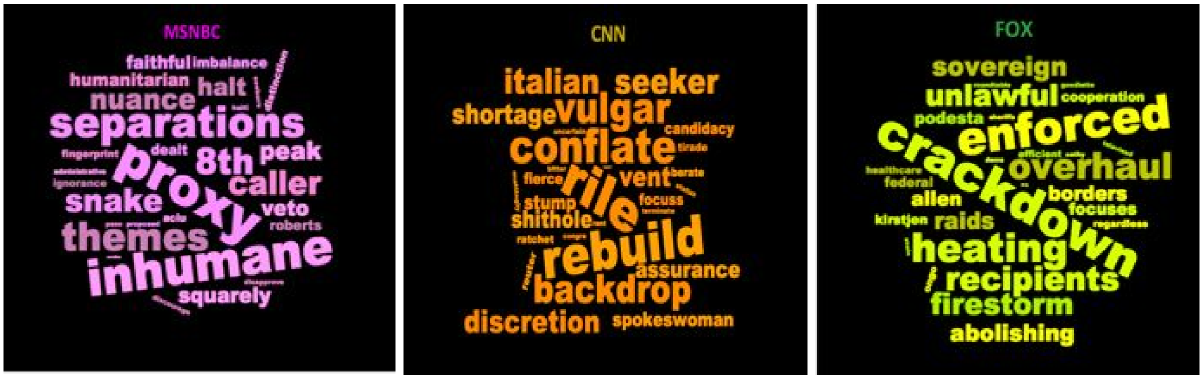
In this clip, there are mixed linguistic messages. The tone and message is negative, but five positive words (highlighted in bold) cause an overall high positive sentiment score. We repeatedly found that positive sentiment scores accurately correlate with high emotion moments, but are less effective as predictors of whether those moments have a positive or negative tone or meaning.

Language is Nuanced

Our next objective was to find a more nuanced way to detect and quantify bias for purposes of analysis. Bias, like other nuanced semantic qualities, is notoriously difficult for computers to recognize. The challenge then is to find a way to account for linguistic nuance. In the next phase of our analysis, we attempted to quantify subtle and not so subtle biases reflected in tone and meaning.

We developed an algorithm to iterate through sections on a given topic to compute a bias score of how many “biased words” appeared in the topic sections. This method for detecting bias also used a system of lexicons. First, we manually created a database of written words from print news and national organizations. This lexicon had limited success with identifying relevant chucks from the transcripts because what we write is different than what we say. Written words are often different than spoken words, and consequently they are not necessarily used on-air. To capture the tone and attitude of language used on television cable news specifically, we also needed to identify the high inference words used most frequently with immigration by CNN, Fox News and MSNBC. We wrote an additional algorithm to see which words appear exclusively in the lexicon of each channel, to get a better understanding of the differences in the language used between channels.

Television Cable News High Inference Language Lexicon Word Clouds



These word clouds reflect some the most frequent high inference words used with immigration exclusively by each cable news network in 2016.

Using these two combined lexicons found on the topic of immigration we created a left-biased and a right-based lexicon.

At this point it is important to note that we wrote a bias classifier just for the topic of immigration. Writing a bias classifier for a different topic would need human input for lexicon generations[3]. Also, this bias analysis was constrained to political bias.

Bias Analysis Using Two High Inference Language Lexicons

We created two lexicons to identify coded and biased language. As noted previously, our first lexicon relied on written sources. Stark language usage differences caused this technique to be limited in analyzing transcripts. We solved for this by combining manually identified high-bias words from news articles about immigration and combining them with words that are used most frequently by CNN, FOX News and MSNBC near the word immigration. Here's a sample of words starting with the letter "A."

[3] A potential way to decrease the human input would be by creating databases of left leaning and right leaning tv shows and looking at the words used by each show. There would still need to be human input for picking the shows.

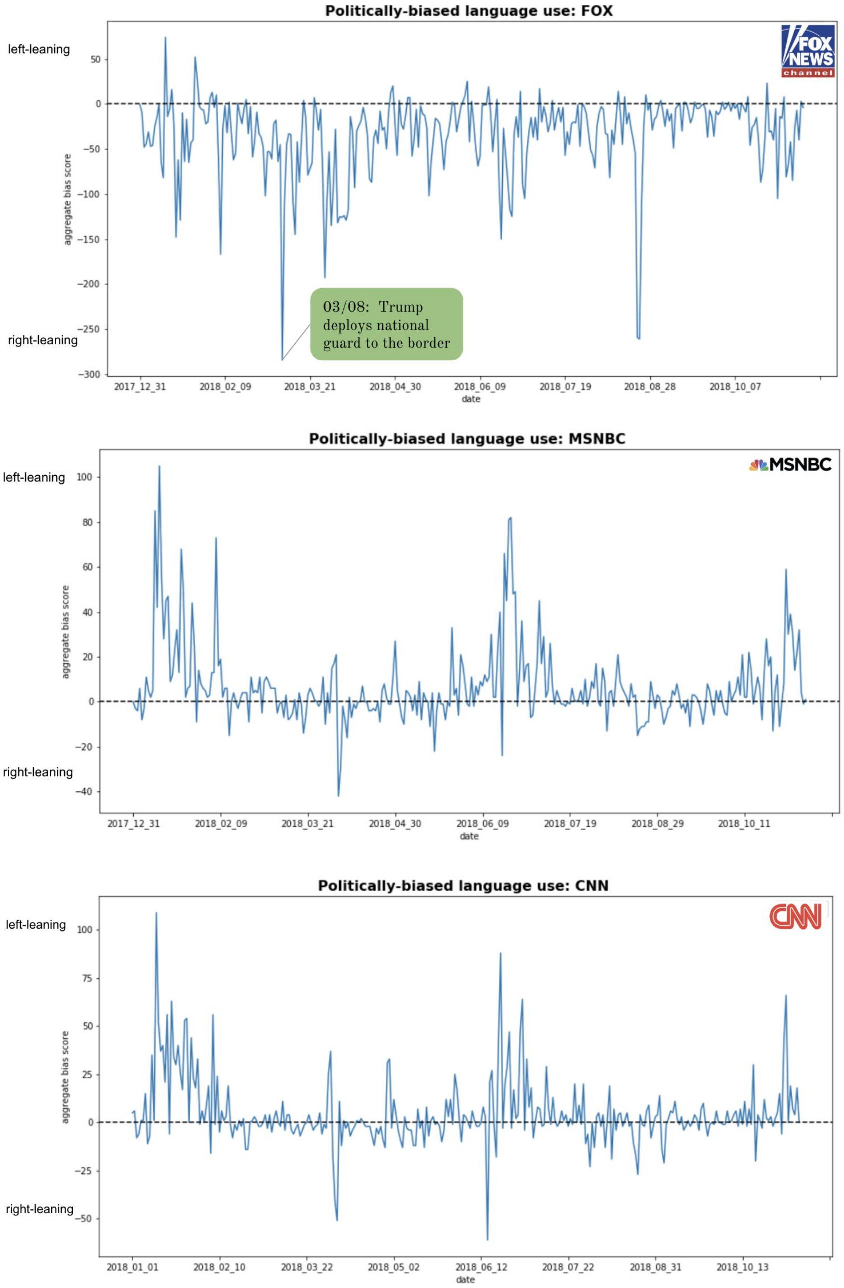
Right biased wording when used with immigration	**Left biased wording when used with immigration**
Abused	Abiding
Acquitted	Accountability
Aggressive	Activate
Alien	Advisors
Ambush	Advocacy
America first	Aggressive
America great	Alienate
America safe	Alt-right
American lives	Amnesty
American worker	Anthem
Anchor baby	Anti-Hispanic
Apprehend	Anti-immigrant
Apprehended	Anti-Latino
Architect	Anti-Muslim
Arrest	Anxiety
Assaulting	Assurance
Avalanche	Asylum seeker

A bias score was generated by measuring the usage of these high inference language words. An instance of a "left-leaning" biased word would push the score in one direction, and an instance of a "right-leaning" biased word would push the score in the opposite direction.

Here are two examples of the results of identifying bias based on the high inference language lexicon:

> **Right Biased Wording When Used With Immigration**
> THE PRIORITIES IS TO **GET RID** OF **CRIMINAL FELONS**, **ILLEGAL ALIENS** WHO HAVE **COMMITTED CRIME** INSIDE THE COUNTRY. NUMBER TWO **ENFORCE** THE IMMIGRATION LAW, AND NUMBER THREE TO **PROTECT THE BORDER**. THAT'S CONSISTENT.
> Fox News Shepard Smith 08/26/2016
>
> **Left Biased Wording When Used With Immigration**
> NOW THE MAN WHO LAUNCHED HIS CAMPAIGN, **SLURRING** MEXICAN IMMIGRANTS WHO CALLED FOR A WALL ON OUR SOUTHERN BORDER, WHICH MEXICO WILL PAY FOR, AND DEPICTS **UNDOCUMENTED** IMMIGRANTS IN THIS COUNTRY AS A **LAWLESS HOARD**, **TERRORIZING CITIZENS**, HE APPEARS TO EMBRACE **ELITE REPUBLICAN ORTHODOXY** AND OBAMA POLICY ADMINISTRATION, WHETHER HE KNOWS IT OR NOT.
> MSNBC, All In With Chris Hayes 08/24/2016

Bias scores for each channel



Right and left bias scores using two high inference language lexicons

The results of running our scoring algorithm on an extended database of transcripts indicate that the bias lexicons work. Fox was correctly identified as largely right-leaning, while CNN and MSNBC were accurately identified as more left-leaning. Our tool enables powerful opportunities for bias-identification. Language bias is a hard problem, as it represents a subjective task and linguistic cues are often subtle. Not only that, bias can often only be determined only through context. There are many opportunities for further analysis and expansion in this realm.

What we learned

What Works

- Extracting information and identifying topic segments based on word frequencies
- Identifying high emotion moments using sentiment analysis
- Using bias analysis to determine trends in political-leanings
- Computer-aided comparisons of coverage patterns of different topics

The Challenges

- Sentiment analysis is not an effective indicator of positive or negative tone, due to limitations in contextual analysis
- Manually-generated lexicons are necessary for more accurate bias analysis results, but can often be influenced by the biases of the human programmer
- Bias lexicons are difficult to create since people from opposing perspectives often use the same words
- Lexicons need to be constantly updated with new words and phrases as news events evolve and change
- Lemmatization makes the initial process of lexicon generation slow

Future Opportunities

- Analyze other topic biases including climate change, abortion, voter suppression, criminal justice, etc.
- Analyze more coded biases including gender, race, ethnicity, etc.
- Improve the classification methodology by developing additional automated ways to parse out syntactical and semantic characteristics of potentially biased language
- Develop specific measurements to quantify partiality or impartiality
- Develop a user friendly interface that easily shares results with transcripts and video clips

The supervised classification approach used in this methodology relies on a combination of manually and automated identification of bias words. The lexicon we generated is a dataset of biased words which can be used for further research in detecting language bias. We show that our approach trained with more explicitly biased content is effective with language known to be clearly biased and also where the language biases are subtler.

The goal of this particular project was to conduct an exploration of whether an AI tool could be used to effectively analyze cable television news for patterns and trends in content, bias and coverage. These overall results confirm the perceived ideological leaning of American television cable news networks.