

Tessa Blakenship   
Jaci Willoughby

IST 736 Text Mining Final Project

Winter 2020

**Fake News!**  
Determining Bias in a News Story

Fake News!  
Determining Bias in a News Story

Tessa Blakenship and Jaci Willoughby

# **Introduction**

Have you ever read a news story online and wondered if it came from a credible source?  We live in a technology age – right? It’s so easy now to doctor photos and create fictional news that its’ difficult to determine if a news story is actually “REAL” or if it was created to fit someone’s personal agenda.   We hope to find a way to review news headlines and stories to determine the source.  Knowing who the source of the story is will allow the public to judge for themselves the credibility of the story.

By now we all have heard about “Fake News” – but for our project we need to first define what Fake News looks like. How should we define what fake news is?​

We all agree that Meriam-Webster and Oxford are reliable sources in defining words- correct? Meriam-Webster has taken the stance that they will not define the term. They feel the term is a combination of the word “fake” and “news” and there is no need to define the term as a whole. (Meriam-Webster, n.d.) BUT - the Oxford Dictionary has defined “Fake News” as a noun that means “False information that is broadcast or published as news for fraudulent or politically motivated purposes”. (Lexico by Oxford, n.d.)

​For our purposes we’ll stick with the Oxford definition

For the project we also need to define what a “Credible” source is.  Which is a little bit harder since we all have our own belief systems and our own standards for credibility.  For some, Fox News is considered a credible source.   For others it’s CNN.   I think we all know that the National Enquirer reports more sensational stories compared to the New York Times.  So how are we going to determine if the news stories in our data are credible?  The Poynter Institute gives us these questions to ask when evaluating sources.   The institute specializes in ethics and fact-checking and is considered the world’s most influential school for journalism. (Poynter, n.d.)

According to a New York Times article – “No one believes anything” in America today. Voters are worn out by the plethora of political news. Rather than trying to figure out the validity of the news, they would rather turn the channel to non-political topics. (New York Times, n.d.) With the decline in trust of the media – we as Americans are finding it harder to trust in other areas of our lives. We don’t trust our Government, we have trouble trusting others with different opinions than our own, and we don’t trust the local authorities to make the right decisions during tough circumstances. By finding a system that can identify credible sources from fake sources, we could improve the trust in the media, leading to improved trust in our lives.​ (USA Today, n.d.)

# **Hypothesis**

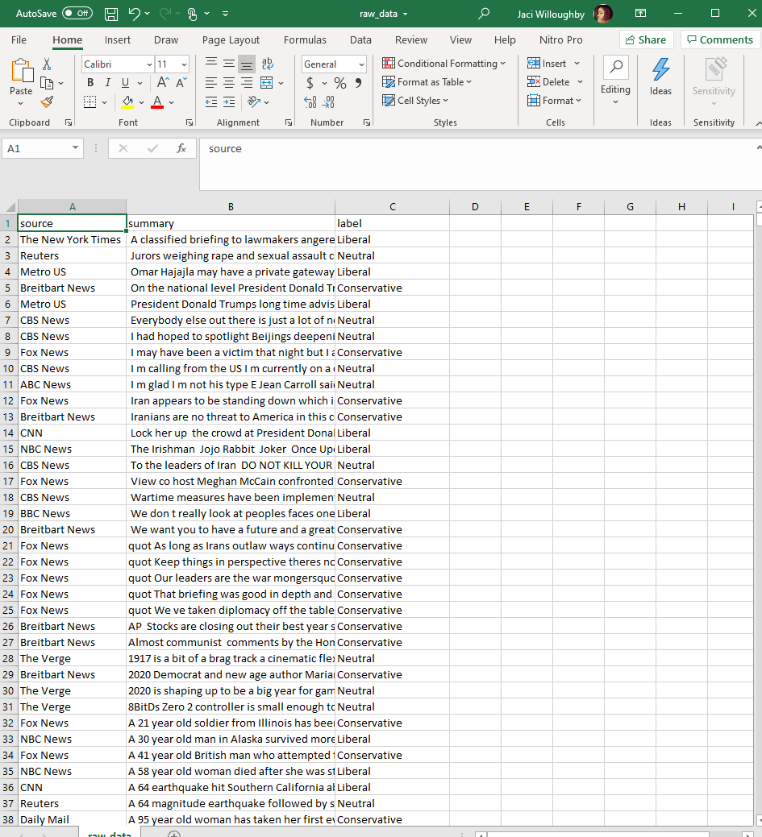
If a news outlet can be predicted based on the content, then the outlet may show bias towards one political view or subject. ​

The data collected spans across different news outlets in the United States, Britain and the Middle East. The information is over the same time period, so that the content is comparable.​

​

**Analysis and Models**

## About the Data – Kaggle News Headline Dataset, Randomly Mined News Summaries and API collected Data



We began the analysis utilizing a News Headlines and Summaries data set from Kaggle. (Murchie, 2020) The data consisted of the following fields:

* News Source
* Author
* Title
* Description
* URL
* API Requested Date
* Published Date
* Content

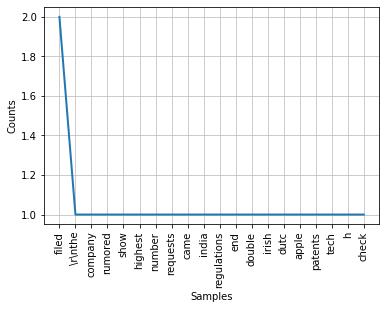
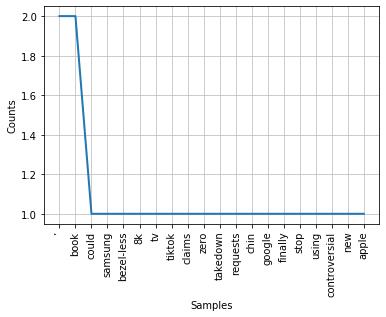
This data set held:

* 3109 news articles
* 311 published between June – August 2019
* 2793 published between October 2019 – January 2020
* 18 News Sources

We then manually mined at random an additional 10 news headlines and summaries from each of the news Sources from February 2019 to included in our Data Set. These were taken from the “Top News” section of each sources website.

Finally, an API was used to mine additional Data to include in the study. (kotartemiy, 2020)

Once the data set was complete, we cleaned it up. First extra columns of information was removed that was not pertinent to proving our hypothesis. Then additional spaces and punctuation marks were removed. The data was reviewed for foreign keys ( ~ , ©, ™, etc…) and these were removed. The data was changed from Sentence Case to all lower case as well.



## Model(s) –

Word Cloud

Word clouds were used to see the frequency of words used in the title and description of the article.

Amazon Mechanical Turk and Kappa

Due to costs we were only able to run a sampling of the Dataset through Amazon Mechanical Turk (AMT). 20 news summaries were selected at random from each Source to create a sample data set. This was uploaded to AMT asking for 5 labels for each news summary. The labels were then added to the sampling data in an Excel Spreadsheet that was used to calculate the Kappa Score for each summary.

Weka

WEKA was utilized to determine classification accuracy. The Filtered Classifier was used to analyze the data. For classification, J48 was first selected as the classifier keeping the default settings, with StringToWordVector selected for the vectorizer, again with default settings intact.

Next, for classification, Naïve Bayes Multinomial was first selected as the classifier keeping the default settings, with StringToWordVector selected for the vectorizer, changing the outputWordCount option to TRUE.

Finally, the classifier was changed to Naïve Bayes Mutlinomial Text with StringToWordVector as above.

Source Distribution – Python

Python was utilized to determine Source distribution – to verify the balance of the dataset.

SentiStrength

SentiStrength (SentiStrength, 2020) is an automatic sentiment analysis tool that is free for academic research. It is available to download for desktop use, or can be used directly via the website to analyze one line of text at a time. The online tool can also detect sentiment in 6 other languages besides English: Finnish, German, Dutch, Spanish, Italian and Russian

The desktop version of the tool was utilized for this project. The Data was converted into a text file in order to process through SentiStrength. Results determined if the sentiment was positive, negative or neutral.

Topic Modeling – Python

Topic modeling was utilized in Python to determine the top 10 topics across all news summaries.

Content K-Means Clusters– Python

The content of the news sources was analyzed with bigram and trigram vectorization. Based on distance, the ideal number of clusters used was 5. The results show that sources that are most closely aligned.

# **Results**

Word Cloud and Topic Modeling

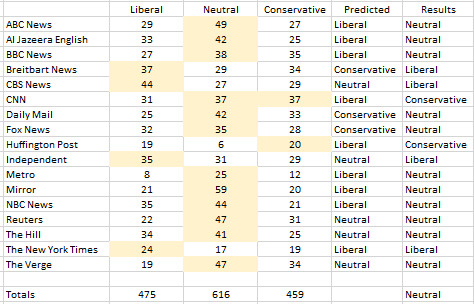
After running the full Data Set through topic modeling, there was no clear difference between the 5 models to indicate bias or sentiment. A decision was made to run each source individually to see if differences would become evident. We then ran the results of each Topic Model through Word Cloud, and the results can be seen below:

|  |  |  |
| --- | --- | --- |
|  | A close up of text on a white background  Description automatically generated | A close up of a piece of paper  Description automatically generated |
| A close up of text on a white background  Description automatically generated | A close up of a piece of paper  Description automatically generated | A close up of a piece of paper  Description automatically generated |
| A close up of text on a white background  Description automatically generated | A close up of text on a white background  Description automatically generated |  |
| A close up of a map  Description automatically generated | A close up of a piece of paper  Description automatically generated | A picture containing text  Description automatically generated |
| A close up of text on a white background  Description automatically generated | A close up of text on a white background  Description automatically generated | A close up of a piece of paper  Description automatically generated |
| A close up of a piece of paper  Description automatically generated | A close up of text on a white background  Description automatically generated |  |

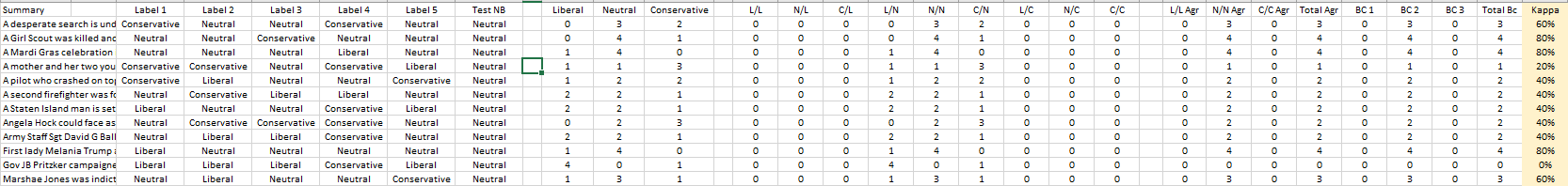
While there’s no significant difference, we do start to see some variance between the sources. Many of the TV stations have topics with Trump, Impeachment, Iran, Iraq, etc … But when you look at The Independence you see that Mum and Man are the top two words. The Verge talks more about Samsung, Tech and Conept. The Huffington Post and NBC have few words that stand out over the others showing that their coverage may be more balanced than the other sources.

Amazon Mechanical Turk and Kappa

Amazon Mechanical Turks were not able to successfully match our predicted labels with our Random Sample from the Data Set. You can see that they’ve labeled the majority of Sources as Neutral. One thing we should consider for future testing was how to reduce personal bias.



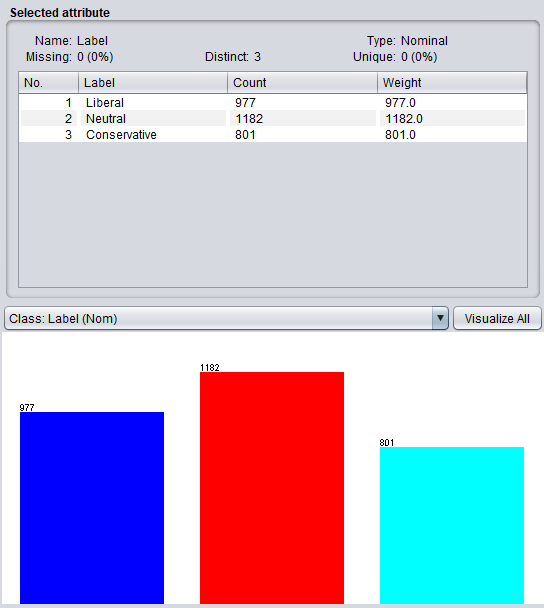
The AMT results were entered into an Excel Spreadsheet to determine the Kappa Scores. The Figures below shows a sampling of the Amazon Mechanical Turk and Kappa results.

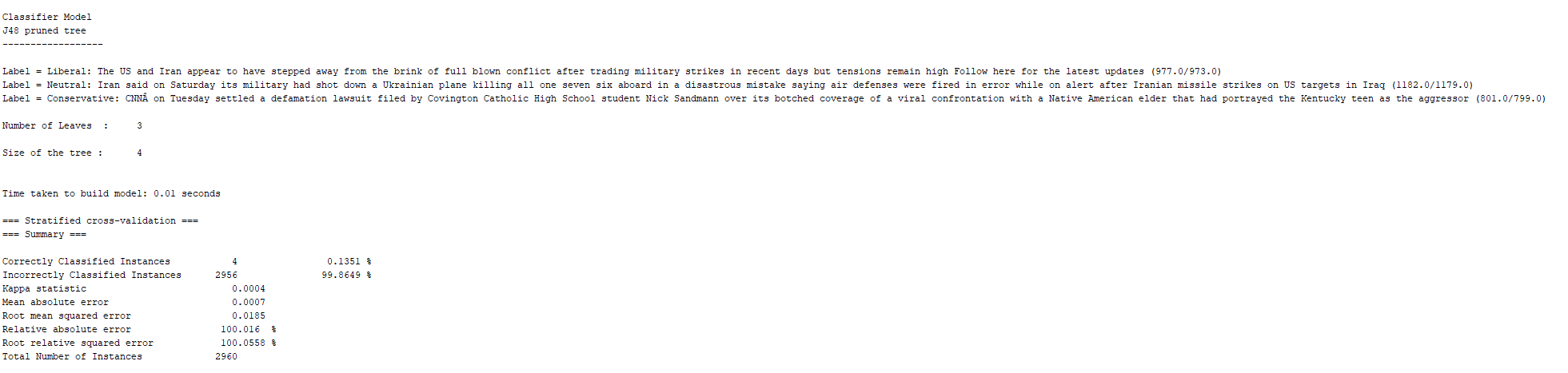


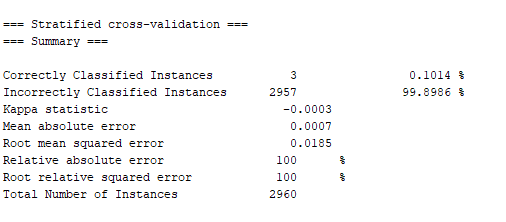
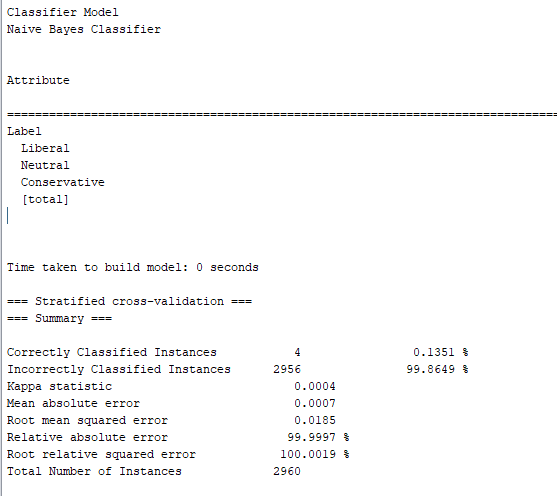
Overall the range of Kappa varied from 0% - 80%. So the agreement is all over the place as well.

Weka

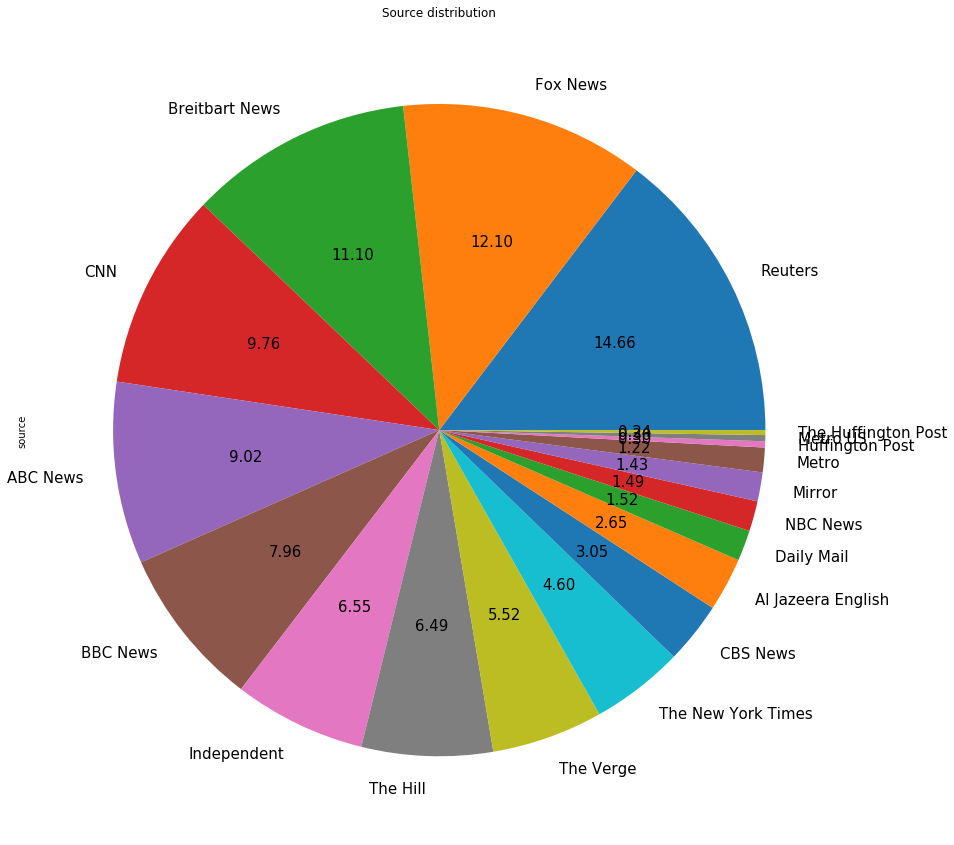
We next ran the full data set to see if we get accurate labels this way – and again the tests came back inconclusive. The computer could not label the news summaries accurately – we had a 99% Incorrect classification on all three types of tests. You can see that our predicted labels are again not as balanced as they should be – and that needs to be considered for future testing. The J48 Pruned Tree, Naïve Bayes and Stratified Cross Validation were unsuccessful in labeling the data correctly – with a 99.8% incorrect classification each.

**





Source Distribution

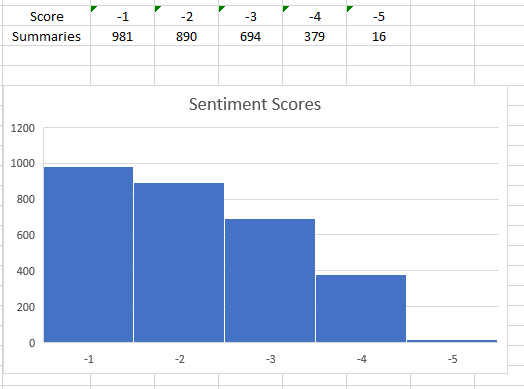


Because we were not having much success in our testing, we decided to see how balanced our Dataset was using Source Distribution. We had 17 Sources and the distribution ranges from 1.12% through 14.66%. This indicates the sources were not as balanced as they should have been – and this might be why there were issues with labeling correctly.

Sentiment Scores

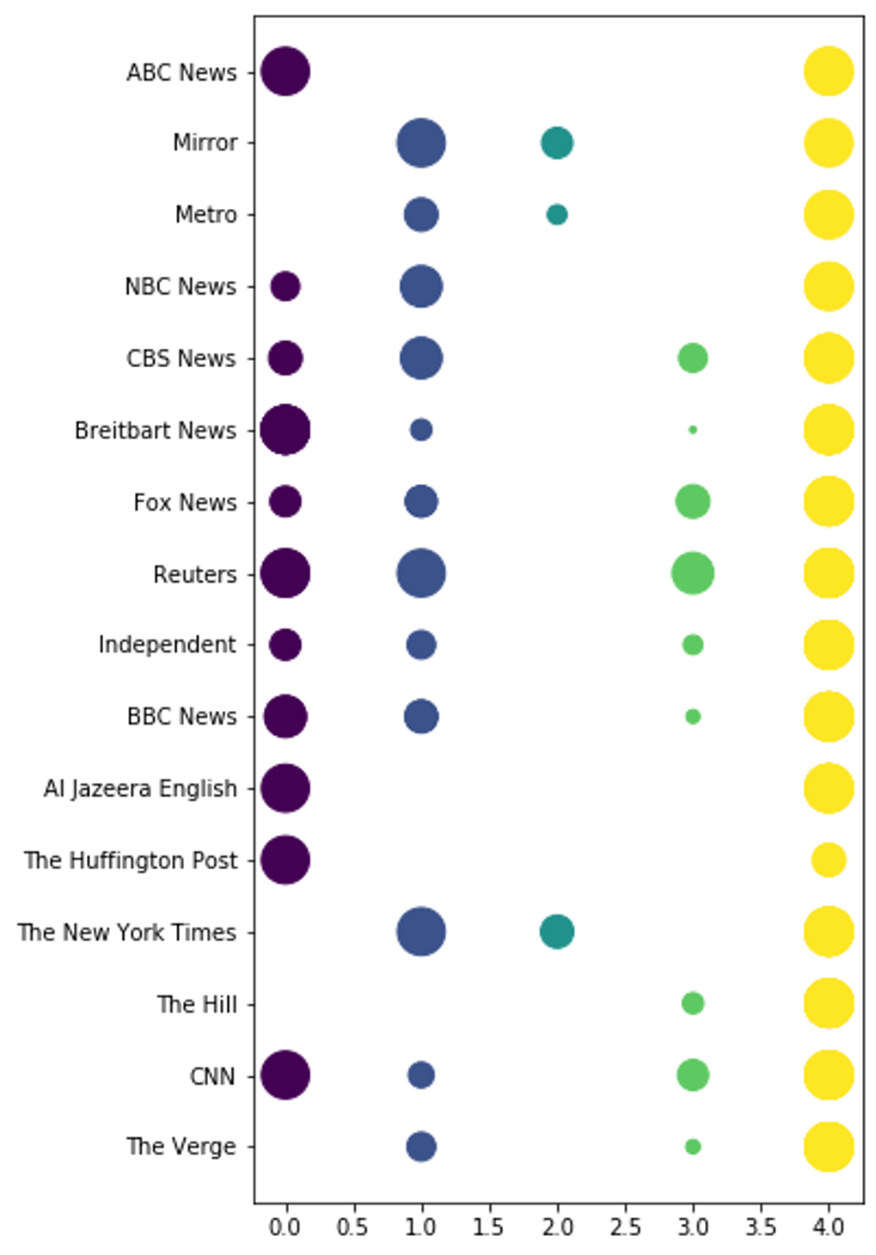
So, if we aren’t able to identify Bias in the news sources, was it possible to predict Sentiment? The answer is YES! And it’s all negative. Running the full data set using SentStrength we were able to see that all our news summaries skewed in the negative range. (You can better see here too that the data was not as balanced as it should be). From all data sources **Sentiment Analysis** came back negative – ranging from -1 to -5.

Here you can see the majority of summaries skewed -1 and -2 … so could be considered closer to neutral in sentiment.

**

K-Means Clusters

The clusters show that there is a base cluster that might just include the stories since they are the same across the time period. Clustering based on a bigram and trigram of the content shows that all of the news sources had the same core information and there were some groups that formed but the majority of the news sources were in multiple clusters. However, cluster 2 includes only three news sources, Mirror, Metro, and The New York Times. These three are all newspapers, but what is interesting is that the Mirror and Metro are both based in the United Kingdom, while The New York Times is located in the United States. We would have expected to see clear clusters.



# **Conclusions**

Topic Modeling may help indicate bias – but more testing is needed. While our results indicate there is a slight difference between the topics reported between sources, a larger sampling needs to be run with more balanced data to make a firm conclusion.

Comparing sentiment based on matching stories may be a better way to find bias. If the same story is positive at one news source, but negative when provided by another source, that might indicate bias. Further analysis would need to be completed to know for sure.

Our results DID indicate that the majority of news stories are negative. This was expected, as the Media tends to focus on what’s wrong in the world with very few “feel good” stories tossed in. We could take a look at the level of negativity in correlation to bias to see if Sentiment can be used to support the notion of bias.

Crowd Sourcing can not be used for classification as personal bias has to be ruled out. AMT does allow additional parameters to be set to help rule out bias – such as selecting only those registered as a Democrat or as a Republican to complete the tasks. Knowing the workers political affiliation ahead of time would allow us to factor in personal bias. Additional funding would be needed to complete these tests.

Finally, clustering indicates that the primary type of delivery may need to be considered when comparing the sources. Three stand-outs in our clusters were all Newspaper Sources, suggesting that grouping by the type of Media would benefit a study on bias in the news.

# **References:**

kotartemiy. (2020, March 1). *GitHub*. Retrieved from Newscatcher: https://github.com/kotartemiy/newscatcher

*Lexico by Oxford*. (n.d.). Retrieved from https://www.lexico.com/definition/fake\_news

*Meriam-Webster*. (n.d.). Retrieved from The Real Story of Fake News: https://www.merriam-webster.com/words-at-play/the-real-story-of-fake-news

Murchie, A. C. (2020, March 1). *Kaggle.* Retrieved from News Headlines & Summary from select soures: https://www.kaggle.com/adammcmurchie/news-headlines-summary-from-select-12-sources

*New York Times*. (n.d.). Retrieved from No One Believes Anything’: Voters Worn Out by a Fog of Political News: https://www.nytimes.com/2019/11/18/us/polls-media-fake-news.html

*Poynter*. (n.d.). Retrieved from 9 questions to help you evaluate the credibility of news sources: https://www.poynter.org/educators-students/2016/9-questions-to-help-you-evaluate-the-credibility-of-news-sources/

*USA Today*. (n.d.). Retrieved from Trust no one? Americans lack faith in the government, the media and each other, survey finds by Alia E. Dastgir : https://www.usatoday.com/story/news/nation/2019/07/23/pew-study-american-trust-declines-government-media-and-each-other/1798963001/