Political Bias in Mainstream News Media

A Preprint

Lily Bhattacharjee[*]
Statistics Undergraduate Student Association
University of California, Berkeley
Berkeley, CA 94720
lbhattacharjee@berkeley.edu

April 27, 2019

***K*eywords** News bias · data science · reporting accuracy

1 Introduction

Journalism – the subtle art of conveying facts about real-world events in a purportedly unbiased manner while pushing an underlying political or social agenda, whether intentional or unintentional, because the writers of news reports are only human. A quick look at tonight's headlines on CNN (a commonly cited leftward-leaning publication) and Fox News (which liberals criticize as being overtly conservative in its curation of articles) reveal the titles "Mueller report won't be the end of Trump's woes" and "Kamala's 'Travesty': Harris' dad slams her remarks on smoking weed and being Jamaican" respectively.

Even though regular viewing of the types of articles that reach online publication, their titles, summaries, representative images, or even potential commenters below can imply that the assertions of partisanship do indeed hold true. In this article, I work on attempting to distinguish between the various ways and biases in which differing news outlets portray hot button issues. Often, people tend to accept the words of major news sources (or even online political blogs) as truth, but with the rise of social media and the now 24/7 news cycle – this has led to some sloppy, fluff, opinion-based reporting in an effort to cut corners in order to remain on top of what's breaking. For example, CNN does highlight opinion pieces with an additional label. Fox does not. It is not always easy to tell how much of a news article is pure fact and how much is second-hand hearsay or analysis, and my project aims to make determining this easier for news readers.

2 Data Analysis

Because there was no freely available central repository for news articles sorted by sources, text, and other metadata, the first challenge in implementing this project was data collection. One of the datasets consisting of known classified "fake" news articles that brief EDA was performed on was taken from Kaggle dataset "Getting Real about Fake News" (link: `https://www.kaggle.com/mrisdal/fake-news`). The column data in that dataset included the following:

- **uuid**: unique article identifier
- **author**: author of story
- **publisheddate**: date published
- **title**: article title
- **text**: article body text
- **languagedata**: language of spam article
- **spam_scoredata**: spam score asssigned from webhose.io

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

- **type**: type of website (label from BS detector)

The remaining data used in these analyses was manually obtained from two major news outlets – *Fox News* (a right-leaning source according to crowdsourced ratings on Media Bias / Fact Check) and the *New York Times* (a left-center source according to the same site). The *New York Times* has a developer-facing API that has a functionality that can accumulate articles chronologically according to a certain query. On March 24, 2018, I queried 50 pages of article results on the topic of "abortion", yielding 500 results by iterating via pagination. The data received included the following columns:

- **headline**: article title
- **lead_paragraph**: first paragraph of the retrieved article
- **snippet**: article preview text before full viewing
- **section_name**: category of writing that the article is classified into e.g. Opinion, U.S., Health, etc.
- **word_count**: article length in words
- **pub_date**: date of publication in the following format YYYY-MM-DDTHH:MM:SS+0000 where H, M, and S represent hours, minutes, and seconds respectively
- **web_url**: link in the *New York Times* domain (`https://www.nytimes.com`) that the article is accessible from

The API does not give automatic access to article text, but using an R script to strip HTML from the pages corresponding to the web urls, body text in raw HTML form is accessible. Regex and text-cleaning techniques were necessary to remove the HTML tags and format the remaining text properly.

Similarly, like most news outlets, *Fox News* does not offer an official API to access article data, so data had to be obtained via web crawling. Selenium, a web automation package with Python bindings, was used to force article pages to load (knowing that paginated queries took the following form: `https://www.foxnews.com/search-results/search?q=query&ss=fn&start=10`) before the HTML was copied and saved into text files. Regex was also needed for cleaning away the tags and code in this case. In total, 10,000 articles associated with the query "trump" were considered in the following analyses, including column data on the article title, introductory paragraph, date published, and article text.

While the News API (`https://newsapi.org`), a publicly available API, does purport to collect articles in real-time corresponding to various queries, my analyses of the data I obtained from it revealed that the API was not collecting stories in an unbiased manner i.e. many top news sources were completely missing from the searches and some of the articles were taken from Facebook links (with Facebook cited as the news source). This, along with the inherent limitations of monthly API call requests, made using the News API exclusively as a source of data biased.

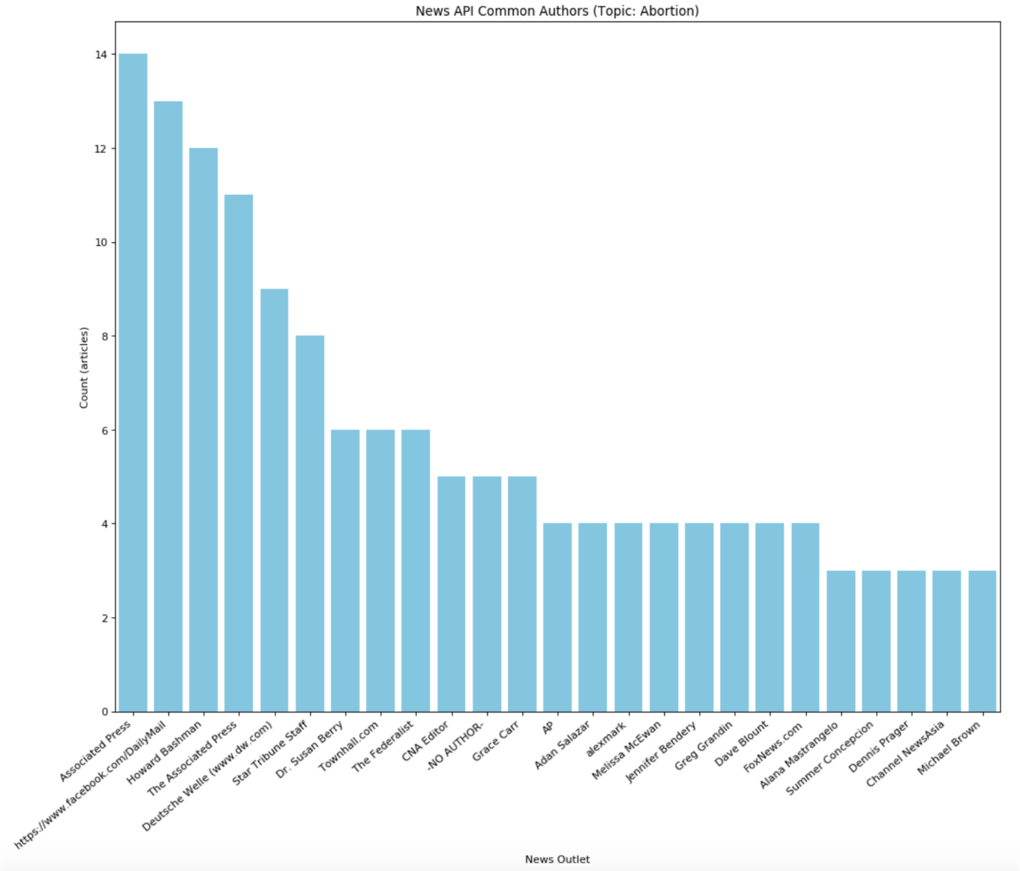
It is important to note that these datasets are in no way a comprehensive outlook on media bias and even the bias of a specific news outlet. The data does not include information or articles from other specific queries on political issues or concern a variety of media sources (other than those in the News API, *New York Times*, *Fox News*, and classified fake news sources in the Kaggle dataset) primarily due to time limitations. Given additional resources to explore future work, it would be a priority to consider the crowdsourced social and economic issues collected on ISideWith (`https://www.isidewith.com/polls`) before moving on to other popular news sources including CNN, NPR, MSNBC, etc. Unfortunately, none of these sources allow easy access to article text either, so gaining access to that is one of the major bottlenecking factors of repeating and extending the results in this project.

3 Visualizations

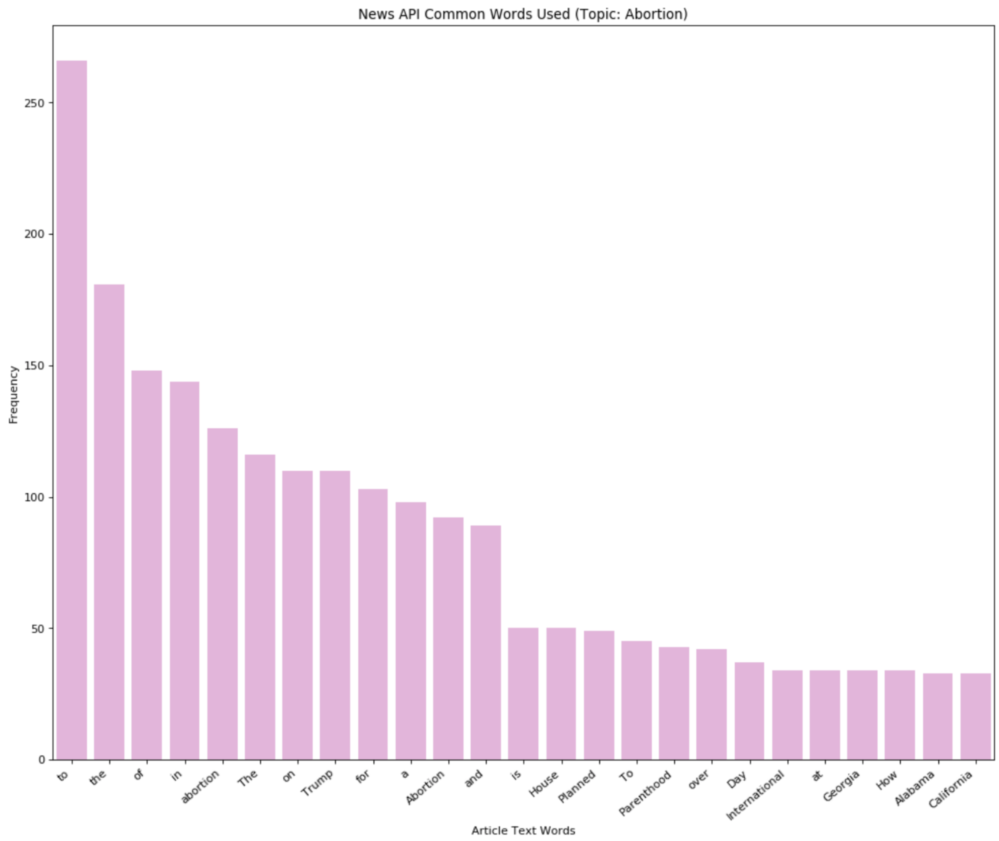
There are five major stages to generating the visualizations that make up this report. Initially, exploratory data analysis (EDA) was performed on the News API to determine whether the API on its own would be a suitable dataset to perform the analysis required for this project. The following stages involved deeper dives into insights provided by one of Python's natural language processing packages – SpaCy – with regards to the *New York Times* articles specifically, the fake news dataset provided by Kaggle, and tonal analysis aided by IBM's Tone Analyzer API. Although this project was not able to determine or create a model to accurately distinguish "real" and "fake" news, both with fuzzy definitions, it does provide a few suggestions for features that can be used to develop a more accurate model for potentially informing readers as they read a particular article.

3.1 News API EDA

The News API returns the author, title, description, date published, and the first few lines of content for news articles associated with the query in a request. According to its website, it returns data from more than 30,000 sources and blogs, but limits users to accessing its database for the last 30 days. Using the query "abortion" the next step was to retrieve around 200 of the most recent articles the News API had obtained and determine whether the dataset was useful for its intended purposes i.e. sampled in an unbiased manner from the pool of possible news sources – most of which had very likely written at least one article on the subject from the past month.



From the bar chart of common authors from this set of abortion-topic articles, it was evident that most large mass media outlets were missing, with the *Associated Press* topping at 14 articles (25, including the varied spelling of *The Associated Press*), *Daily Mail* (British-based tabloid / celebrity-focused news) following at 13, and Howard Bashman – an appellate litigator with a decidedly liberal leaning – following in third. Performing an analysis of this data would be significantly skewed because it is not an accurate representation of the mass news field as it is currently; Howard Bashman, for example, is not a well-known news figure.



Keeping this judgment in mind, when taking a closer look at the article text column in the dataset, it is notable that some of the most common words that are not stop words (useless words in NLP that can be ignored, included articles and widely-used prepositions), we notice that some common terms are "abortion," "Trump," "Planned" / "Parenthood" (referring to the national reproductive healthcare nonprofit controversially funded by the government. None of these articles are confirmed to be spam or untrue, but the notably high usage of stop words with respect to more meaningful ones can be hypothesized to be indicative of a higher frequency of opinion-centric articles as opposed to factual reporting of events.

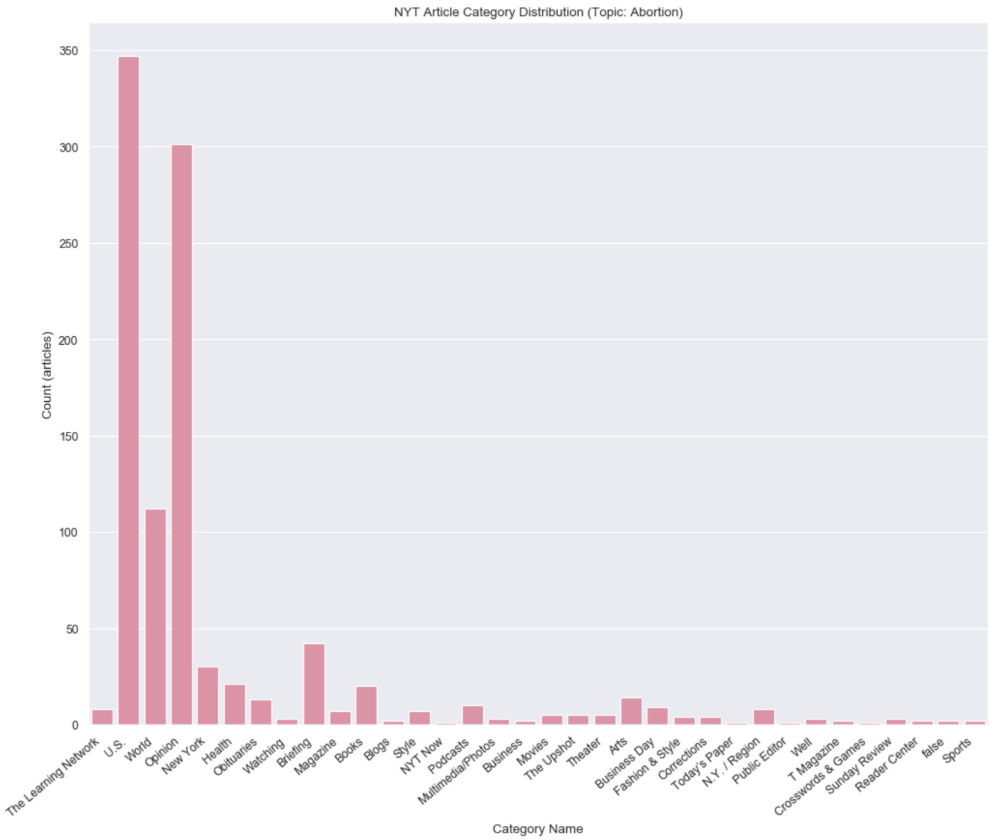
3.2 *New York Times* EDA

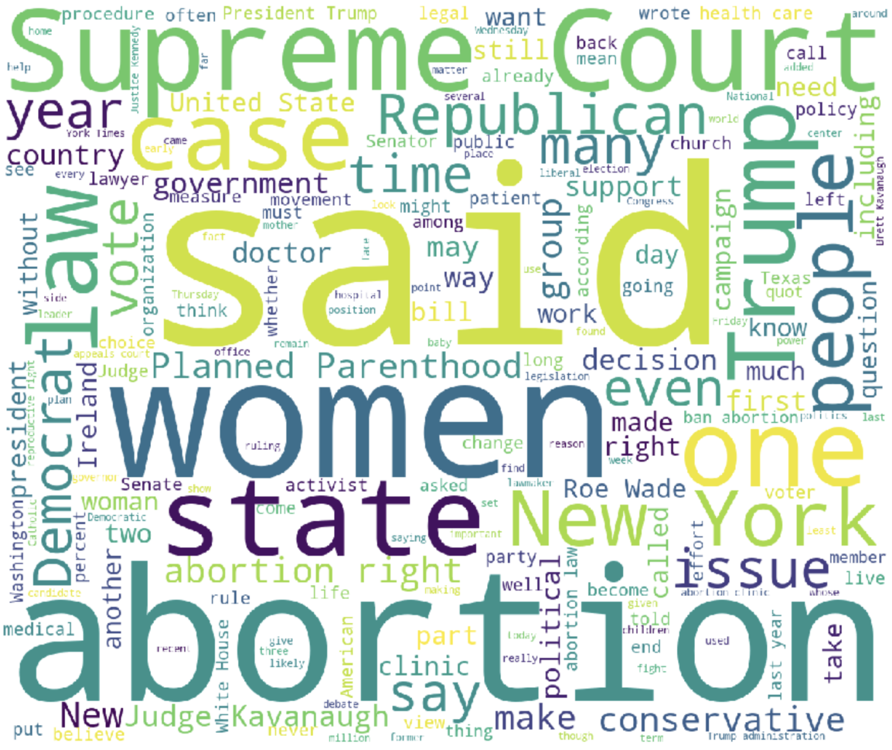
As mentioned previously, the NYT API did not directly return article text, so a full EDA as performed after obtaining the original text via the links embedded in the original JSON data and cleaning the raw HTML. The final dataset included around 500 articles from the last month (same time period at the News API) regarding the same topic – "abortion." Most of the articles fell into the "U.S" and "Opinion" categories, as visible from the bar chart below. The word cloud following it was created using the Python wordcloud library, counting and resizing phrases in the combined article dataset according to the frequency of their occurrence. Some of the most common phrases in these articles are evidently "Supreme Court," "case," "women," "abortion," "state," "New York," and "Democrat" / "Republican" – a potential indication of the partisanship of various opinions and decisions on abortion legality.

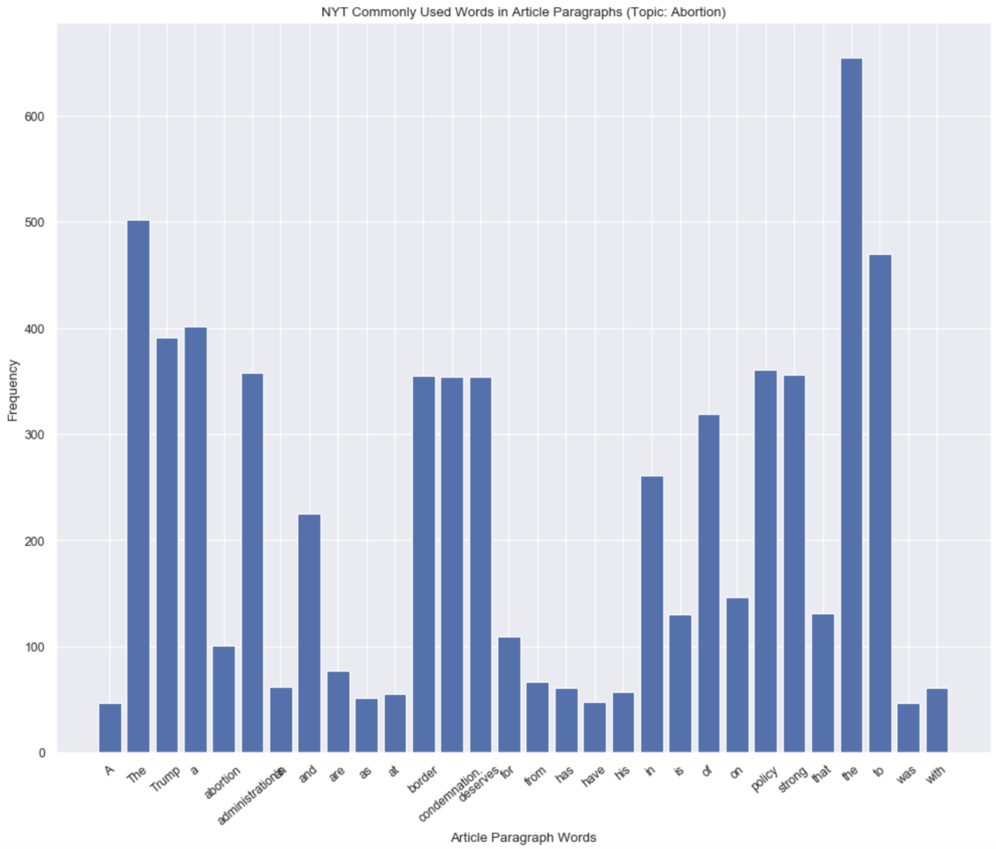
The bar charts below are different from the word cloud because they represent the frequency of word usage rather than phrase usage in three different article data columns – paragraphs, snippets, and headlines. These charts are inclusive of stop words, so "the" is one of the most frequent words used in all three cases, although it is not the most common in headlines, most likely because headlines are selective and short with words. Therefore, they tend to use stop words less often, eschewing fluff for brevity. Some of the most commonly used words in article paragraphs are "Trump," "administration," "border," "condemnation," and "policy." It is interesting to note that these words tend to be longer and more complex in meaning on average than those used in the articles from the News API sample.

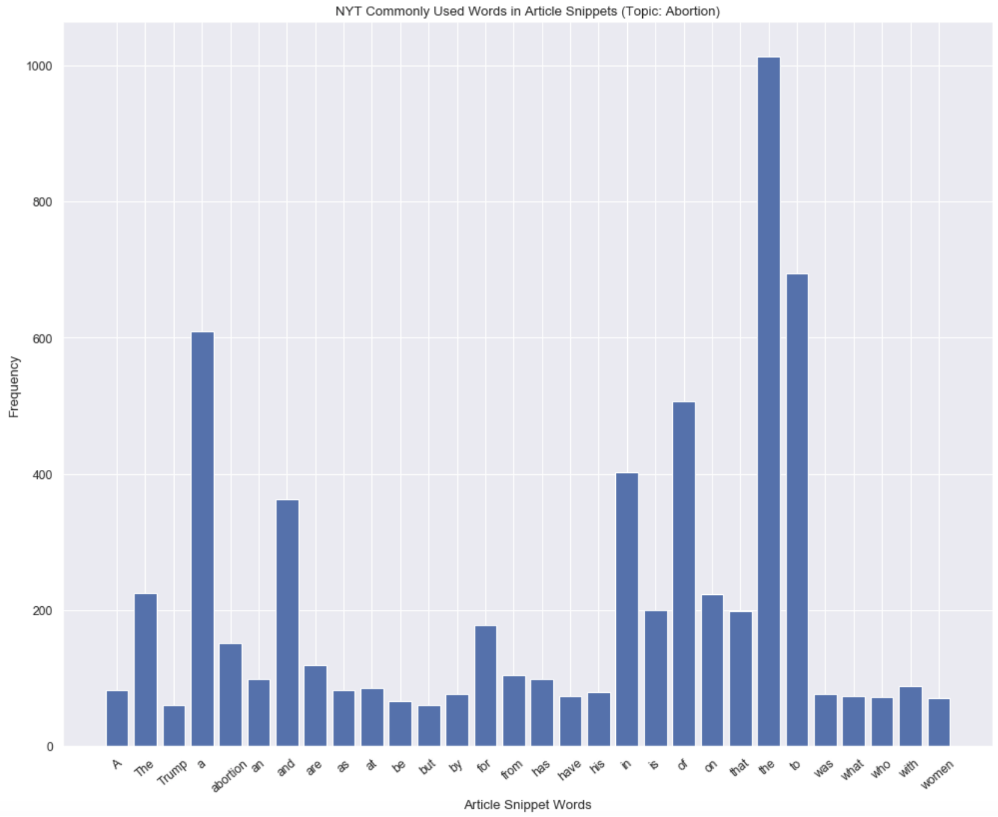
On the other hand, the words used in article snippets – the preview that appears before a user clicks on the article – are majority stop words. From the top 20 most frequently used, as plotted in the bar chart below, the only non-stop words are "Trump," "abortion," and "women" and they are not used comparatively frequently. It is important to note that a sizable minority of articles (20%) did not have available snippets because they were either video links or simply chose not to display a preview, so this may have had an effect on the distribution of the data displayed.

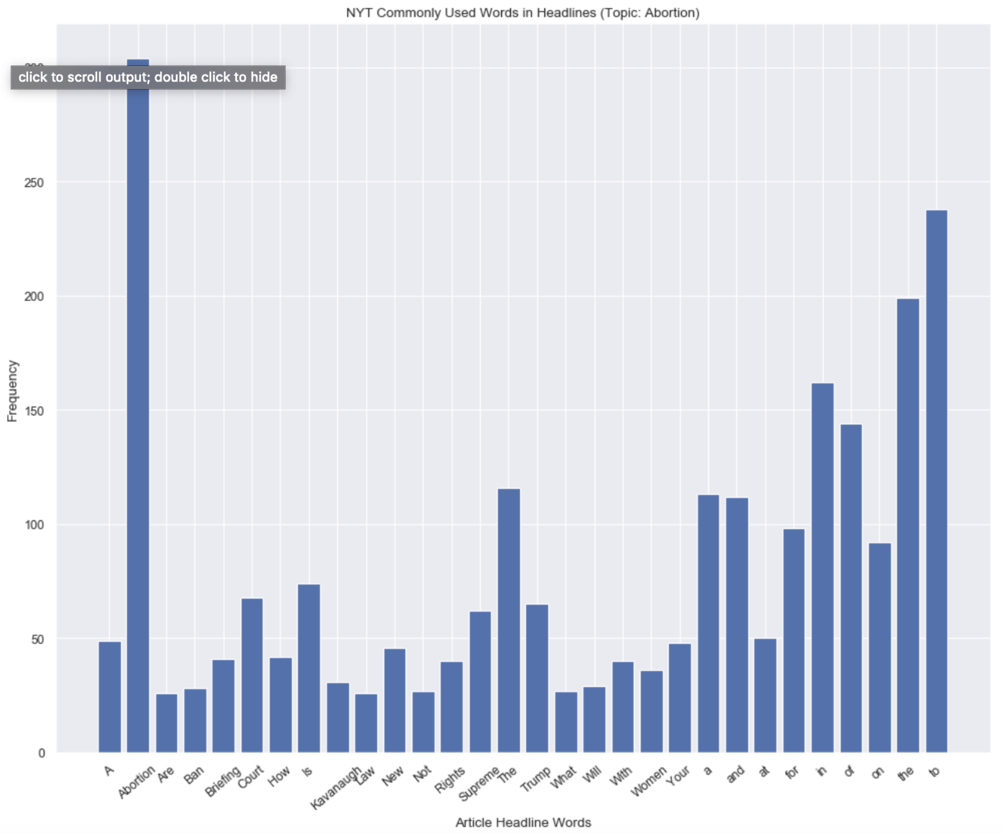
In headlines, "Abortion" is by far the most common word, and disregarding stop words completely, the remaining of the top 20 most common words used are present at approximately the same frequencies e.g. "Kavanaugh," "Trump," "Women," etc. It is possible that some frequencies were artificially decreased due to the case-sensitive counting process e.g. "Women" would be counted differently from "women," so this is an improvement that should be made with further analysis. The headline counts have the largest diversity in words used between all three text categories.

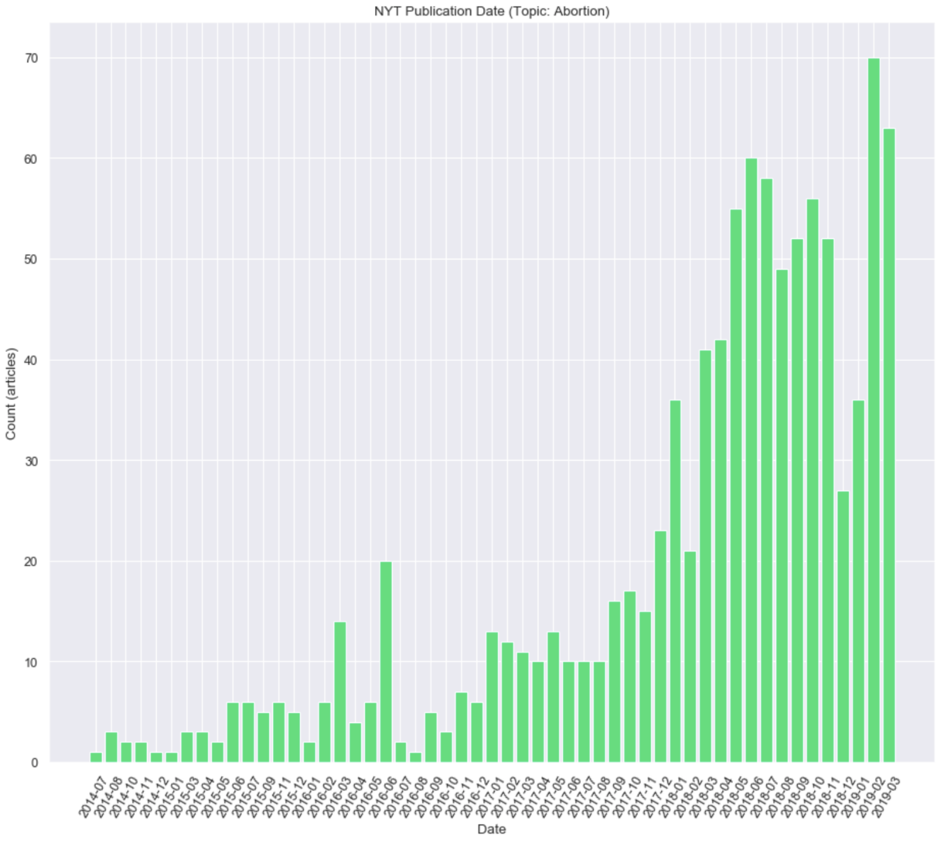

NYT Article Category Distribution (Topic: Abortion)
Count (articles)
0
50
100
150
200
250
300
350
The Learning Network
U.S.
World
Opinion
New York
Health
Obituaries
Watching
Briefing
Magazine
Books
Blogs
Style
NYT Now
Podcasts
Multimedia/Photos
Business
Movies
The Upshot
Theater
Arts
Business Day
Fashion & Style
Corrections
Today's Paper
N.Y. / Region
Public Editor
Well
T Magazine
Crosswords & Games
Sunday Review
Reader Center
false
Sports
Category Name

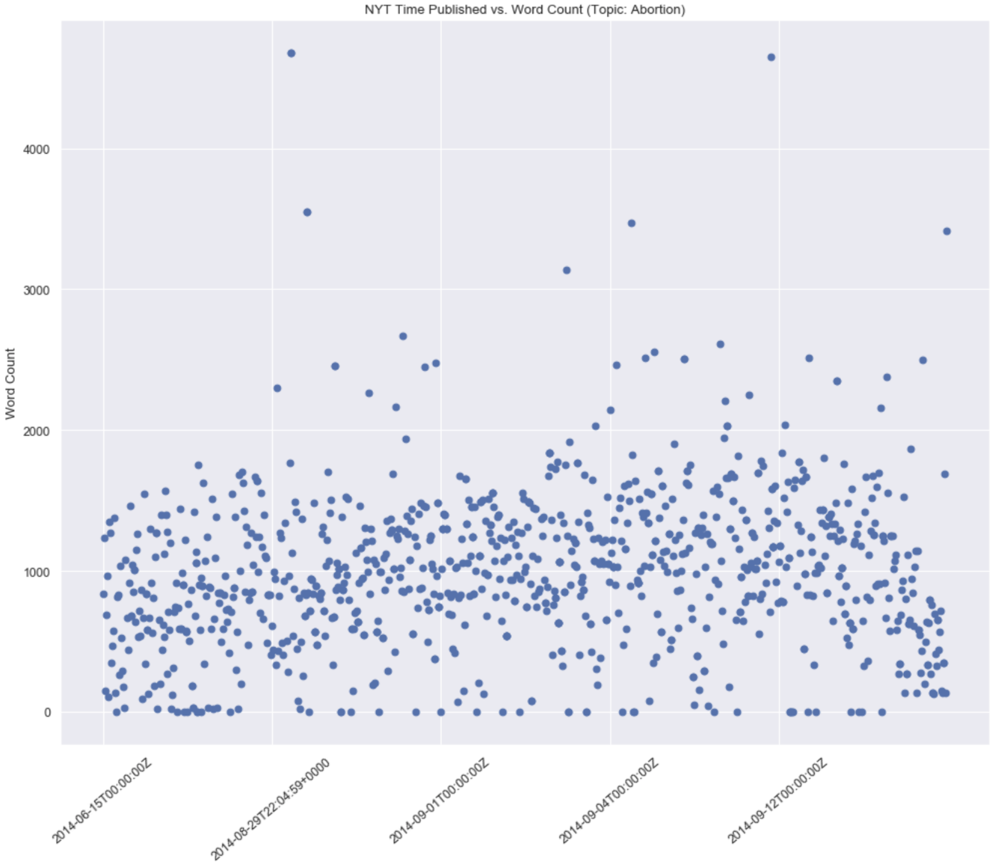

said
abortion
women
Supreme Court
state
case
law
Trump
people
one
New York
Republican
many
time
issue
Democrat
vote
Planned Parenthood
abortion right
United State
year
government
country
support
doctor
decision
even
conservative
Judge Kavanaugh
Roe Wade
say
clinic
make
political
president
Washington
Ireland
campaign
question
including
group
bill
first
right
made
much
know
day
may
way
work
still
want
need
call
policy
church
public
Senator
patient
going
already
procedure
often
President Trump
legal
wrote
health care
back
mean
left
Texas
quot
ban abortion
change
activist
asked
Senate
woman
two
another
percent
party
well
become
called
effort
member
live
voter
medical
rule
life
part
told
end
take
last year
White House
American
view
thing
never
believe
put
New
organization
without
whether
think
choice
Judge
lawyer
measure
movement
must
among
might
according
abortion law
organization

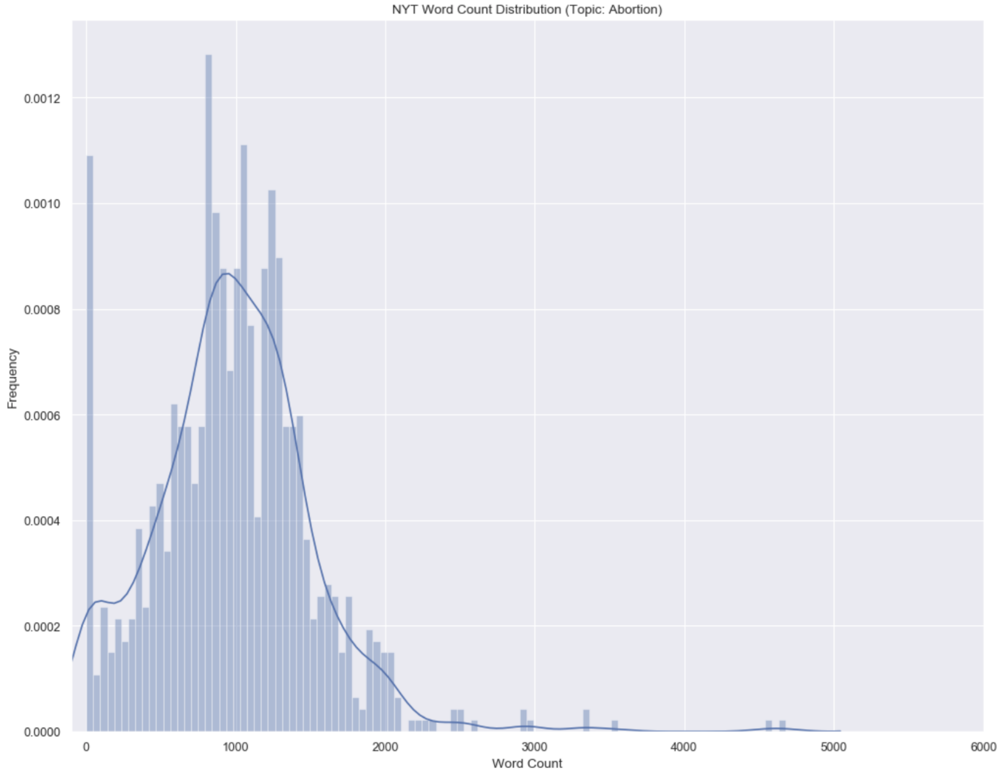

NYT Commonly Used Words in Article Paragraphs (Topic: Abortion)
Frequency
0
100
200
300
400
500
600
A
The
Trump
a
abortion
administration
an
and
are
as
at
border
condemnation.
deserves
for
from
has
have
his
in
is
of
on
policy
strong
that
the
to
was
with
Article Paragraph Words


NYT Commonly Used Words in Article Snippets (Topic: Abortion)
Frequency
0
200
400
600
800
1000
A
The
Trump
a
abortion
an
and
are
as
at
be
but
by
for
from
has
have
his
in
is
of
on
that
the
to
was
what
who
with
women
Article Snippet Words


NYT Commonly Used Words in Headlines (Topic: Abortion)
click to scroll output; double click to hide
Frequency
0
50
100
150
200
250
A
Abortion
Are
Ban
Briefing
Court
How
Is
Kavanaugh
Law
New
Not
Rights
Supreme
The
Trump
What
Will
With
Women
Your
a
and
at
for
in
of
on
the
to
Article Headline Words


NYT Publication Date (Topic: Abortion)
Count (articles)
0
10
20
30
40
50
60
70
2014-07
2014-08
2014-10
2014-11
2014-12
2015-01
2015-03
2015-04
2015-05
2015-06
2015-07
2015-09
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12
2018-01
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-09
2018-10
2018-11
2018-12
2019-01
2019-02
2019-03
Date


NYT Time Published vs. Word Count (Topic: Abortion)
Word Count
4000
3000
2000
1000
0
2014-06-15T00:00:00Z
2014-08-29T22:04:59+0000
2014-09-01T00:00:00Z
2014-09-04T00:00:00Z
2014-09-12T00:00:00Z
Time

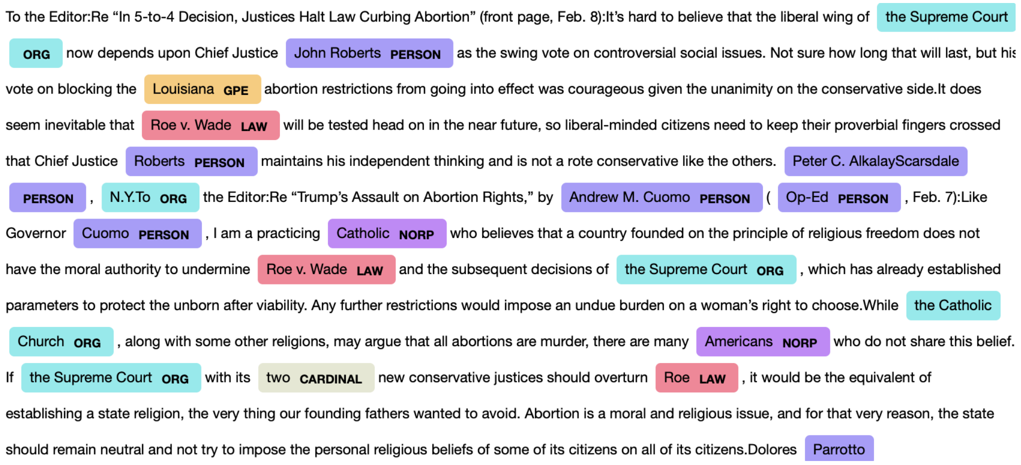

NYT Word Count Distribution (Topic: Abortion)
Frequency
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
1000
2000
3000
4000
5000
6000
Word Count

Besides the text-heavy aspects of the NYT dataset, there were some numeric columns that yielded a few insights about the types of articles NYT tends to publish. As evidenced from the scatter plot displaying time vs. word count over a period of 5 months, most articles hover around 1000 words as a mean, disregarding the few zero- or close to zero-word articles that are actually embedded video clips. There are some outliers that overshoot 3,000 (and in two cases, 4,000 words, but those two articles in particular are categorized as opinion pieces). Hence, over time, NYT published at an increasing rate for the topic "abortion" during the accessible time period with noticeable spikes at 2016-06, 2018-06 (close to midterm elections), and there is a slight trend upward in word count over time.

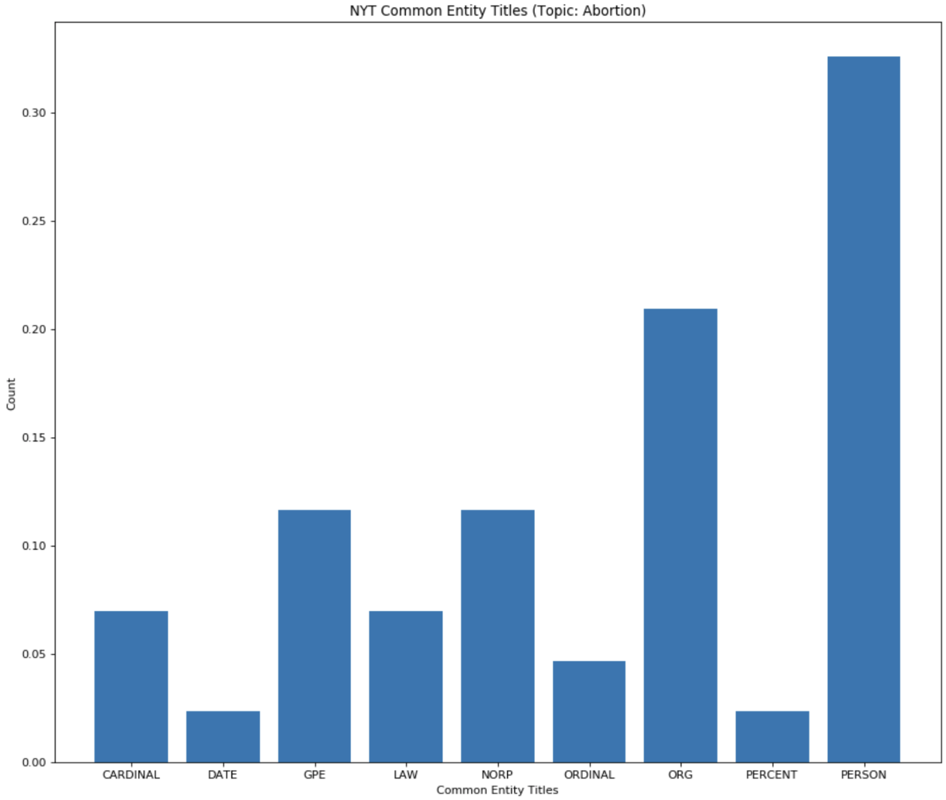
The word count graph below this chart verifies the analysis above, with the number of words in any given article following this skewed right distribution. Noticeable modes are 0 and 1000, corresponding to videos and normal-length articles respectively.

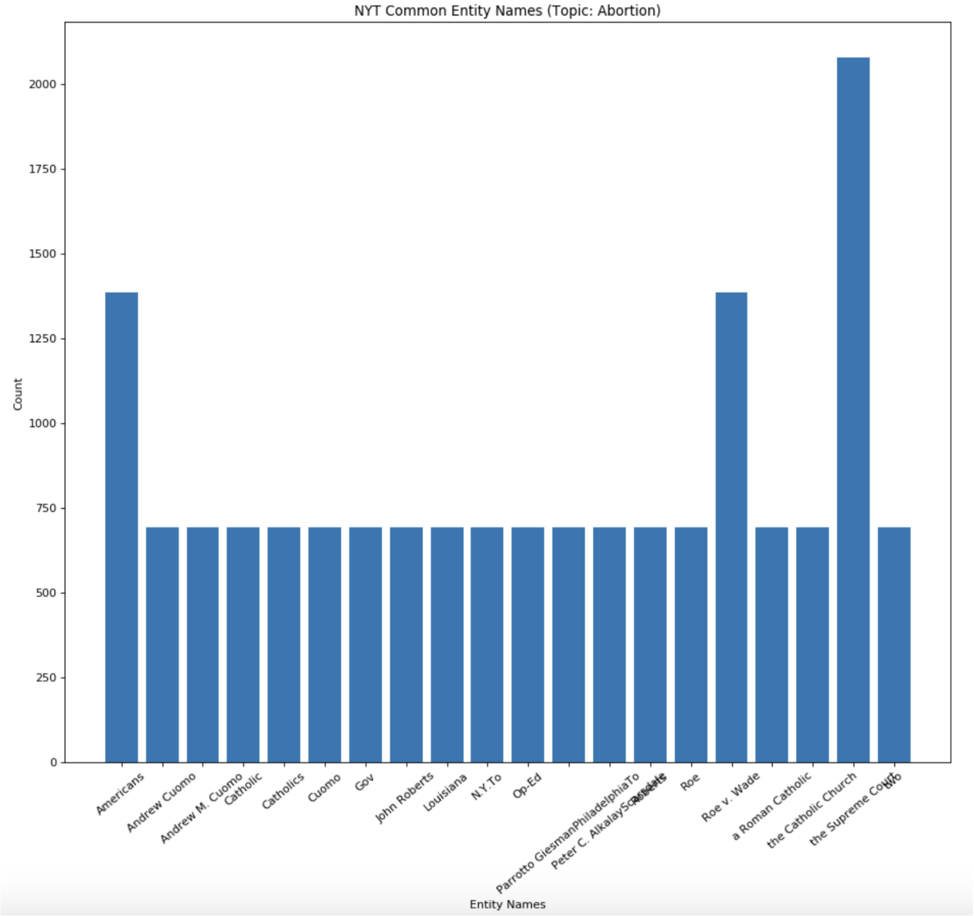
3.3 Exploring *New York Times* Article Text with SpaCy

In this section, the same NYT articles as the previous section are analyzed with Python's NLP package SpaCy, which is useful for visualizing and quantifying sentence structure in large texts. The second graphic in this section with highlighted words in a sample paragraph of text represents SpaCy and its ability to distinguish between known types of nouns in text e.g. mentioned public figures, organizations, laws (these are called entities). Most of the entities are classified into the following categories according to the bar chart below: "Person," "Organization," "Geo-Political Entity" (GPE) in that order. The most common entity names are "Americans" (Person category), "the Catholic Church" (Organization category), and "Roe" (a well-known law easing access to abortion resources and legalizing it across the nation).



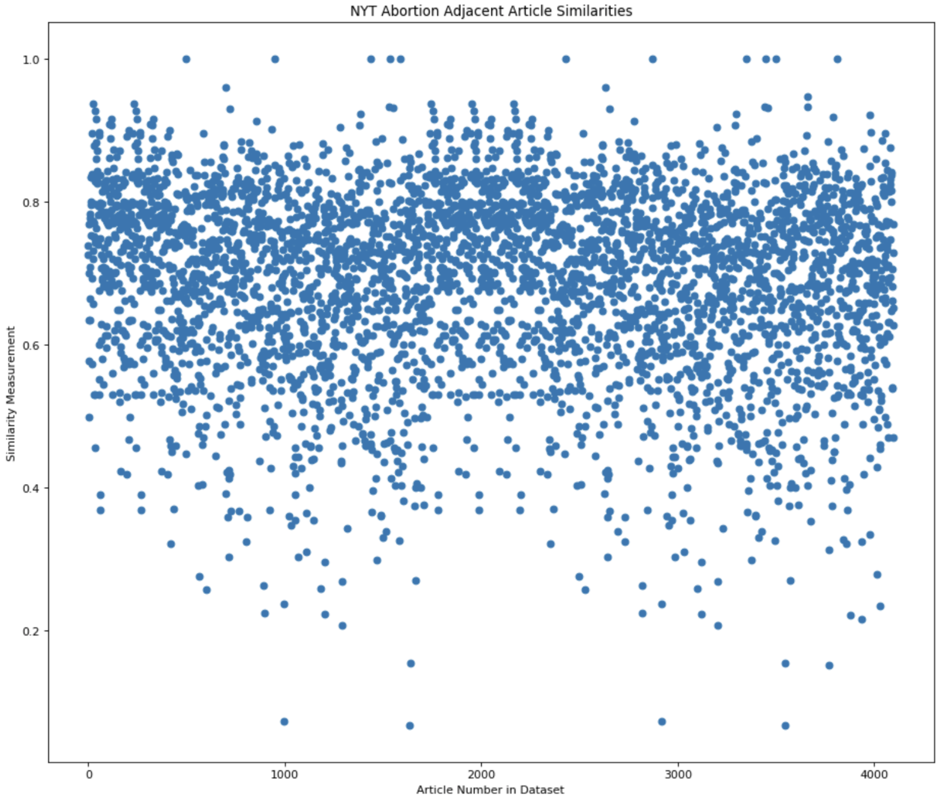
To the Editor:Re "In 5-to-4 Decision, Justices Halt Law Curbing Abortion" (front page, Feb. 8):It's hard to believe that the liberal wing of the Supreme Court ORG now depends upon Chief Justice John Roberts PERSON as the swing vote on controversial social issues. Not sure how long that will last, but his vote on blocking the Louisiana GPE abortion restrictions from going into effect was courageous given the unanimity on the conservative side.It does seem inevitable that Roe v. Wade LAW will be tested head on in the near future, so liberal-minded citizens need to keep their proverbial fingers crossed that Chief Justice Roberts PERSON maintains his independent thinking and is not a rote conservative like the others. Peter C. AlkalayScarsdale PERSON , N.Y.To ORG the Editor:Re "Trump's Assault on Abortion Rights," by Andrew M. Cuomo PERSON (Op-Ed PERSON , Feb. 7):Like Governor Cuomo PERSON , I am a practicing Catholic NORP who believes that a country founded on the principle of religious freedom does not have the moral authority to undermine Roe v. Wade LAW and the subsequent decisions of the Supreme Court ORG , which has already established parameters to protect the unborn after viability. Any further restrictions would impose an undue burden on a woman's right to choose.While the Catholic Church ORG , along with some other religions, may argue that all abortions are murder, there are many Americans NORP who do not share this belief. If the Supreme Court ORG with its two CARDINAL new conservative justices should overturn Roe LAW , it would be the equivalent of establishing a state religion, the very thing our founding fathers wanted to avoid. Abortion is a moral and religious issue, and for that very reason, the state should remain neutral and not try to impose the personal religious beliefs of some of its citizens on all of its citizens.Dolores Parrotto

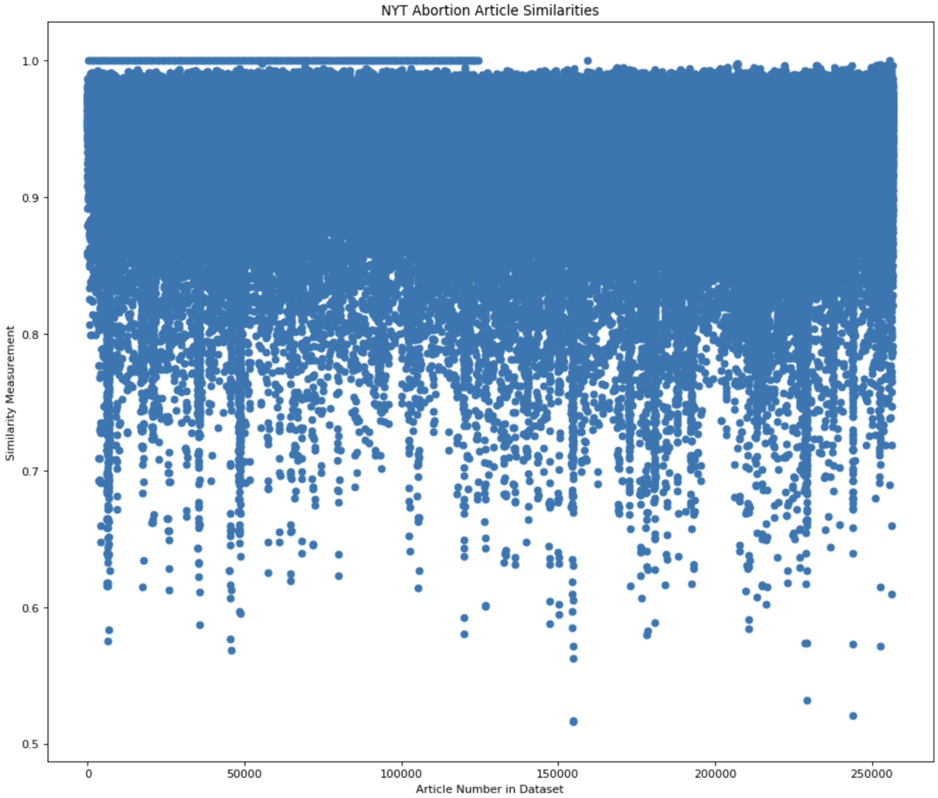




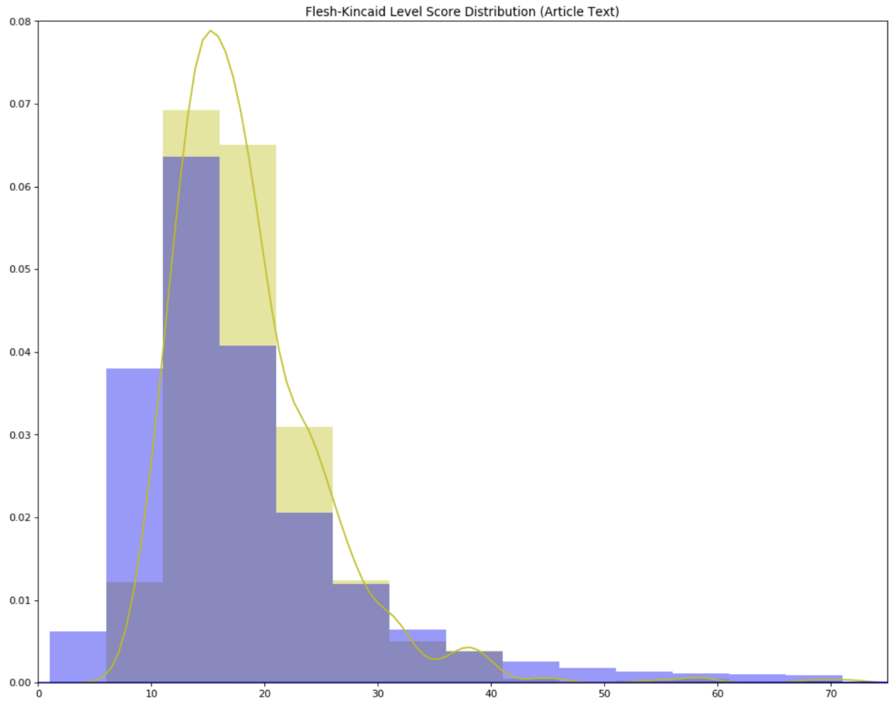
SpaCy also allows for the calculation of the structural similarity of different pieces of text using its nlp similarity method. Iteratively running this method, I was able to calculate the similarities of adjacent articles (or article pairs published right after each other) and the scatter plot below exemplifies the relative structural consistency of article text over time – similarity hovers around 0.6-0.8 with a few outliers at 0 that may be accounted for by the presence of no text / video 'articles.'

There are also outliers on the other end at 100% similarity, although the articles compared in those cases do not contain the same text. The scatter plot below the one for adjacent articles performs the similarity method between all article pairs, reinforcing the statement that not only are adjacent articles similar to each other, but NYT writing style and sentence structure has remained very consistent at 0.8 similarity over time. The 1.0 similarity lining the top of the scatter plot can likely be attributed to the original article being compared with itself.


NYT Abortion Adjacent Article Similarities
Similarity Measurement
1.0
0.8
0.6
0.4
0.2
0
1000
2000
3000
4000
Article Number in Dataset


NYT Abortion Article Similarities
Similarity Measurement
1.0
0.9
0.8
0.7
0.6
0.5
0
50000
100000
150000
200000
250000
Article Number in Dataset

As an aside to the SpaCy analysis, an initial hypothesis I had in terms of distinguishing false from real news is that fake news may be less readable overall than articles from legitimate sources. A common measure of readability is the Flesh-Kincaid metric, which uses the formula $206.835 - (1.015 \times ASL) - (84.6 \times ASW)$. In this case, ASL stands for average sentence length and ASW for average number of syllables per word. This metric is expected to – overall – return the grade level of reading for the text.



From the graphic above, in which the blue distribution represents the Flesh-Kincaid scores of the Kaggle Fake News dataset and the yellow represents the scores of the NYT dataset, although there is a slight lower bias for the known fake news i.e. these false articles tended to be more readable than NYT articles, this is almost negligible. Both have scores with modes in the range 10-20 and are college-level readable according to the scoring formula, so this metric cannot be used accurately to classify between the two categories.

3.4 Kaggle Fake News Dataset EDA

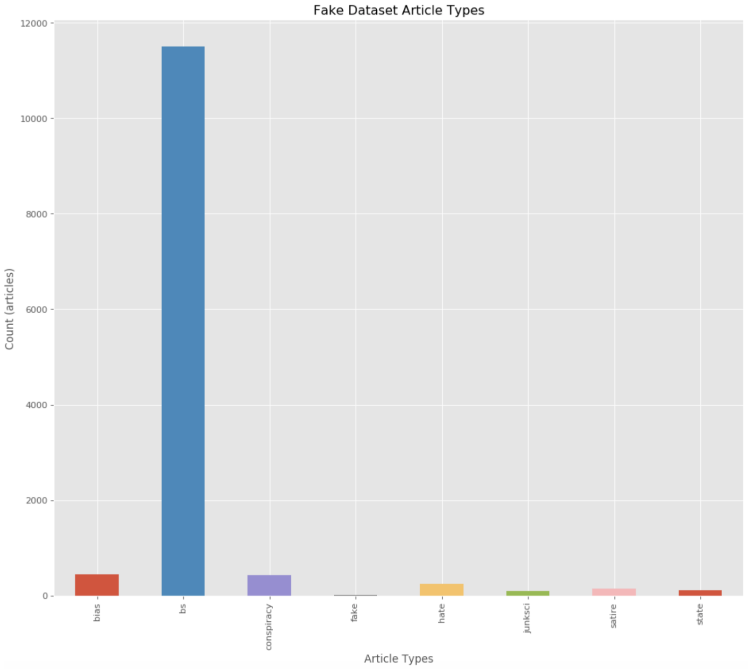
Throughout this paper, the terms "fake" and "real" news may have been confusing because they are difficult to define for a program, although not so difficult for a perfect fact checker to categorize. In this dataset, "fake" news is split into 7 major categories, defined as the following:

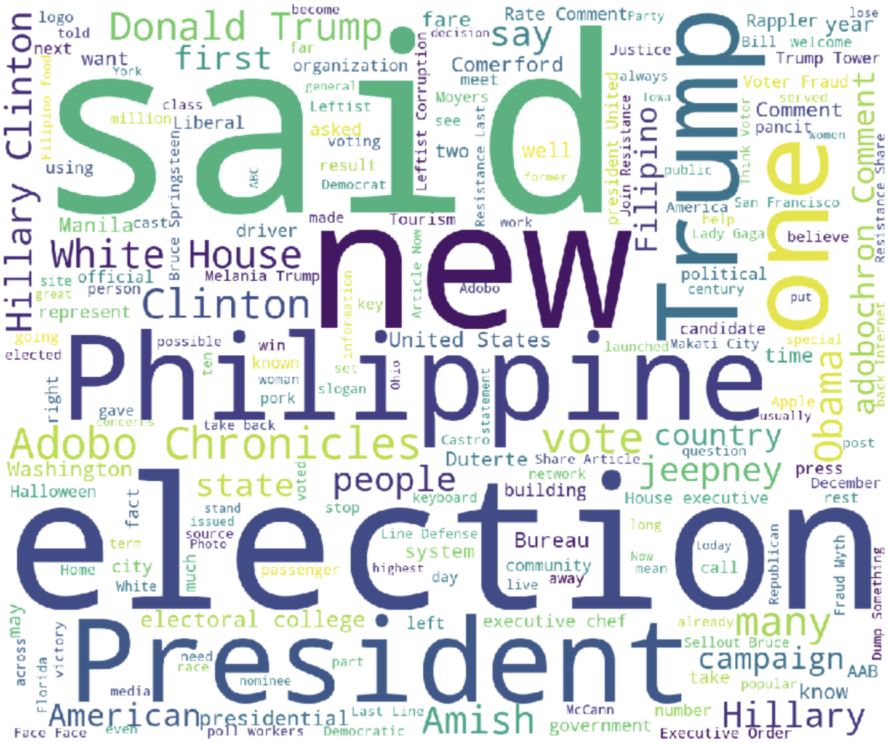
- **bias**: news that subtly pushes a political agenda / portrays certain political figures or events with a spin to influence the readers
- **bs**: factually incorrect / completely-made-up news
- **conspiracy**: news based on conspiracy theories e.g. flat earth
- **fake**: news that doesn't necessarily fit into the other categories mentioned but is still factually incorrect
- **hate**: news that expresses hatred towards a particular social / cultural group
- **junksci**: news based on inaccurate or false science e.g. challenges to global warming based on cyclical temperature changes
- **satire**: news that resembles Onion-style reporting
- **state**: news controlled financially / editorially by the government

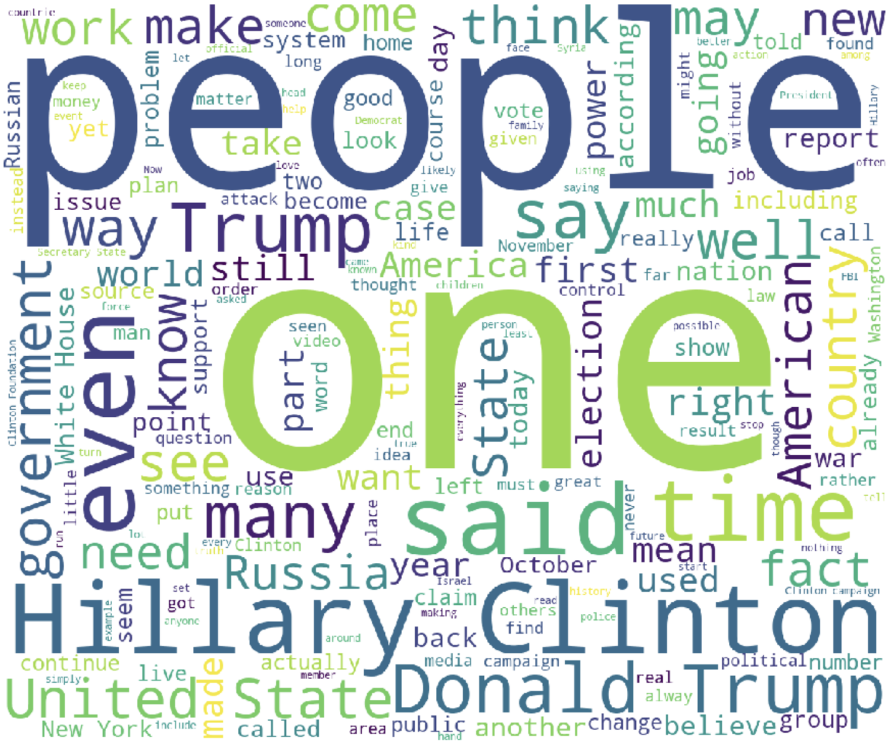
Because "bias" is the largest category, it is the only one considered in the generation of the word cloud in the second graphic. The "fake news" concerns a variety of topics, so it is difficult to tell if any of the words in the cloud are likely to be irrelevant to the articles they were pulled from. It is interesting that one of the words used most frequently is

"Philippine" considering that it is a misspelling of the "Philippines," so it is possible that misspelled / misused / rarely used words like the aforementioned, "Adobo," and "jeepney" are more common amongst biased news (note differences from the NYT word cloud).

In terms of fake news overall, there are not many obviously out-of-place words, although more in-depth analysis can be performed with phrases rather than words to determine whether adjacent collections of words are used more often than the single words themselves. For example, the words "one" and "people" are frequently used in this dataset as they are approximately the same size, but the word cloud does not make clear whether these words are used together, or perhaps even joined with other words.



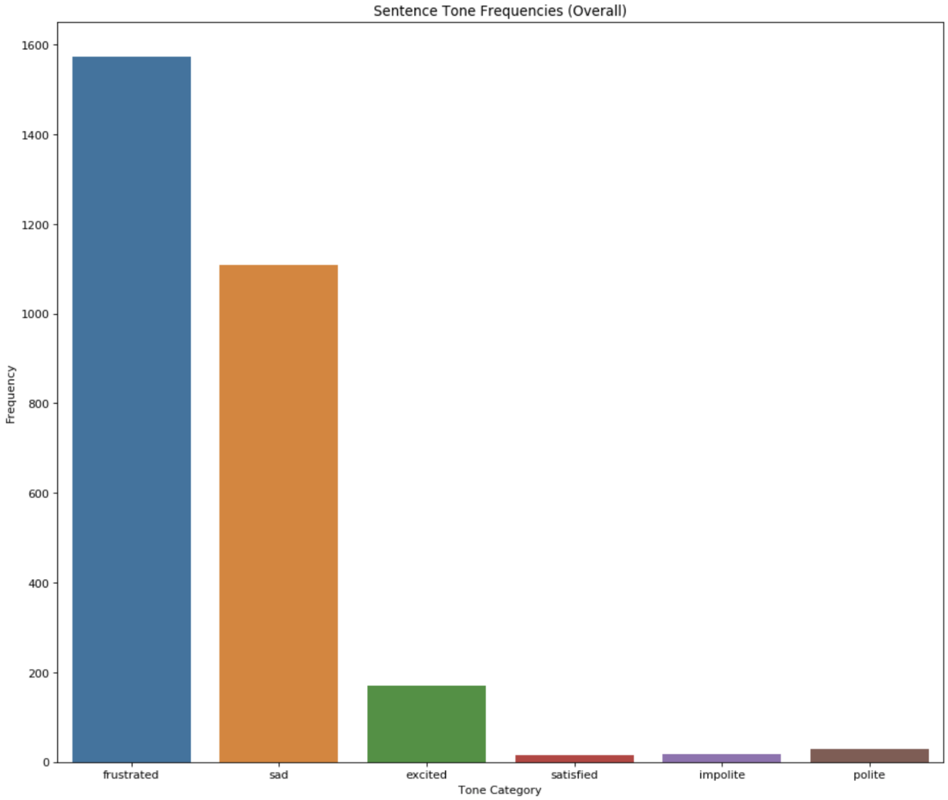


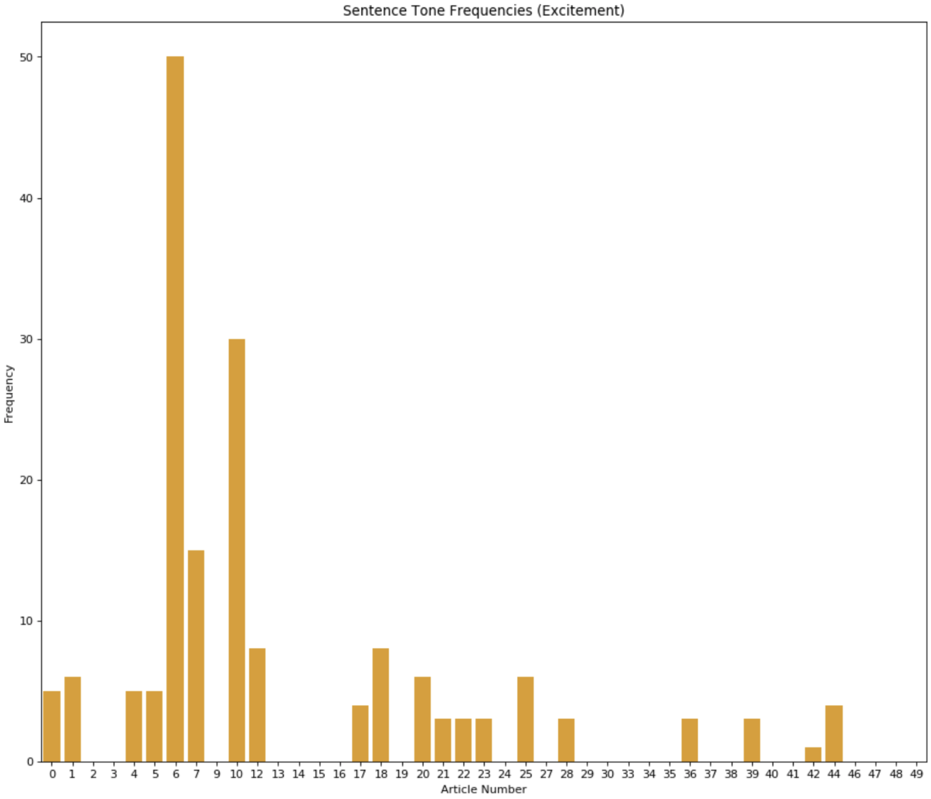


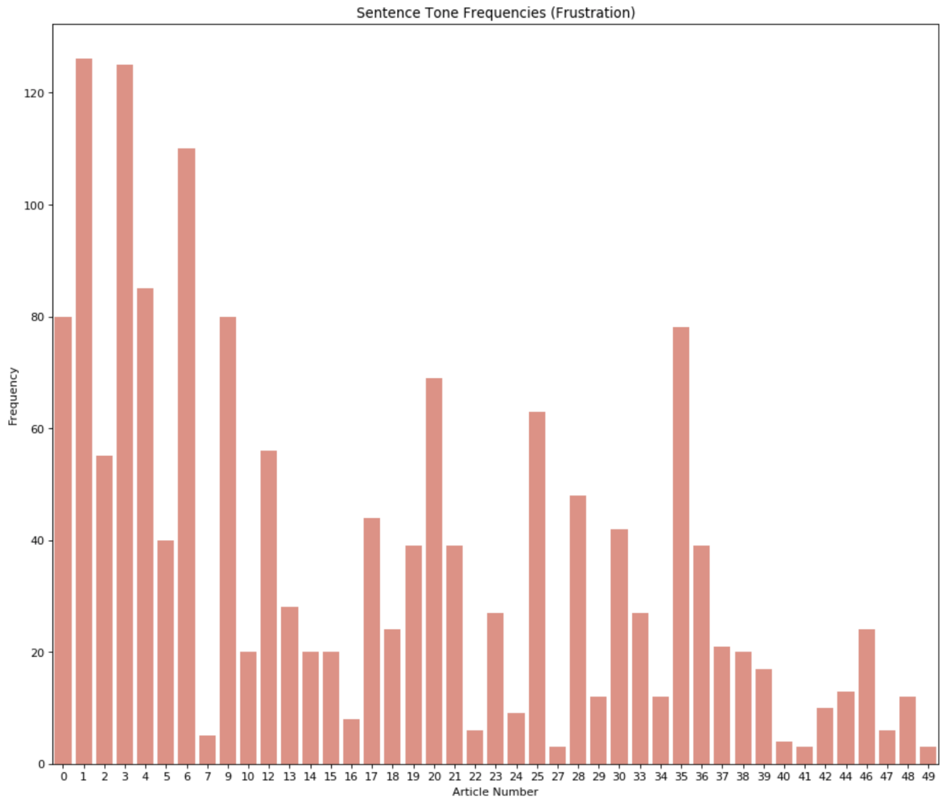
3.5 Exploring the *New York Times* with IBM Tone Analyzer

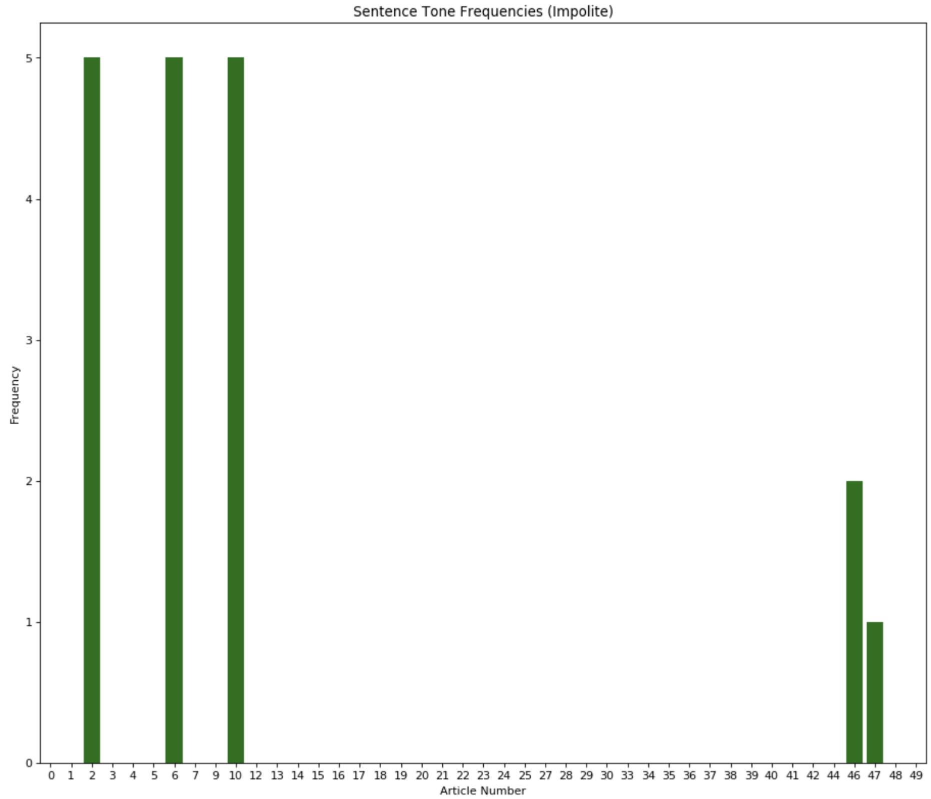
Some of the last visualizations I created used the NYT abortion articles and the IBM Tone Analyzer API, an IBM Cloud software designed to detection emotions in text. Given a chunk of text, the API partitions it into sentences and subsequently assigns tones that closely approximate the emotions expressed by each individual sentence, with the options being "excitement," "frustration," "impolite," "polite," "sad," and "satisfied." Running the API on 50 randomly sampled individual articles from the NYT dataset, the tone results were represented in the visualizations below. The process was actually quite complicated – with the initial body text stored in the form of a Python dataframe being converted into JSON format, making an HTTP request to the API, processing and writing the response text to a text file, and re-converting the JSON in the file to an analysis-friendly Python form.

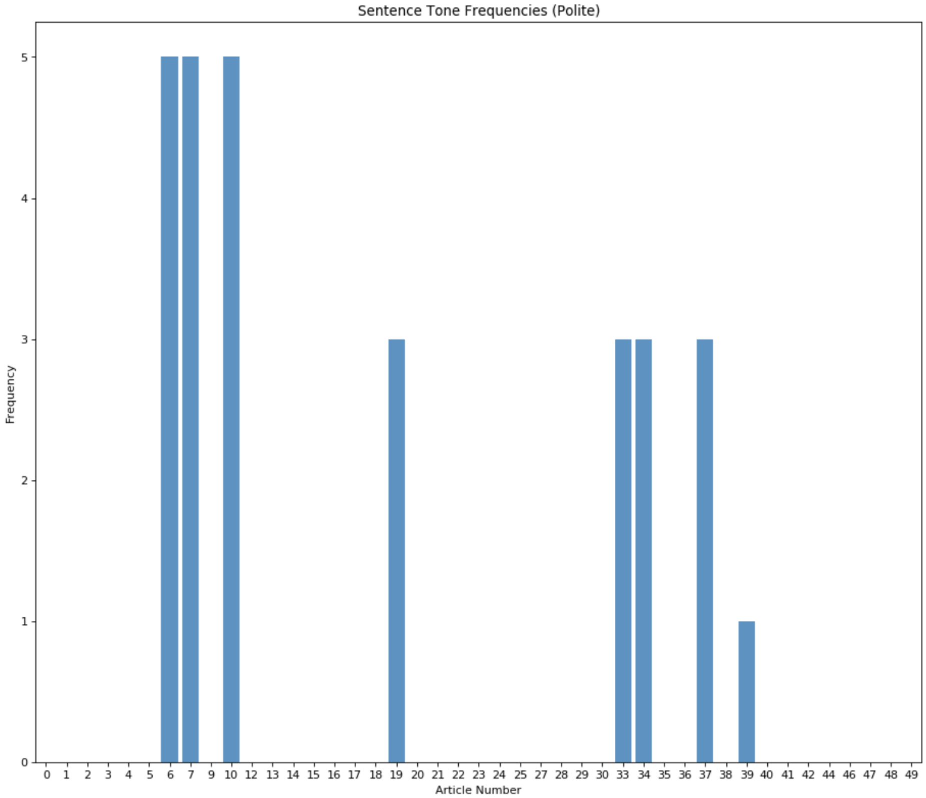
We notice that the majority of NYT article sentences are categorized into a "frustrated" tone while the rest fall into "sad." The graph distributions following the initial bar chart represent the counts of particular emotion categories among specific articles. For example, Article 6 has the largest number of "excited" sentences (50), Articles 1, 3, and 6 have the most "frustrated" sentences, Articles 2, 6, and 10 have the most "impolite" sentences, Articles 6, 7, and 10 have the most "polite" sentences, Articles 1 and 35 have the most "polite" sentences, and Article 1 has the largest number of "satisfied" sentences. An extension to this analysis would be performing a similar tone processing of the fake news dataset to see if the tones tend to differ between the two groups because in this case the analysis is quite one-sided.

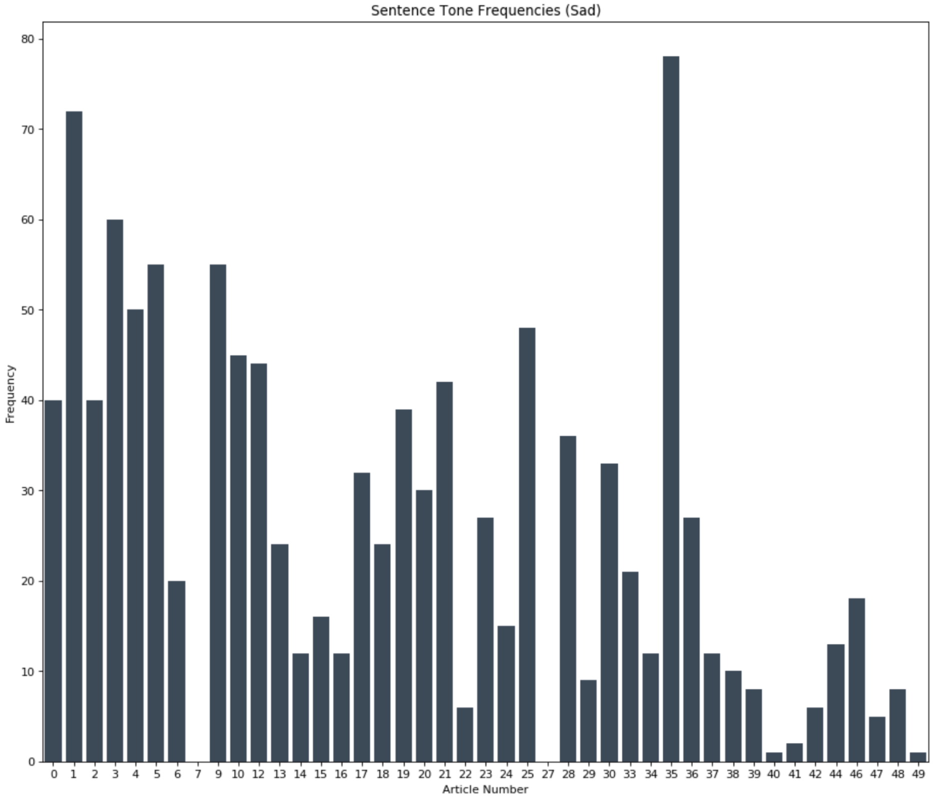

Sentence Tone Frequencies (Overall)
Frequency
1600
1400
1200
1000
800
600
400
200
0
frustrated
sad
excited
satisfied
impolite
polite
Tone Category

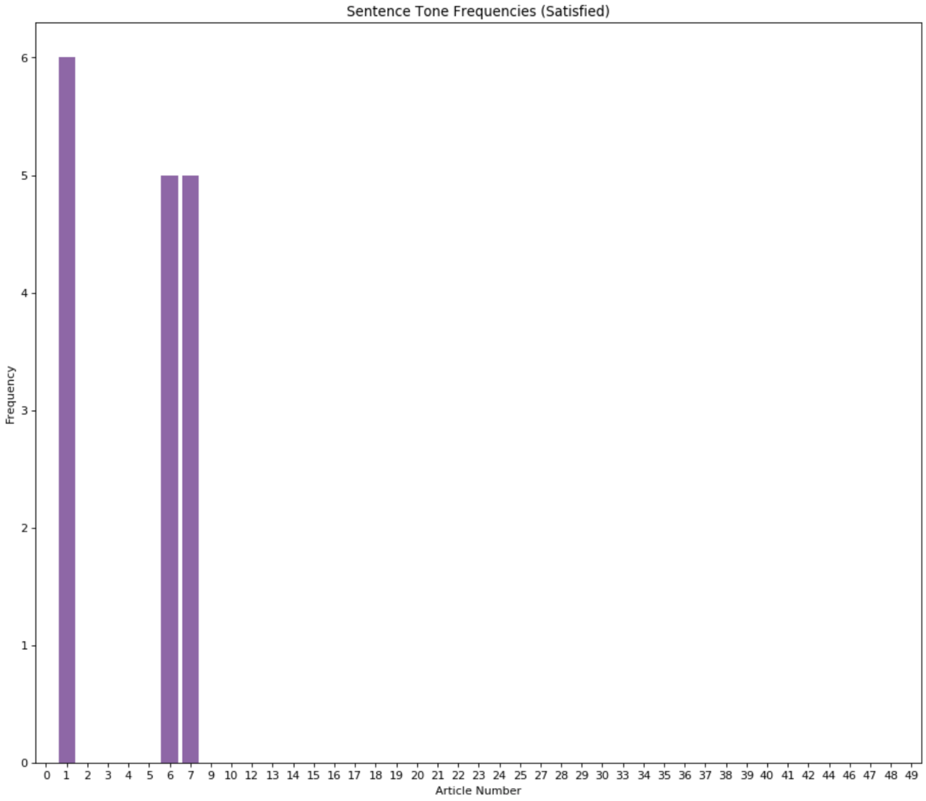

Sentence Tone Frequencies (Excitement)
Frequency
50
40
30
20
10
0
0 1 2 3 4 5 6 7 9 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 28 29 30 33 34 35 36 37 38 39 40 41 42 44 46 47 48 49
Article Number


Sentence Tone Frequencies (Frustration)


Sentence Tone Frequencies (Impolite)


Sentence Tone Frequencies (Polite)
Frequency
0
1
2
3
4
5
0 1 2 3 4 5 6 7 9 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 28 29 30 33 34 35 36 37 38 39 40 41 42 44 46 47 48 49
Article Number


Sentence Tone Frequencies (Sad)
Frequency
0
10
20
30
40
50
60
70
80
0 1 2 3 4 5 6 7 9 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 28 29 30 33 34 35 36 37 38 39 40 41 42 44 46 47 48 49
Article Number



4 Conclusion

I started this project with the optimistic expectation that I would be able to determine an overarching metric for distinguishing or at least flagging news articles that were likely to contain false or misleading information, an iterative fact checker that pulled all of the sentences from a page that the user was viewing live and outputted a percentage warning based on how trustworthy the information presented was likely to be. Although the analysis I performed over this semester was in-depth and time-consuming, the results are far from this goal. From papers that have worked on similar topics, research groups have made progress in measuring the factual adherence of a news article, but testing accuracy remains low. From my work here and derived from my analysis of my visualizations, I have a few suggestions for metrics that can contribute to separating misleading / false information purported as facts:

- **word count**: Overall, the mean word count of article text from the Kaggle Fake News Dataset was < 500 words whereas reputable sources like the NYT hovers around 1,000 words on average for the most part. Allowances should be made for detecting video 'articles' that do not fit this word count requirement
- **use of mispelled / misused / rare words**: Known "flagged" fake news sources are more likely to contain these words, while the NYT relies on more conventional words, as evidenced from the word cloud, to get its point across.
- **time published**: sources that publish reputable news tend to publish consistently and for long periods of time and during work hours (9AM-5PM depending on time zones) rather than being unable to trace back / being published at odd times

Some metrics that may not work well include the following:

- **reading score**: This metric is not enough on its own. The Flesh-Kincaid score has holes in its calculations because it is based off of syllables and sentence length – the resultant text does not necessarily have to be difficult to read. However, there is a noticeable lower bias for known fake news, so further exploration into whether or not fake news is more readable is a viable path.
- **article similarity**: Although NYT articles are pretty similar to each other structurally on the whole, there is a sizable number of outliers that would be misclassified if known factual articles were compared to an unknown

real / fake article for classification. This might need to be combined with a more accurate metric to be useful in any way.

Some future work that will build upon these findings include the following:

- obtaining data from a wide variety of news sources from across the spectrum of crowdsourced political bias – *CNN*, *MSNBC*, *Breitbart*, *The Federalist*, etc.
- collecting articles on various controversial topics – in this paper we focus mostly on "abortion" but there are far hot button issues ranging from immigration to economic and international relations
- including analyses of the authors that write these articles and how their biases have remained consistent or changed over time. An interesting thing that CNN does is that it mentions whether the writer for a particular article is known to be a conservative commentator, affiliated with a liberal think tank, etc. A lot of the time, these authors are pushing the agendas of the organizations they are a part of, and the public deserves to know that before they read the article.

References

Source 1: “Identifying Political Bias in News Articles”
(https://www.ieee-tcdl.org/Bulletin/v12n2/papers/lazaridou.pdf)
Source 2: “Political Ideology Detection Using Recursive Neural Networks”
(https://people.cs.umass.edu/ miyyer/pubs/2014_RNN_framing.pdf)
Source 3: “Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis”
(https://www8.gsb.columbia.edu/media/sites/media/files/JustinRaoMediaBias.pdf)
Source 4: “Sentiment Analysis of Political Tweets: Towards an Accurate Classifier”
(http://www.anthology.aclweb.org/W/W13/W13-1106.pdf)
Source 5: “On the Challenges of Sentiment Analysis for Dynamic Events”
(https://arxiv.org/pdf/1710.02514.pdf)
Source 6: “Fake News Detection on Social Media: A Data Mining Perspective”
(https://www.kdd.org/exploration_files/19-1-Article2.pdf)
Source 7: “The Quest to Automate Fact-Checking”
(http://cj2015.brown.columbia.edu/papers/automate-fact-checking.pdf)
Source 8: “Fact-Checking Polarized Politics: Does the Fact-Check Industry Provide Consistent Guidance on Disputed Realities?"
(https://www.americanpressinstitute.org/wp-content/uploads/2016/02/Marietta-Barker-Bowser-2015-Forum.pdf)
Source 9: “Bias in Newspaper Photograph Selection”
(http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.870.963rep=rep1type=pdf)