News Media During an International Crisis: What Twitter Data Says About COVID-19

Emilio Luz-Ricca

Abstract

This project explores the presence of news media on Twitter during the novel coronavirus. In a period fraught with high uncertainty and strict social distancing orders, news media provides reliable coverage of global events while allowing individuals to remain in the safety of their homes. Using named entity recognition and sentiment analysis, common subtasks of natural language processing, in conjunction with data visualization, I explore a novel data set of 39,791 tweets from ten major news sources collected during the month of March, 2020. Looking at the data by topic, news source, and over time, I highlight trends in the inclusion of named entities and the sentiment ingrained in coverage. Guided by the results of my analysis, I suggest further lines of research and outline potential methods for approaching future research. This work contributes to the field of crisis informatics by taking the first steps towards describing the role of news media during periods of international crisis.

1. Introduction

Twitter is an online microblogging platform that hosts a variety of content, with 330 million active users in 2019.[CITATION NEEDED] When approaching social media as sources of raw text data, Twitter is most often used for their relatively liberal policy on data availability. Specifically, Twitter offers tools for developers and researchers through their public, but somewhat restricted, application program interface (API). Once data has been collected, tools from computational linguistics and natural language processing (NLP) can be applied using robust open source packages (Stanza, NLTK, spaCy, TextBlob, etc.) written for popular programming languages (usually python).

In particular, the emerging field of crisis informatics has often used Twitter data in conjunction with NLP techniques to explore the complex interactions between organizational entities and individuals during times of crisis.[CITATION NEEDED] Common lines of research within the field of crisis informatics include the filtering of data on social media to identify and summarize accurate information, building systems to aid crisis management organizations, and describing the ways in which citizens and authorities cooperate on social media during crisis situations.[CITATION NEEDED] Unfortunately, source data in crisis informatics research is biased towards English Twitter data, reflecting both the stranglehold that social media sites have on "big text data" and the language-support limitations within NLP.[CITATION NEEDED] While I do not provide solutions to these problems within this project, it is important to acknowledge this bias as a prevalent issue within crisis informatics and NLP research as a whole.

The age of the novel coronavirus (COVID), a pandemic that has devastated countries globally and disrupted social patterns for much of 2020, certainly should be regarded as a crisis situation. Whereas many case studies within the field of crisis informatics focus on small-scale or relatively isolated crises, COVID is a crisis of unprecedented scale and with global implications. COVID provides an opportunity to reformulate underlying theory within crisis informatics as well as explore novel interactions amongst a broader range of participatory actors. In this way, COVID represents a uniquely unifying experience within the diverse global community.

However, the scale of COVID also presents a distinct issue: the data is too "big." That is, there is so much to study concerning social media and COVID that approaching all possible aspects in one study is infeasible. As such, I choose to only look at one class of actors: traditional news media. News media has a significant presence on Twitter, with four news sources appearing in the top 50 most followed Twitter accounts (CNN Breaking News, CNN, The New York Times, and BBC Breaking News with 58, 48, 47, and 44 million followers respectively). Furthermore, a 2018 study from Pew Research Center indicates that Twitter is a highly news-focused social media site, with 71% of users getting news from Twitter (it is worth noting that they are not necessarily getting their news from news outlet accounts).[CITATION NEEDED] Situational influences pertaining specifically to COVID might even further inflate this statistic. For instance, the strict stay at home orders enacted in many countries limited the potential supplies of information to citizens. In another more recent Pew study, 89% Americans indicated that they got their news about COVID from national news outlets (not exclusively), with 67% for international outlets.[CITATION NEEDED] The proportion for national news outlets matches that of public health officials (also 89%) and is higher than that of Donald Trump and his task force (67%) or state and local elected officials (81%).[CITATION NEEDED] News media is easily accessible through online news articles, social media posts, or television, and is a generally reliable source of information which, during a period marked by high uncertainty, is valuable.

For this project, I use the ten most followed news sources on Twitter that are not "breaking news" accounts. This mainly has to do with the fact that accounts that deliver "breaking news" rarely link to actual news articles and are thus less comparable to more traditional forms of news media and to other news sources on Twitter. The sources used in the final data set are: CNN, ABC News, The New York Times, The Washington Post, Reuters, The Economist, The Wall Street Journal, TIME, The Associated Press, and BBC News (World). A basic overview of each source's Twitter presence and background are included in Table 1. In future tables and figures, I will refer to each news source by their Twitter handle as shown in the "Twitter_Handle" column of Table 2.

	Twitter_Handle	Country	Num_Followers (M)
The New York Times	nytimes	US	46.8
CNN	CNN	US	48.5
BBC (World)	bbcworld	UK	28.2
The Economist	theeconomist	UK	24.8
Reuters	reuters	UK	22.0
The Wall Street Journal	WSJ	US	17.8
TIME	TIME	US	17.2
ABC News	ABC	US	15.7
The Washington Post	washingtonpost	US	15.9
The Associated Press	AP	US	14.2

Table 1: Information on the news sources used in this project. The "Num_Followers" column is the number of followers for each account on 6/12/2020.

The data used in this project includes all tweets from the aforementioned sources from 3/4/2020 to 4/1/2020 (one month of data). To understand how the data collection period fits into the chronology of COVID-related events, I include a basic timeline in Figure 1.

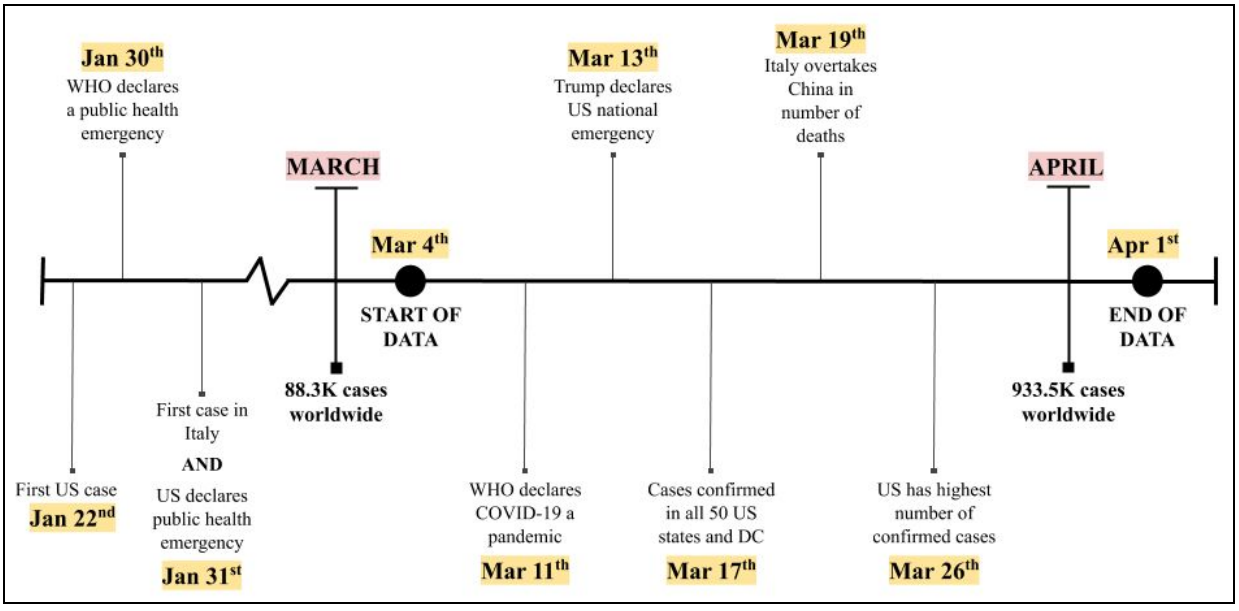

Figure 1: A timeline of key events occurring before and during the data collection period.

Raw text data is almost never useful on its own, which is where tools from NLP and computational linguistics come in. Commonly used tools include: automated content analysis and topic modeling (unsupervised machine learning approaches that attempt to identify latent

frames or underlying topics within text data), sentiment analysis (NLP tasks that categorize the polarity, positive/negative/neutral, or subjectivity, subjective/objective, of a given text), named entity recognition (statistical and machine learning methods concerned with identifying and tagging "named entities" within a text), and more. While topic modeling and content analysis approaches generally produce more interesting standalone results, both are quite computationally expensive and are generally unfit for data exploration. Named entity recognition (NER) and sentiment analysis (SA) are perhaps less analytically robust methods, but both still provide interesting results and can help reveal trends within text data.

SA implementations can take many forms, from lexicon-based to machine learning approaches, with some implementations combining the two with hybrid approaches.[CITATION NEEDED] NER implementations generally rely on supervised machine learning algorithms trained on hand-annotated corpora. While both of these methods perform well on more formal text data (grammatically correct and little to no lexical variants, as defined by CITATION NEEDED), Twitter data can be a difficult application setting because of the unstructured/informal nature of the text data. Challenges associated with using standard NLP packages for Twitter data are covered more in Section 2.2.

While perhaps not as accurate as human-based approaches (i.e., human coding or hand annotation), NLP approaches offer much more time-efficient methods of analysis. This is important as research within crisis informatics is partially concerned with providing automated, real-time systems that can aid in crisis management efforts.[CITATION NEEDED] In my examination of news media on Twitter, I take a very basic descriptive approach. After all, COVID represents a novel situation in many ways, which calls for exploration as a preliminary step. This analysis is not rigorous by any means, instead serving as the initial steps within the context of further research concerning COVID. Certainly, there is much more to study here: the role of individuals on Twitter, the interplay between news media and individuals, collaborative efforts for relief during COVID, the general dissemination of accurate (or inaccurate) information, etc. Additionally, the development of non-English tools within NLP, such as Stanford's python package Stanza which supports 66 languages, will aid in broader research efforts concerning global trends. That being said, the purpose of this project is primarily to aid in further research efforts by revealing interesting lines of analysis. The emerging field of crisis informatics has just been provided possibly the largest crisis of recent history; it is extremely important that we use this as an opportunity to improve systems, theory, and techniques so as to effectively understand and aid in future crisis situations.

The rest of the paper will be structured as follows: Section 2 introduces the methodology for both data collection and analysis; Section 3 presents the results via a number of visualizations; Section 4 discusses the results; Section 5 acknowledges the limitations associated with the methodology and research design, providing recommendations for further research. Additional figures and tables are included in the "Additional Materials" Section, which also includes a link to a GitHub repository including some of the code written for this project.

2. Methods

2.1 Data Collection:

This data was collected using a python script that I wrote, which uses Twython (a wrapper library for the Twitter API). A summary of the data is provided in Table 2.

	Num_Tweets	Percent_Unique	Percent_Retweets
nytimes	3207	0.963	0.198
CNN	4615	0.796	0.039
bbcworld	1186	0.995	0.317
theeconomist	3179	0.681	0.022
reuters	12030	0.796	0.034
WSJ	2633	0.930	0.009
TIME	2398	0.626	0.047
ABC	4649	0.810	0.051
washingtonpost	4130	0.959	0.055
AP	1764	0.966	0.472
Overall	39791	0.831	0.078

Table 2: A basic description of each data set. "Percent_Unique" denotes the portion of the tweets that are duplicated, i.e., published more than once with identical text by the same news source.

While I distinguish between unique and duplicate tweets in Table 2, I chose to use duplicate tweets for all forms of analysis and for the entire exploration phase. In approaching this data, I aimed to characterize these news sources based on how they presented themselves to users. As such, I considered duplicate tweets a key component of how each news source chose to disseminate information, and opted not to remove them. The number of retweets is important to consider, as the text for some retweets is truncated by default by the Twitter API. All retweets, including truncated retweets, were used in the analysis, as removing either would have impacted sources disproportionately (i.e., 47.2% of the tweets from AP are retweets but less than 1% are retweets from TIME). Further discussion of the impact of truncated tweets is provided in Section 5.1.

2.2 Characteristics of the Text Data:

Text data is often categorized in the literature as either well-structured or noisy.[CITATION WOULD BE NICE] For instance, news article data would be considered well-structured text data as they generally do not contain spelling or grammatical errors and do not use lexical variants such as acronyms ("lol" instead of "laugh out loud") or intentional spelling errors ("coooool" for

emphasis instead of "cool"). However, tweet data from general users is usually considered quite noisy for the opposite reasons. Unfortunately, the text data that I collected falls somewhere in between these two extremes. This tweet data contains emojis, "@" and "#" references, and is sometimes structured in strange ways to adhere to the format of a tweet. There is not much need for lexical normalization before analysis, but at the same time it is not perfectly structured data. In this way, the data is linguistically "semi-structured."

A large issue with dealing with semi-structured data is the lack of resources available that have been optimized for this particular task. Some of the preprocessing steps that I applied before analyzing the data are not entirely verifiable in the literature. Along the same lines, implementations within NLP are generally not well suited for semi-structured text data as they are most often trained on (machine learning approaches) or verified against hand annotated texts that fall into one of the two extremes. Certainly, more research is needed to better understand which tools perform best with semi-structured data and how preprocessing decisions might affect results.

2.3 Analysis Methods:

To get a better understanding of potential trends within this data, I applied a few simple analysis methods. I chose to use methods that were computationally inexpensive and relatively time-efficient, given overall time constraints for this project. The methods that I applied were NER and SA. I had intended to also apply word collocations and word frequencies, but time did not permit this step of the analysis. I used spaCy for NER, mostly because of how time-efficient their statistical implementation is. While there are other implementations that are much better suited to Twitter data[CITATION NEEDED], initial tests with Stanford's python NLP library, Stanza, indicated that the slightly better identification and tagging performance would not be worth the much longer processing times (I tested both implementations on the TIME data set and evaluated results by inspection). The named entity (NE) tags supported by spaCy's NER implementation and their corresponding descriptions are included in the additional materials. For SA, I opted to use TextBlob's lexicon-based approach. While it has been shown that lexicon-based approaches consistently perform worse than machine learning or hybrid approaches when applied to tweet data[CITATION NEEDED], the reproducibility and efficiency of a lexicon-based approach is preferred for this project. TextBlob's approach assigns a polarity value between -1 (very negative) and +1 (very positive) to a given text by considering the words in the text that appear in the lexicon. A polarity of 0 is interpreted as neutral. A link to TextBlob's lexicon is included in the additional materials.

2.4 Preprocessing:

Each analysis method required a different level of preprocessing, with some libraries handling certain aspects of preprocessing on their own. The preprocessing steps that I applied for each method are detailed below:

The preprocessing steps that I took for NER were generally conservative. I tried to apply the best possible preprocessing for spaCy's NER, without significantly affecting the semantic or syntactic structure of the text. A visual representation of the entire pipeline is included in Figure 2.

1. First I removed emojis. Sometimes emojis would be misconstrued as NEs, which I deemed incorrect.
2. Next, I fed the text data to spaCy.
3. Finally, I filtered out "@" and "#" references, which is a very conservative approach. I opted not to remove these references from the text entirely, as they were often structurally important (as opposed to emojis) and might affect spaCy's ability to detect NEs. Instead, I filtered out these references if they were detected as NEs, because they were often inconsistently identified and/or incorrectly tagged ("@mattbrier, ORG" is a NE, but the organization tag is inaccurate). I also filtered out any NEs with text matching with "coronavirus," "covid," or "covid-19" for similar reasons (I made all characters lowercase before checking).

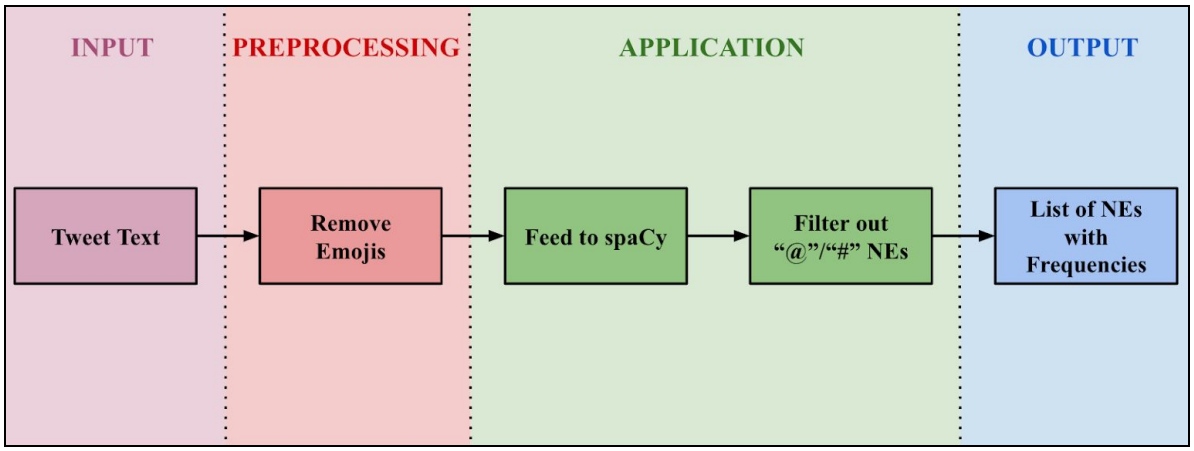

Figure 2: The full NER pipeline with each element sorted into its respective processing stage.

Because of the nature of TextBlob's SA implementation, the preprocessing steps were minimal.

1. I considered removing emojis and "@"/"#" references. After a few initial tests, I determined that it would not be necessary to remove these elements. Using the BBC (World) data set, I saw minimal differences in results both when just removing "@"/"#" symbols from the text and when completely removing the tokens containing "@"/"#" from the text. This indicates that "@"/"#" references are not in the lexicon, even when removing these special symbols. Emojis are also not in the lexicon. So, I decided to remain conservative once again, and I did not interfere with these elements during preprocessing. This again helped to preserve the linguistic structure of the text (I was largely concerned with generating falsely

negated statements, i.e., "...not @willrichards happy..." might become "...not happy...", which may not represent the intention of the initial text).

2. In the end, I directly fed the text data to TextBlob with no preprocessing.

2.5 Classification of Tweets:

In several portions of the exploration, I looked at the differences between tweets about COVID and tweets not about COVID. I used a simple approach for classification: if the lowercase text of a tweet contained the substring "covid" or "coronavirus," I classified that tweet as "about COVID." With regards to the statistical F_1 test, conceptually, this approach is high in precision but low in recall. A further discussion of this classification method and other possible approaches is included in Section 5.3.

All code for the collection method, preprocessing steps, and the application of both NER and SA are included in the linked GitHub repository. The dehydrated data sets are included in the repository as well.

3. Results

To begin the exploration portion of this project, I brainstormed several possible lines of analysis that might be interesting to follow using this data. While there was not enough time to follow all of these lines, I include some additional possibilities for future research in Section 4.

3.1 SA Results:

First, I approach the results from SA. I am not splitting by COVID/non-COVID tweets yet, only looking at the general trends within the data set. In Table 3, I randomly sample five tweets as examples of the types of tweets categorized as positive, neutral, and negative. I include examples that give an idea of not only the most polar tweets (+1 and -1), but also tweets that are polarized to a lesser extent (+0.5 and -0.5). Next, Figure 3 shows the distribution of polarities. I use a logarithmic scale for this figure because the tweets in the data set are skewed towards a polarity of zero (interpreted as neutral).

User	Polarity	Tweet_Text
CNN	1.0	US stocks finished a turbulent week with gains on Friday, logging their best day since October 2008
ABC	0.5	MORE: Among those looking instead for one who "can bring needed change" (35% of voters), Sanders prevailed, 53-44%, per preliminary exit poll results.
reuters	0.0	Big banks reassure staff about potential job cuts
washingtonpost	-0.5	Opinion: The GOP just smuggled another awful provision into the big stimulus bill
reuters	-1.0	Japan says virus has made economy's condition 'severe', worst view in seven years

Table 3: Examples of positive, neutral, and negative tweets as classified by TextBlob's standard lexicon-based SA approach.

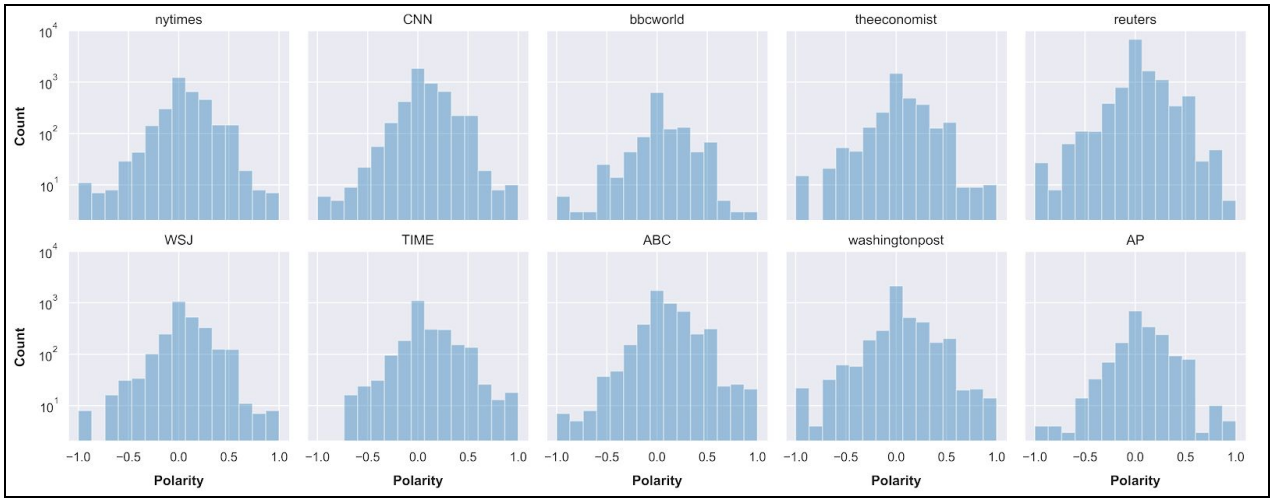

Figure 3: Polarity distributions for each news source on a logarithmic scale.

To better understand the extent to which zero-polarity tweets influence the data set, I look at the proportion of zero- to nonzero-polarity tweets in Figure 4. Interestingly, the percentage of each data set occupied by zero-polarity tweets is not uniform. In particular, Reuters, BBC (World) and The Washington Post have significantly more neutral tweets than the other news sources. I split the data sets into COVID/non-COVID in Figures 5a and 5b to see if there are any differences between the two subsets. For most news sources, both subsets tend to agree in their proportions. However, for Reuters and The Associated Press, 6.2% and 10.7% less tweets, respectively, were neutral in their about COVID subset.

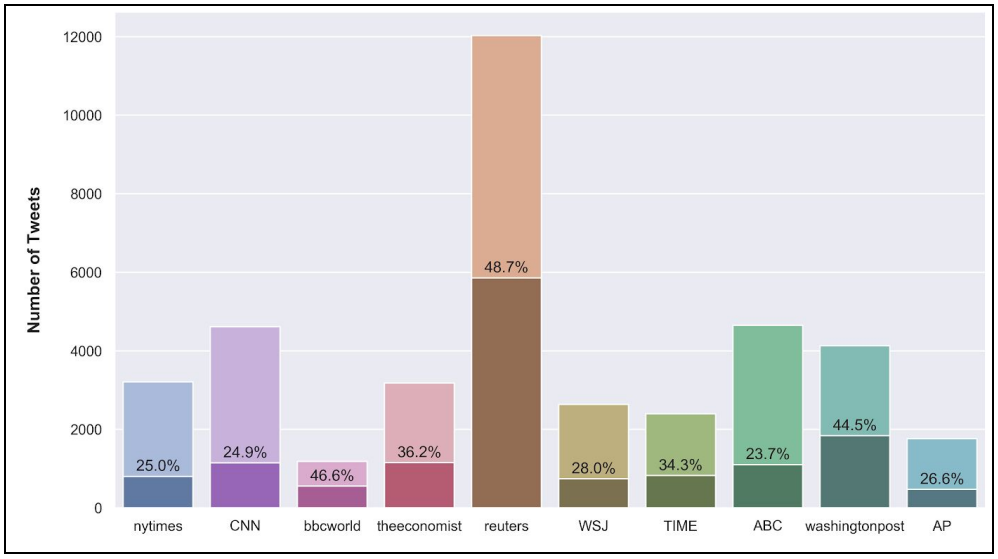

Figure 4: The proportion of zero- to nonzero-polarity tweets. The darker portion denotes the zero-polarity tweets and the percentage specifies what portion of each data set is zero-polarity.

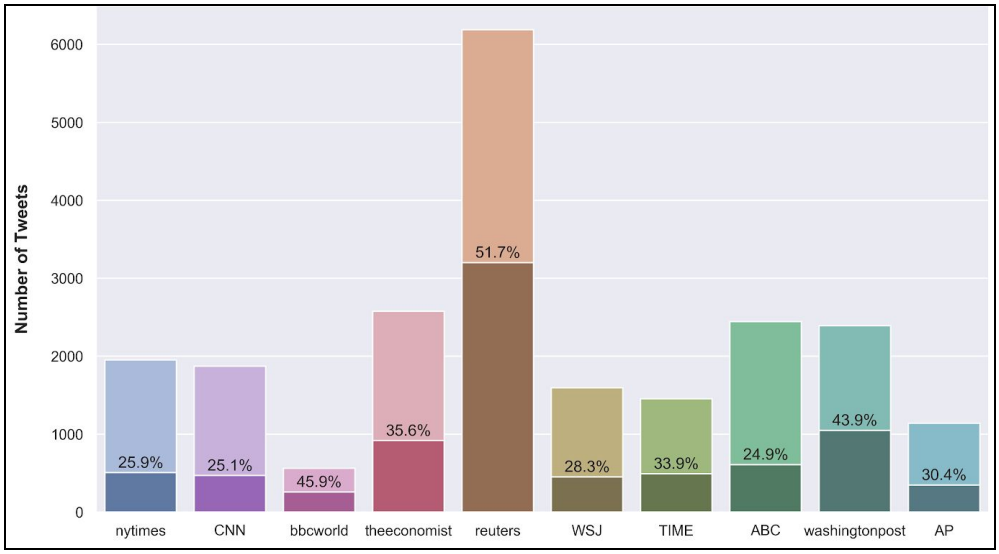

Figure 5a: The proportion of zero- to nonzero-polarity tweets in subset not about COVID.

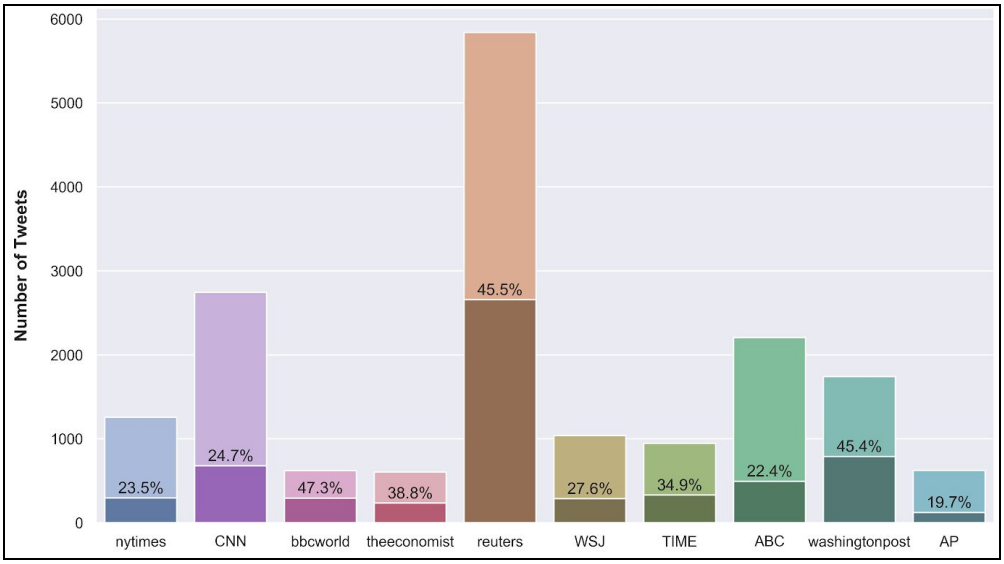

Figure 5b: The proportion of zero- to nonzero-polarity tweets in subset about COVID.

When looking at the breakdown between positive, negative, and neutral tweets, the story is similar (Figure 6). For all news sources, nonzero-polarity tweets are much more often positive than negative in polarity. This is particularly true for The New York Times, CNN, ABC News, and The Associated Press, where over 50% of tweets are positive and positive tweets represent more than twice as much of the data sets as negative tweets.

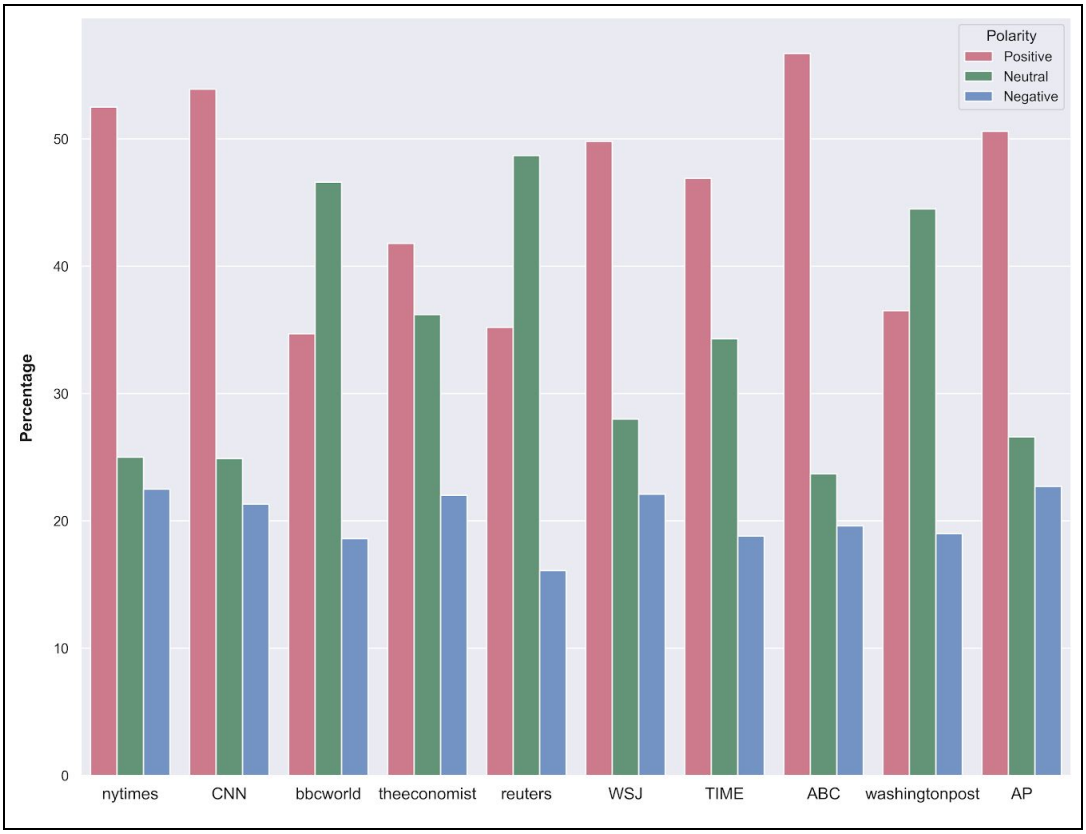

Figure 6: The breakdown of positive, neutral, and negative polarity tweets for each news source.

Next, I examine polarity trends over time in Figure 7. To smooth out the lines, I downsample the data using three-day averages polarities. It is worth noting that polarity variations on this figure are quite small: most sources' averages stay within a polarity range of 0.1. Surprisingly, none of the sources dip into the negative polarities, although BBC (world) comes close about halfway through the time period. While still quite close to neutral, it is interesting that all sources are very slightly positive for the entire time period.

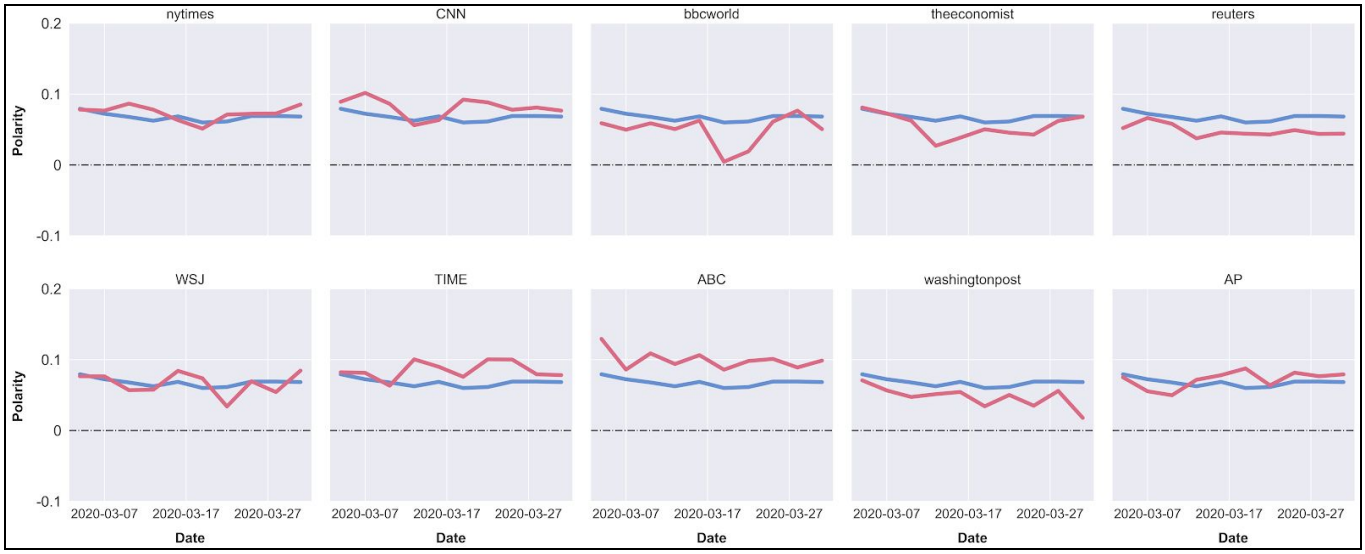

Figure 7: Polarity trends over time for each news source. The red line shows the three day averages for the given news source, while the blue line shows the average polarity across all news sources at each time point. The dotted line shows where zero is.

In Figure 8, I split the line plots from Figure 7 by COVID/non-COVID tweets, using the same downsampling approach. Even with the downsampling, the lines are relatively erratic with little to no discernible patterns. For most news sources, both subsets exhibit similar polarity trends. In the plot for TIME, however, there is a noticeable difference between the two lines, especially in the first few weeks of the time period. Also, the non-COVID subset for BBC (World) dips into the negative polarities for a single time thing halfway through the period, a phenomenon not seen in any of the other SA line plots.

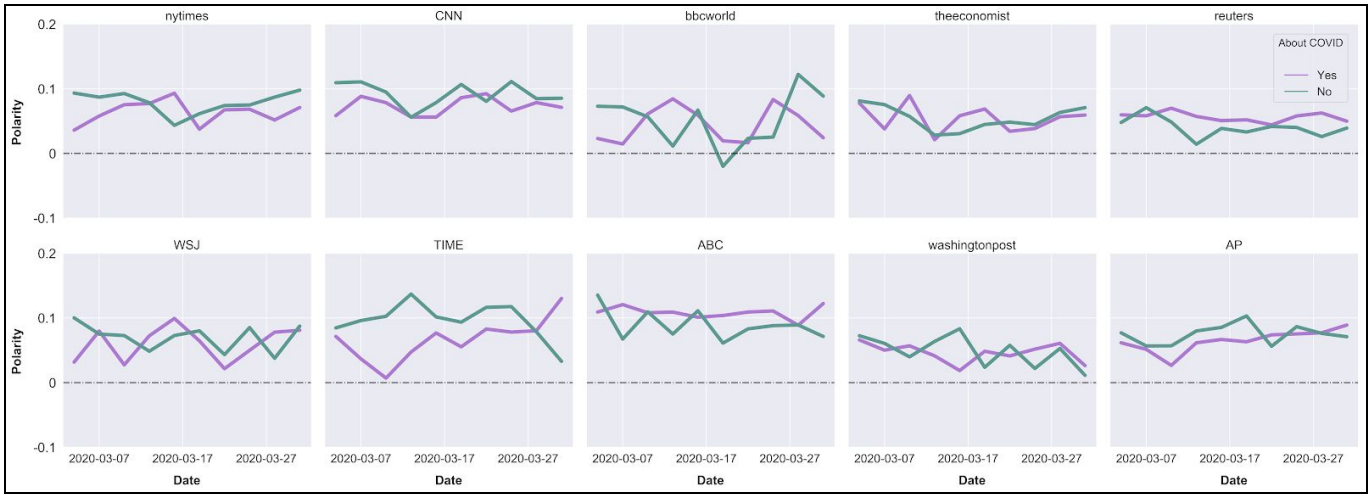

Figure 8: Polarity trends split by COVID/non-COVID. Again, three-day averages are used in both lines. The dotted line shows where zero is.

To better understand the extent to which tweets in this time period are dominated by COVID coverage, I plot the percent of tweets that are about COVID in Figure 9, using the same downsampling method as in Figures 7 and 8. It is easy to see that certain news sources published consistently more tweets about COVID, CNN and BBC (World) in particular. In contrast, The Economist published significantly less tweets about COVID, never getting above about 30% (the starting value for many news sources). Despite these differences, each news source saw increases in the percentage of tweets published about COVID when comparing the beginning of the time period to the end.

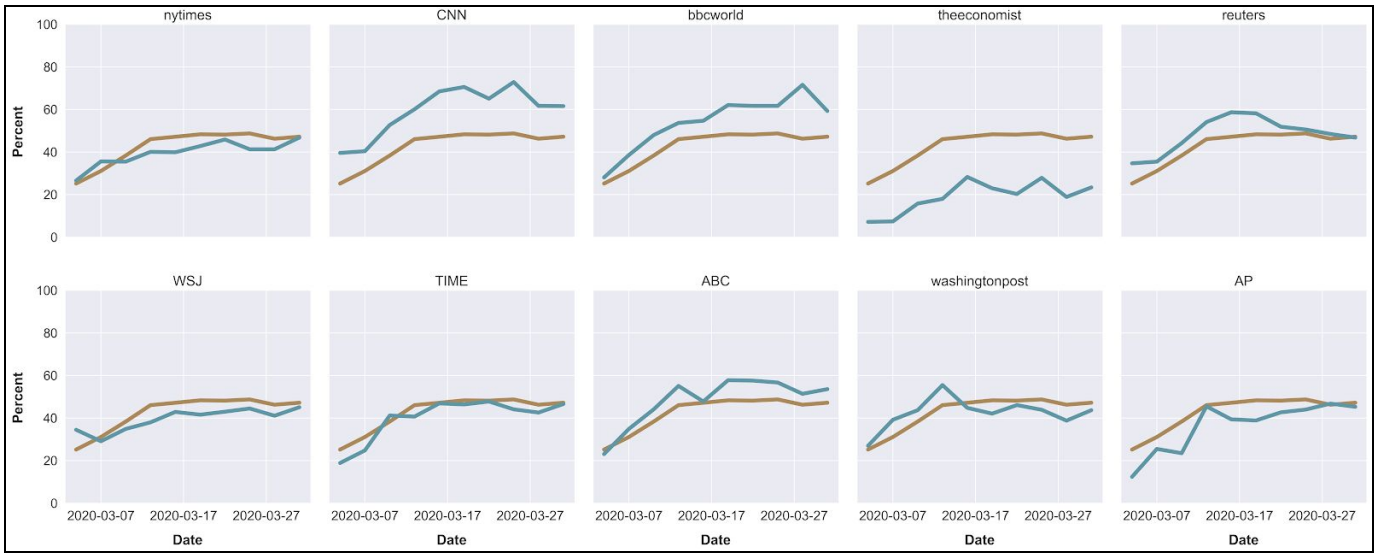

Figure 9: The percent of articles about COVID for each source. Again, three-day averages are used. The blue line shows the trends for each news source, while the brown line shows the average trend.

3.2 NER Results:

Next, I look at the results from the NER and demonstrate how the two forms of analysis could be used together. To begin, I explore the tag breakdown in the overall data sets in Figure 10. Most sources have relatively similar tag breakdowns, although certain sources seem to favor certain types of NEs. For instance, Reuters and BBC (World) reference "GPE" NEs more than any other news sources, whereas ABC News seems to have a tendency towards referencing NEs with the "PERSON" tag. The "OTHER" category contains the tags that appeared infrequently, which are: "EVENT," "PRODUCT," "QUANTITY," "LANGUAGE," "LAW," "FAC," "ORDINAL," "WORK_OF_ART," "PERCENT," "LOC," "TIME," and "MONEY."

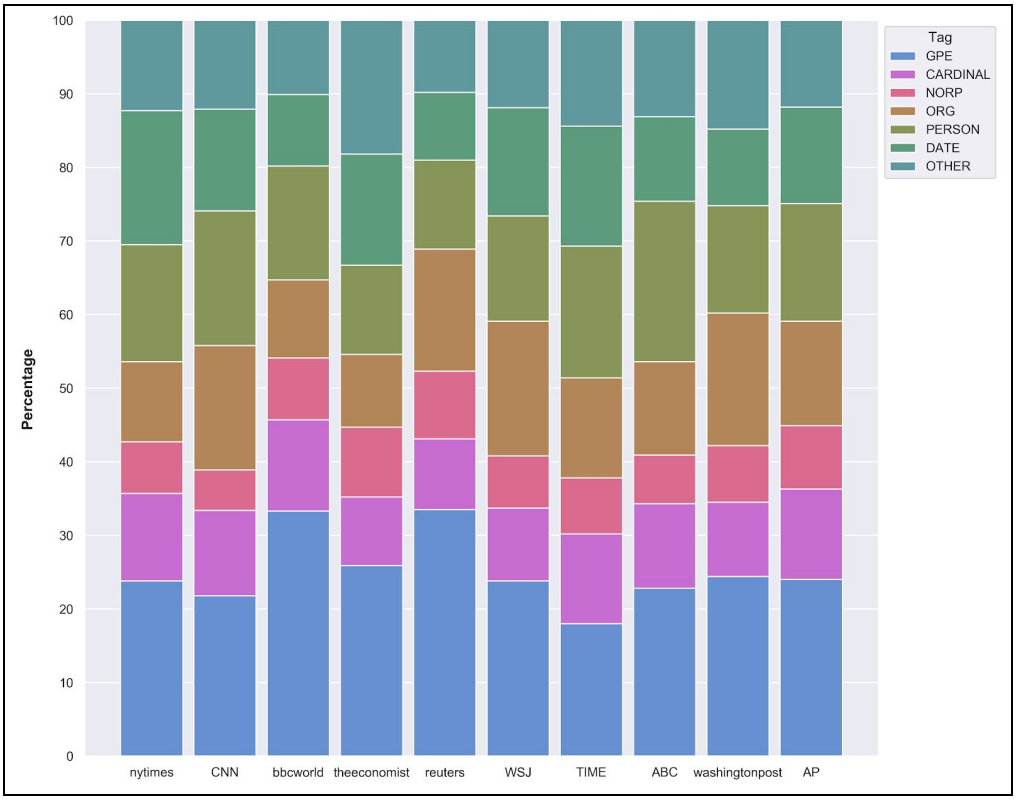

Figure 10: The tag breakdown as categorized by spaCy's default NER system for each news source. "GPE" is "countries, cities, states"; "CARDINAL" is "numerals that do not fall under another type"; "NORP" is "nationalities or religious or political groups"; "ORG" is "companies, agencies, institutions, etc."; "PERSON" is "people, including fictional"; "DATE" is "absolute or relative dates or periods." The "OTHER" category contains the 12 tags that appeared the least frequently (less than 5% in eight or more sources).

While Figure 10 shows overall trends, Figure 11 once again splits by COVID/non-COVID tweets, but no longer splits by news source. The difference in "GPE" tag frequency immediately jumps out: from non-COVID to COVID tweets, "GPE" tags increase by 7%. Contrast this with the similar decrease in "PERSON" tags from non-COVID to COVID tweets of 6%.

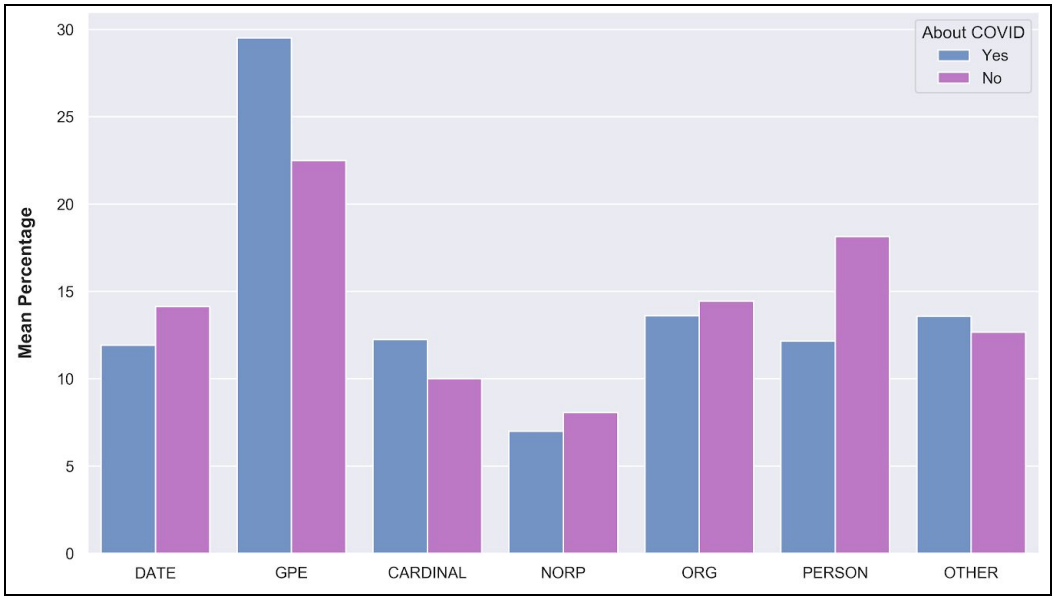

Figure 11: Tag breakdowns in news sources overall, split by COVID/non-COVID tweets.

3.3 Combining NER and SA:

To demonstrate how SA and NER results might be used together, I focus on "PERSON" NEs and look at how the perception of frequently referenced public figures shifts over time. First, I pull out the top ten "PERSON" tags for each week. Figure 11a shows the overall top tens whereas Figure 11b is only COVID tweets. In both figures and in almost every week, Donald Trump dominates the top ten. The only exception is the first week of Figure 11a, which was the week of Super Tuesday.

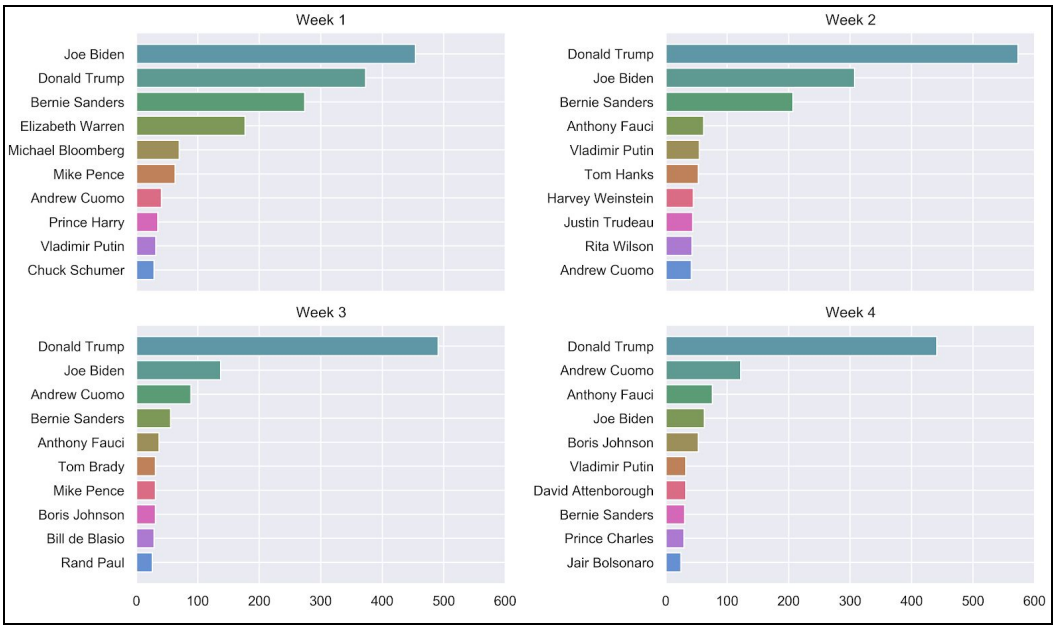

Figure 11a: The top ten most referenced "PEOPLE" NEs for all tweets.

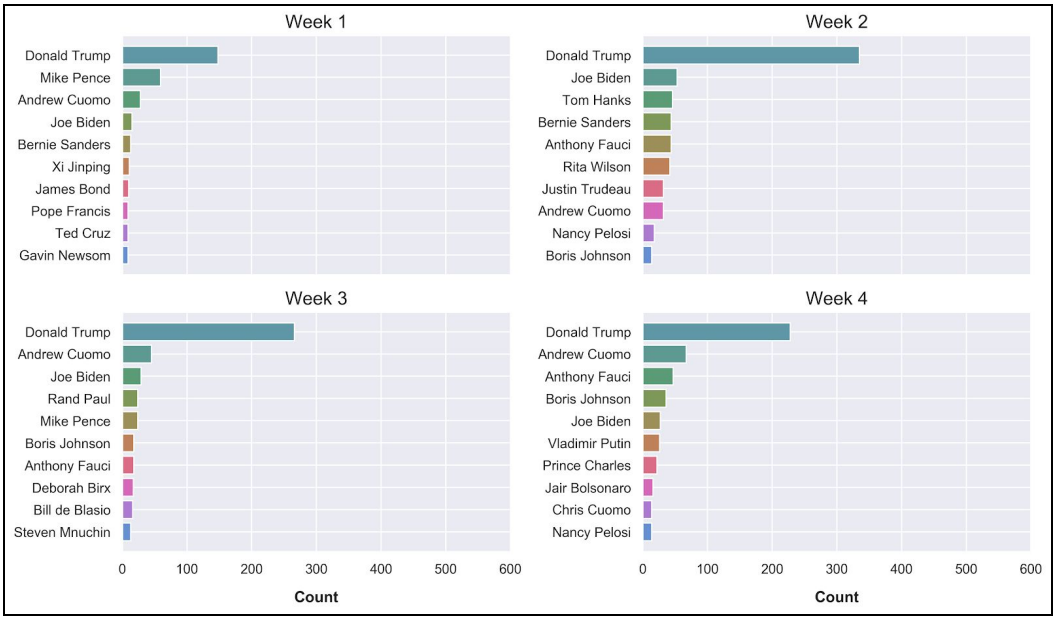

Figure 11b: The top ten most referenced “PEOPLE” NEs for COVID tweets.

Figure 11c shows the same information as Figure 11b, but splits NEs by gender. In the top ten most frequently mentioned “PERSON” NEs for COVID tweets, the ratio of mentions of men to women is about 18 to 1 (removing Trump from the calculation still leaves the ratio at 8 to 1).

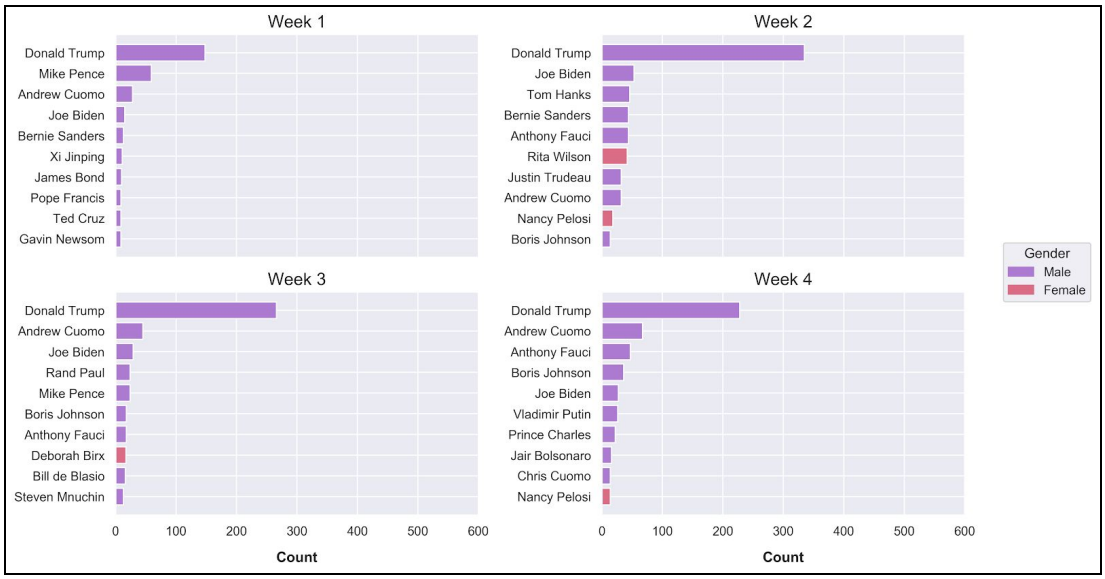

Figure 11c: Figure 11b, split by gender.

The five most frequently mentioned "PERSON" NEs for COVID tweets across the time period are: Donald Trump (977 mentions), Andrew Cuomo (172), Joe Biden (124), Anthony Fauci (115), and Mike Pence (103). In Figure 12, I look at the polarity of tweets containing these top five NEs over the entire time period. Again, I use the same downsampling approach as in Figures 7-9. While the polarities of tweets about Trump stay stable and close to neutral, the polarities for Fauci and Pence are quite volatile, even dipping into negative values at points. It is interesting that we do not necessarily see the same level of polarity consistency that we saw in Figure 7. I discuss these results further in Section 5.2.

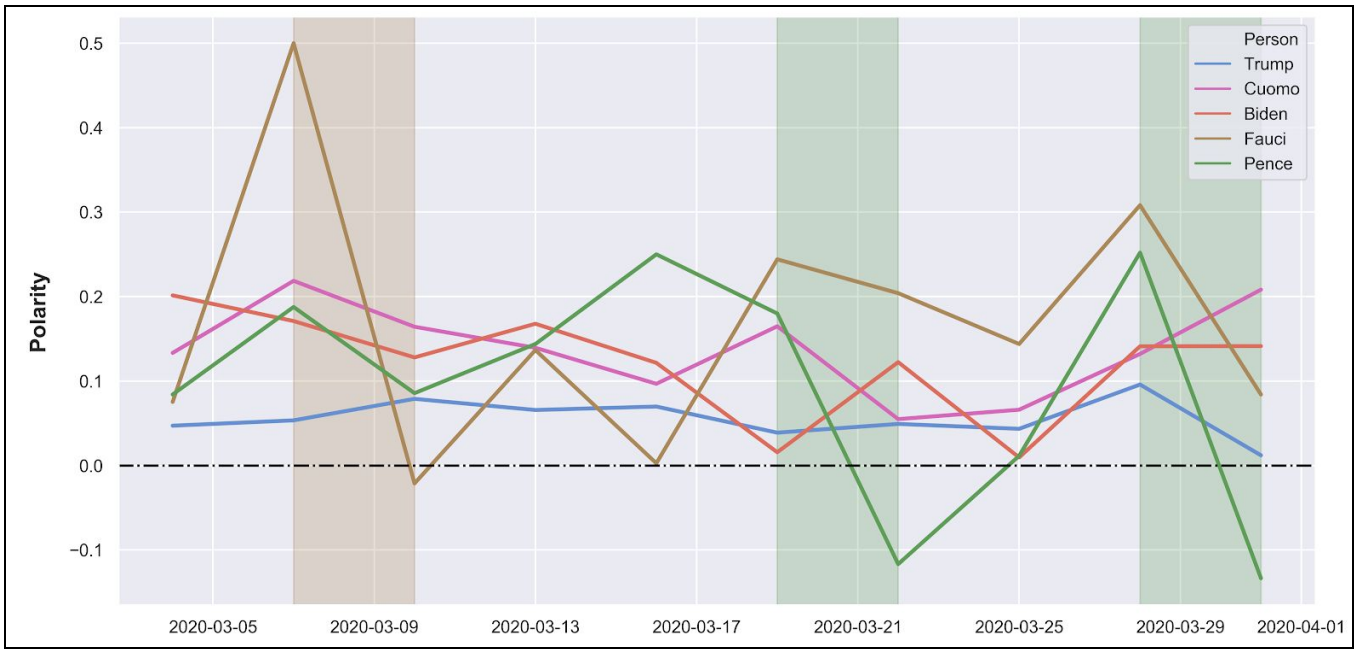

Figure 12: The polarity of the top five NEs. The dotted line shows where zero is. The shaded regions highlight major polarity shifts.

4. Discussion

In this section, I discuss the results from Section 3 and recommend further lines of research.

4.1 SA:

One of the most surprising results from the SA is the tendency towards positive polarities (Figure 7). While these positive polarities do tend to be quite close to neutral, the slight positivity during a somewhat bleak period is unexpected. Collecting a more representative sample of tweets over a longer period would help to place this trend into context, as it is possible that tweets from news sources tend to be positive in general. Perhaps the trends in Figure 7 are actually lower than average, in which case COVID might be having a net negative effect on sentiment. Along the same lines, the fact that certain news sources' ratio of neutral to non-neutral

tweets did not align when comparing COVID/non-COVID subsets may suggest a slight bias in COVID coverage. However, extricating the limitations of the lexicon-based approach (as covered in Section 5.2) from true results is an important step towards more accurate findings. A more exhaustive comparison between news sources with respect to polarization might also yield unique results. Just by Figure 7, it is already clear that not all sources maintain similar levels of polarity (i.e., ABC News tends to be more positive than average whereas The New York Times tends to be more negative). Finding the situations where news sources most agree or disagree would contribute to the general understanding of how news media can impart unifying or discordant messages during periods of crisis.

While it is difficult to discern any particular patterns in Figure 8, the simplicity of the categorization method for sorting into COVID/non-COVID tweets may be having a confounding effect (i.e., the differences are minimized between categories because of high rates of misclassification). Or, maybe the patterns are more nuanced, obliterated by the crude three-day averages. High polarity tweets do exist within the data sets (Table 3), but are comparatively uncommon. A study focused solely on these high polarity tweets (i.e., above +0.5 or below -0.5 in polarity), with an emphasis on characterizing the types of tweets that tend to be high polarity, might reveal the intentions of the different news sources.

In general, the SA results that I have explored here lead into more robust methods, such as automated content analysis, which could drive further research focused on framing and bias. As participatory agents during crisis situations, news media can influence public opinion, so it is important to understand how different news sources approach crises through coverage. More specifically, gaining a greater understanding of the role of news media in disseminating crisis-related information through the study of COVID coverage could bolster underlying theory within crisis informatics.

4.2 NER:

In the exploration of NE tag breakdowns, the shift from an emphasis on “PERSON” tags to “GPE” tags is unexpected. Perhaps this reflects the international effect of COVID, which would explain the increase in “GPE” tags (countries, states, etc.). It is intriguing that “PERSON” tags are less prominent in COVID tweets, though. A further breakdown of the roles of “PERSON” NEs might shed light on shifts in ideals during periods of crisis (i.e., less emphasis on celebrities in exchange for more references to figures of authority and political figures). To draw more valid comparisons here, a better baseline for non-COVID tweets would be needed, which is addressed in Section 5.1.

In COVID tweets, men are mentioned at a far higher rate than women. Again, a better baseline for comparison would help here, as it’s possible that this just picks up on inherent sexism within news media coverage or unbalanced representation with public figures in general. Moreover, many of the top ten most mentioned public figures are on the US COVID task force, which is overwhelmingly occupied by men. In order to extricate the effect of COVID on gender

bias in coverage, a more international sampling of news sources would also be necessary (seven of the ten sources are from the US, Table 1). Still, this seems like a promising area for further research.

4.3 SA and NER:

By putting together the results from SA and NER, I hope to show how several methods of analysis could be used together to begin to answer more profound questions. The results shown in Figure 12 are certainly preliminary, but show some promise. I do worry that the volatility in the polarities for Fauci and Pence may be due to relative data sparseness. Both Fauci and Pence have the least amount of mentions among the figures followed in Figure 12, which could indicate that the shaded regions are not entirely representative (i.e., maybe there were only a few tweets for either of them on a given day and from only a couple of unique news sources). However, Cuomo and Biden don't have that many more mentions than Fauci or Pence, so this may not actually be responsible for the volatility. Collection from a larger number of sources and over a longer time period could help to remedy this issue of sparseness, if it is the culprit in this instance.

Systematically correlating the shaded regions with specific actions or events related to Fauci or Pence is difficult. Both are spokespeople on the COVID task force, so polarity shifts pertaining to either of them could be attributable to changing opinions of the task force as a whole. To try to isolate the direct sentiment towards figures, a more effective approach might be to identify and measure the words most associated with each figure using word collocations. Measuring the polarity of the entire tweet is inexact and could potentially confound results. At the same time, tweets are quite short which severely limits the amount of text data to work with. I discuss the possibility of bringing in article data in the next section.

Focusing on the pairwise correlation or co-occurrence of common "PERSON" NEs within tweets could also reveal unanticipated patterns. As prominent members of the same presidential administration, I am surprised to see that Pence and Trump's polarity trends don't match up more.

4.4 Miscellaneous:

When I was originally planning out this project, I had hoped to collect full article data in addition to Twitter data. Many tweets from news sources link to a corresponding article, which opens up the possibility for comparison between a news source's presence on Twitter and on their native site. In the end, I was not able to include this in my project. Using python web scraping libraries like beautifulsoup, this could be an insightful extension to this project. It would be interesting to see if the messages presented on each platform agree. Given the deluge of COVID-related coverage and high levels of uncertainty, differing messages across platforms might exacerbate concerns with an already fraught situation. Additionally, working with full article data would be easier than with Twitter data. While the volume of text data would be much

greater, most NLP tools are well-suited for the structured nature of article data and would perform much better with little need for intervention.

While working on data cleaning and preprocessing, I noticed that some sources tended to repeat tweets at a high rate. As can be seen in Table 2, TIME and The Economist are especially guilty of this. I decided not to do anything special with duplicated tweets, but taking a closer look specifically at these tweets that are duplicated (identical text) or high in token similarity could produce interesting results. For calculating token similarity, a simple bag of words approach would likely be sufficient (many of the high similarity tweets only differ in text by a few tokens). Since repeating tweets is a conscious decision on the part of a news source, examining this subset might reveal certain hidden values for a given source (i.e., the stories that they most want to emphasize or biases in coverage).

5. Limitations and Future Directions

Because of the time constraints for this project, I had to make many research design concessions. In this section, I acknowledge the limitations of the work carried out in this project and discuss possible improvements. Additionally, I recommend several potential future directions for this area of research.

5.1 Data Collection and Truncated Data:

One of the biggest challenges in this project was the lack of a strong baseline. Throughout the project, I compare COVID and non-COVID tweets, but even non-COVID tweets are from a time period where COVID dominated the conversation. As such, using these non-COVID tweets as the baseline for comparison is potentially skewed. To best understand how news media has participated in the public dialogue surrounding COVID, a baseline that is not during COVID would be needed. In CITATION NEEDED (Muslim SA article), AUTHORS use a robust 20-year baseline. While this exact approach is not possible (Twitter is only 14 years old), the comparisons that I make here would be much more valid if a few years of tweets from news sources were used as a baseline.

For a long time, Twitter was known for its 140-character limit for tweets. In 2017, Twitter moved to a 280-character limit. Even with the longer character limit, most users do not tend to publish tweets above the 140-character limit.[CITATION NEEDED] This is unfortunately not the case for the news sources studied in this project. Figure 13 shows the distribution of character lengths for tweets from each news source. Some sources, such as Reuters and The Washington Post, do tend generally publish tweets of less than 140 characters. But for many other sources (CNN, The New York Times, The Associated Press, ABC News), the distribution is much flatter, with a significant number of tweets over 140 characters.

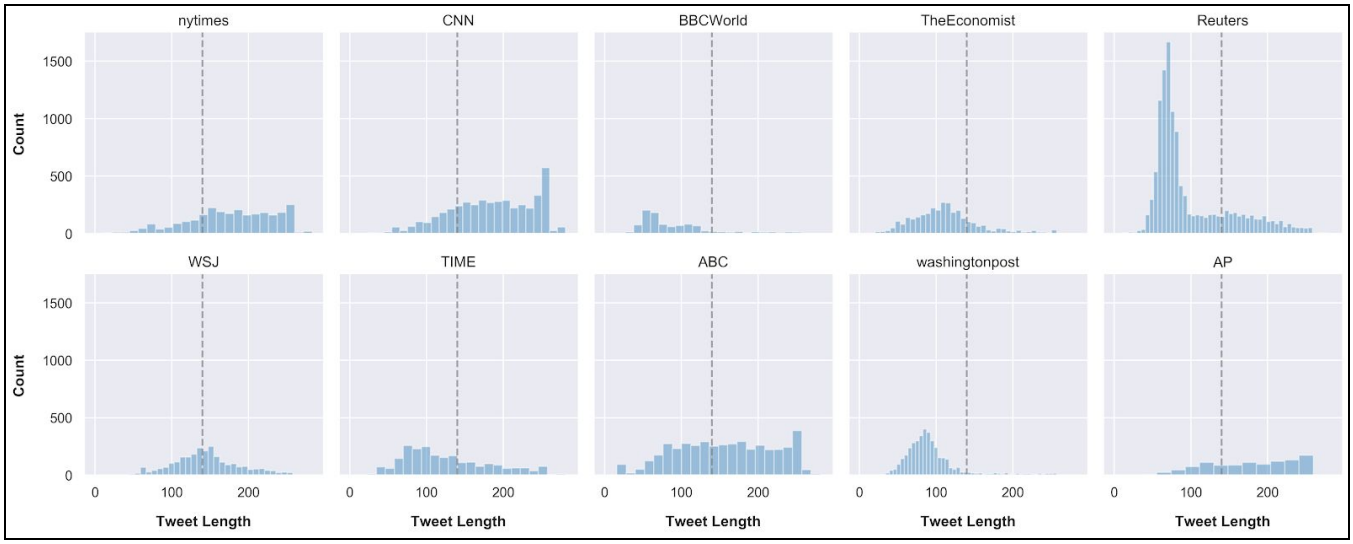

Figure 13: The distribution of character lengths for tweets from each news source. The dotted line shows where 140 characters is on the x-axis.

By default, the Twitter API truncates tweets longer than 140 characters in returned JSON files. When I collected my data in March, I made sure to request full text tweets when querying the API. However, this did not stop retweets from being truncated, which I did not realize until much later. Unfortunately, it seems at this time that there is no functionality included in the Twitter API for retroactively collecting non-truncated tweets. I then had to decide whether or not to keep these truncated tweets in the data set. To get an idea of the character losses that I was experiencing due to truncation, I calculated the mean character length of tweets over 140 characters for each news source and used that value to approximate the percent loss (using median character lengths produced nearly identical results). I show the percent losses due to truncation for each news source in Figure 14. Aside from The New York Times, BBC (World), and The Associated Press (the three sources with the highest percent retweets, seen in Table 2), character losses are quite low. Overall, the data set experienced a 2.8% character loss due to truncation. In the end, I decided to keep in truncated tweets. I also fixed the data collection script, which should now collect only non-truncated tweets.

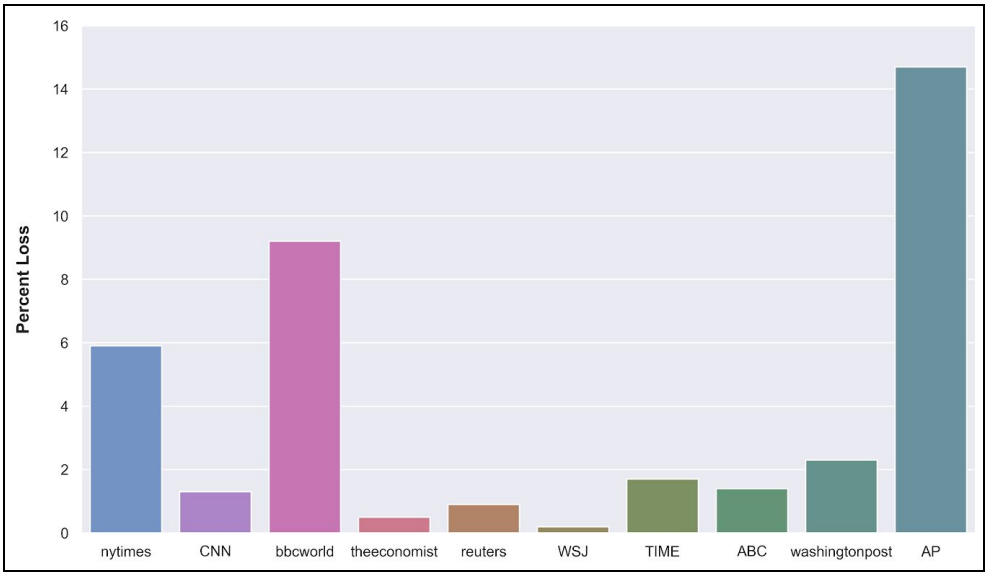

Figure 14: Expected character losses for each source, calculated using the mean character length for non-truncated tweets over 140 characters.

5.2 Analysis Methods:

My design philosophy for this project was focused on using efficient implementations of relatively simple analysis methods, which TextBlob and spaCy fit with well. However, by choosing efficiency over performance, results may have been affected in a number of ways. Here, I discuss alternative approaches to preprocessing, SA, and NER that might produce better results.

Because of the lack of literature focusing on news media on Twitter, it was difficult to make justified preprocessing decisions. For the most part, the preprocessing that I pursued for my two methods seemed sound and did not affect performance in initial tests, but of course this could all benefit from more rigorous testing as in CITATION NEEDED (preprocessing for SA). Although I do not think lexical normalization is particularly necessary for this text data, dealing with emojis and "@"/"#" in a more nuanced manner might improve the accuracy of downstream analysis methods.

The Association for Computational Linguistics (ACL) has held annual workshops for the past five years dedicated to improving the performance of NLP tasks when applied to noisy user-generated texts (NUT).[LINK TO WNUT SITE] Many of these workshops included a shared task for NER in NUT, some even focusing on text data from Twitter. Conditional random field (CRF) and long-short term memory approaches (LSTM) found success in the 2015 and 2016 tasks, respectively. In large part, the need for more robust NER methods for Twitter data stems from the novelty of NEs and the unstructured nature of the data. It is possible that LSTM or CRF approaches would improve accuracy, but given the semi-structured nature of the data used in this project, I expect that less robust methods may still perform well. For instance, by inspection of initial testing results, I noticed that the python library Stanza performed better than spaCy in

identification and tagging of NEs. CITATION NEEDED (testing NER systems) compares well-established NER systems using hand-annotated documents, ranking by F_1 score. A similar approach towards testing systems on semi-structured data would be valuable. Knowing whether Twitter data from news sources is more similar to well-structured text data or general unstructured Twitter data would help guide further design choices in this area of research.

As with NER, it has been shown that certain SA approaches perform better than others when applied to Twitter data. CITATION NEEDED (comparing SA for Twitter) shows that lexicon-based systems for SA perform significantly worse than machine learning or hybrid approaches. CITATION NEEDED (inducing lexica) introduces a method for the induction of specialized lexica using unlabeled corpora and domain-specific word embeddings, which might be a good approach for this context. In both approaches, however, more text data is required than is used in this project (CITATION's approach performed well with corpora larger than 10^7 tokens). Also, a supervised machine learning approach would require additional efforts to curate accurate hand-annotated training data. The unsatisfying reality is that more research is needed here as well. Again, I would like to emphasize that more work towards characterizing this semi-structured data would help minimize the amount of work needed to find the most suitable approaches.

In particular, I wonder if the lexicon-based approach might be responsible for the high proportion of zero-polarity tweets. Given the large number of tokens that are not part of the lexicon used for TextBlob's SA implementation[LINK TO LEXICON], it is possible that there is a skew towards falsely neutral results. It is also possible that tweets are, on the whole, relatively neutral (or quite close to neutral). Once again, further research is needed to see whether or not this result holds under different approaches.

5.3 Tweet Classification:

The approach that I took for determining whether or not a tweet was about COVID is extremely naive, which may affect certain results. To improve classification accuracy, a machine learning approach might be better suited for the task. Drawing on the multitude of previous research on word vectors, a semantic similarity approach may also be more successful. More formal testing would be required to confirm the superiority of these methods over my naive classification, but I suspect that both would achieve much higher recall and higher F_1 scores overall.

References

Additional Materials

1. NER Tags for spaCy

These are the tags and descriptions supported by spaCy's NER implementation, taken directly from spaCy's documentation.

- "PERSON": people, including fictional.
- "NORP": nationalities or religious or political groups.
- "FAC": buildings, airports, highways, bridges, etc.
- "ORG": companies, agencies, institutions, etc.
- "GPE": countries, cities, states.
- "LOC": non-GPE locations, mountain ranges, bodies of water.
- "PRODUCT": objects, vehicles, foods, etc. (Not services.)
- "EVENT": named hurricanes, battles, wars, sports events, etc.
- "WORK_OF_ART": titles of books, songs, etc.
- "LAW": named documents made into laws.
- "LANGUAGE": any named language.
- "DATE": absolute or relative dates or periods.
- "TIME": times smaller than a day.
- "PERCENT": percentage, including "%".
- "MONEY": monetary values, including unit.
- "QUANTITY": measurements, as of weight or distance.
- "ORDINAL": "first", "second", etc.
- "CARDINAL": numerals that do not fall under another type.

2. TextBlob SA Lexicon:

The lexicon used in TextBlob's SA approach can be viewed at the TextBlob GitHub page.

3. Code and Data:

The GitHub repository "emiliolr/news-covid" contains much of the code used for this project as well as the dehydrated tweet data sets.