

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 early stages of the coronavirus pandemic (COVID). In a period fraught with high uncertainty and strict social distancing orders, news outlets provide 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, I explore a novel data set of 39,791 tweets from ten major news sources published during the month of March 2020. Examining the results, I find that news sources tend to differ in their sentiment and inclusion of named entities, sometimes substantially. While tweet polarities tend to be relatively consistent and slightly positive across news sources, extracting tweets that follow public figures reveals more volatile trends with larger ranges of polarities. I recommend several future directions for research, emphasizing the value of a stronger baseline for comparison, data collection over a longer period and from more sources, and implementations more suited to the semi-structured nature of this particular data. This work contributes to the field of crisis informatics by taking the first steps towards describing the role of news media during a period of international crisis.¹

1. Introduction

Crisis informatics is a “multidisciplinary field combining computing and social science knowledge of disasters; its central tenet is that people use personal information and communication technology to respond to disaster in creative ways to cope with uncertainty” (Palen and Anderson, 2016). The novel coronavirus (COVID), a pandemic that has devastated countries globally, 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.

Twitter is an online microblogging platform that hosts a variety of content, with 330 million monthly active users in 2019.² When using social media as a source of text data, Twitter is often used for their comparatively liberal policy on data availability. Specifically, Twitter offers tools to access their data through a public, but still somewhat restricted, application program interface (API). Once data has been collected, techniques 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.

¹ All data and code is available at <https://github.com/emiliolr/news-covid>

² <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

In particular, crisis informatics research has often used Twitter data in conjunction with NLP techniques to explore the complex roles of organizational entities and individuals during times of crisis (Reuter et al., 2018). Studies within this area of crisis informatics research often focus on the filtering of social media data to identify and summarize accurate information, building systems to aid crisis management organizations, and describing the ways in which citizens and authorities cooperate through social media during crisis situations (Reuter et al., 2018). Unfortunately, source data in crisis informatics research is heavily 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 (Reuter et al., 2018). While I do not provide solutions to either of these problems, it is important to acknowledge this bias as a prevalent issue within crisis informatics and NLP research as a whole.

The scale of COVID 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 examine one class of participatory actors: traditional news media. It has been shown that the way in which news media chooses to frame certain issues can influence the public’s perception of those issues, especially when those frames are highly positive or negative (de Vreese et al., 2011). Furthermore, 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). 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 (Shearer and Matsa, 2018). Situational influences pertaining specifically to COVID might further enlarge this statistic. For instance, the strict stay at home orders enacted in many countries limits the potential sources of information for individuals. In another more recent Pew study, 89% of Americans indicated that they consumed news about COVID through national news outlets, with 67% for international outlets (Gottfried et al., 2020). The statistic for national news outlets matches that of public health officials (also 89%) and is higher than that of the US COVID task force (67%) or state and local elected officials (81%) (Gottfried et al., 2020). News media is easily accessible through a variety of platforms and is a reliable source of information which, during a period marked by high uncertainty, is extremely valuable.

For this project, I use the ten most followed news sources on Twitter that are not “breaking news” accounts. I decided to omit “breaking news” accounts as they rarely link to actual news articles and are thus less comparable to other news source accounts 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 is included in Table 1. In future tables and figures, I will refer to each news source by their Twitter handle as indicated by the “Twitter_Handle” column of Table 1.

| | 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 as of 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 illustrate how the data collection period fits into the chronology of COVID-related events, I include a basic timeline in Figure 1.

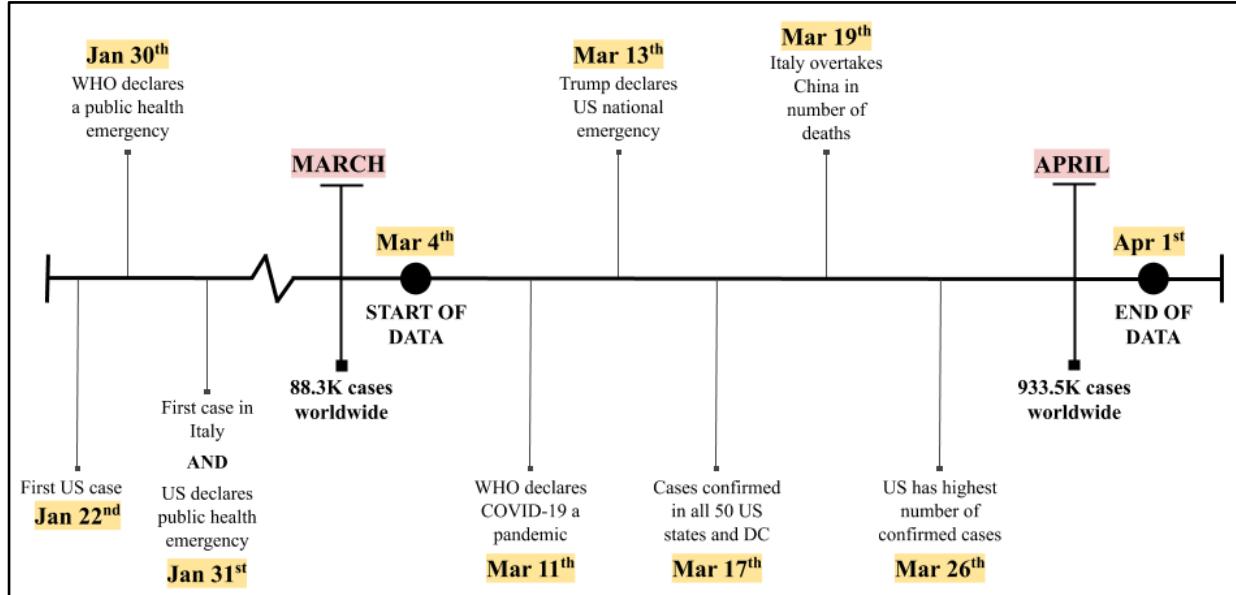


Figure 1: A timeline of key events occurring before and during the data collection period.

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 (an NLP technique that assigns a polarity

(positive, negative, or neutral) or subjectivity (subjective or objective) score to a given text), named entity recognition (statistical and machine learning methods concerned with identifying and tagging “named entities” within a text), among others. While topic modeling and content analysis approaches usually produce more interesting standalone results, both are quite computationally expensive and are generally unfit for use in data exploration. Named entity recognition (NER) and sentiment analysis (SA) are less analytically powerful methods, but both still produce compelling results and can help reveal deeper trends within text data.

SA implementations can take many forms, from lexicon-based to machine learning approaches, with some implementations combining the two, resulting in hybrid approaches (Thakkar and Patel, 2015). NER implementations typically rely on supervised machine learning algorithms trained on hand-annotated text data. While both of these NLP methods perform well on more formal text data (grammatically correct and little to no lexical variants, as defined by Han et al. (2013)), Twitter data can be a difficult application setting because of the unstructured nature of the text data. Challenges associated with using standard NLP packages for Twitter data are covered in more depth in Section 2.2.

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 phase. This analysis is not rigorous by any means, instead serving as the initial steps within the context of further research concerning COVID. In particular, I begin to describe the way in which news outlets have approached COVID through coverage on Twitter, with an emphasis on the subjects that they choose to cover (NER) and the ways in which they choose to present said subjects (SA).

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 dissemination of accurate (or inaccurate) information, etc. Additionally, the development of openly available non-English tools within NLP (i.e., 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 dealt the largest crisis of recent history; it is extremely important that we use this as an opportunity to improve theories, techniques, and systems so as to effectively understand and aid in future crisis situations.

2. Methods

2.1 Data Collection

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_Uncore | 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_Uncore” denotes the portion of the tweets that are duplicated, i.e., published more than once with identical text by the same news source.

I chose to include 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 Twitter 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 (cut off at 140 characters) 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 structured or noisy (Han et al., 2013). For instance, tweet data from general users is usually considered quite noisy as it can contain spelling or grammatical errors and often includes lexical variants such as acronyms (i.e., “lol” instead of “laugh out loud”) or intentional spelling errors (i.e., “cooooool” for emphasis instead of “cool”). Conversely, news article data would be considered structured for its syntactic regularity and lack of lexical variants. 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 irregularly to adhere to the format of a tweet (the 280-character limit, in particular). 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 were not entirely justifiable due to a lack of published literature about semi-structured data. Along the same lines, implementations within NLP are generally not well suited to semi-structured text data as they are most often optimized for 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 downstream analyses. I cover these issues more in Section 5.2.

2.3 Analysis Methods

To elucidate potential trends within the data, I applied a few simple analysis methods. I chose to use methods that were computationally inexpensive and fairly time-efficient. Given these parameters, the two methods that I settled on were NER and SA. 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 (Baldwin et al., 2015), initial tests with the NLP library Stanza demonstrated that the improved identification and tagging performance would not be worth the considerably longer processing times (I tested both Stanza and spaCy 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 Twitter data (Thakkar and Patel, 2015), the reproducibility and efficiency of TextBlob's lexicon-based approach made it a better option 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 its 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 relatively conservative. In general, I tried to apply preprocessing steps that would improve NER accuracy while still preserving the syntactic and semantic 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. I opted not to remove these references from the text entirely, as they were often structurally important and might affect spaCy's ability to detect NEs. Instead, I filtered out these references if they were detected as NEs, because they were inconsistently identified and/or incorrectly

tagged (for example: “@mattjohnson, ORG” is correctly identified as an 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).

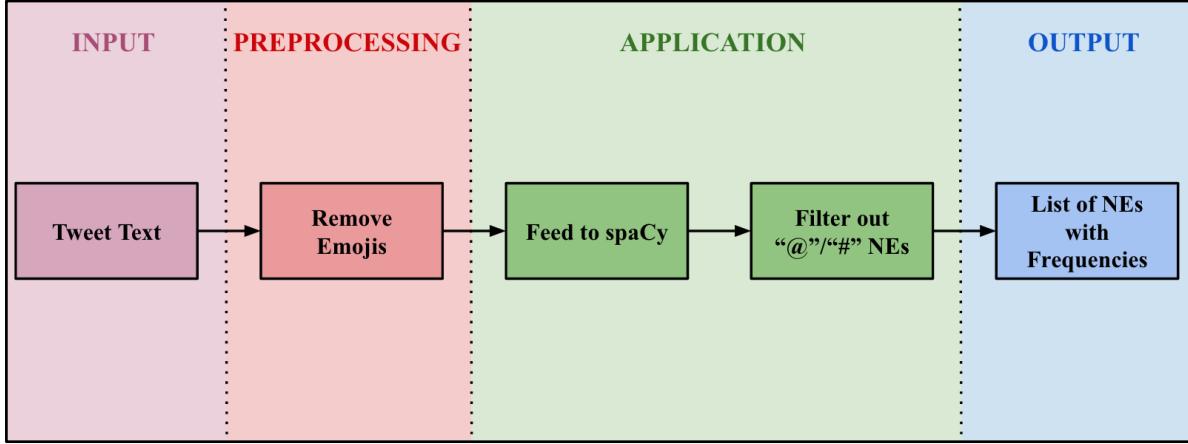


Figure 2: The full NER pipeline with each element sorted into its respective processing stage.

The necessary preprocessing steps for SA were minimal. 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 removing “@”/“#” symbols from the text and when completely removing the tokens containing “@”/“#” from the text. This indicates that “@”/“#” references are usually not in the lexicon, even when removing these special symbols. Emojis are also not in the lexicon. Ultimately, 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). 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 a “about COVID.” With regards to the statistical F_1 measure, a measure used to determine the effectiveness of a classification algorithm, 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.

3. Results

In this section, I introduce the results of the SA and NER through a number of visualizations.

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 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 less polarized tweets (+0.5 and -0.5). Next, Figure 3 shows the distribution of polarities for each news source. I use a logarithmic scale in this figure because most tweets fall close to 0 polarity, leading to a large spike in the centermost bin.

| Polarity | | Tweet_Text |
|----------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| User | | |
| 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.

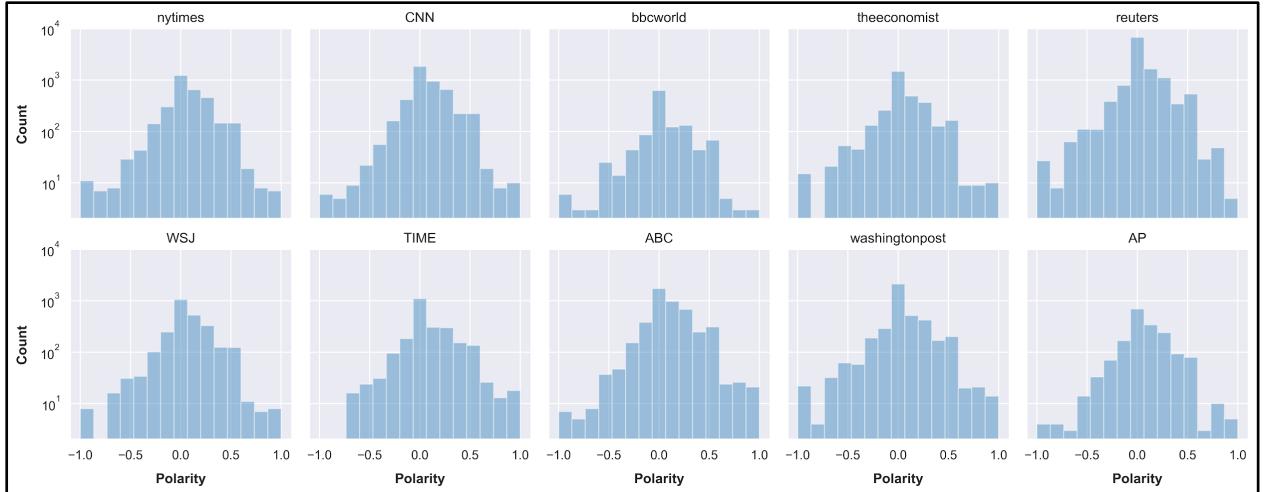


Figure 3: Polarity distributions for each news source on a logarithmic scale.

To better understand the extent to which zero-polarity tweets impact each data set, I look at the proportion of zero- to nonzero-polarity tweets in Figure 4. Interestingly, the percentage of each data set that is zero-polarity is not uniform across sources. In particular, Reuters, BBC (World) and The Washington Post have significantly more neutral tweets than the other news sources. I split the data sets by 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 COVID subset (Figure 5b).

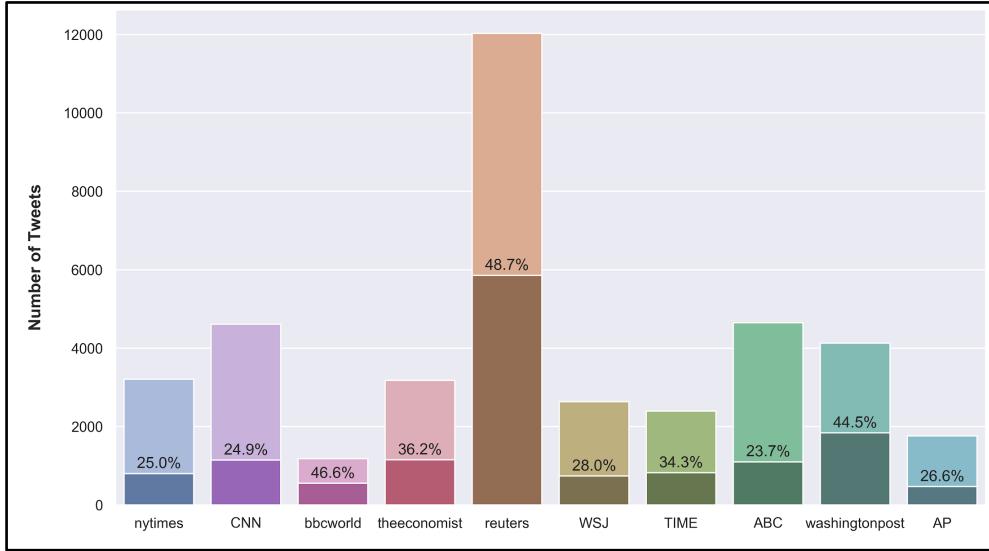


Figure 4: The proportion of zero- to nonzero-polarity tweets. The darker portion denotes the zero-polarity tweets and the percentage specifies what portion of each data set is zero-polarity.

Figure 6 shows the breakdown of positive, negative, and neutral tweets for each news source. For all news sources, nonzero-polarity tweets are 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 there are more than twice as many positive tweets as negative tweets.

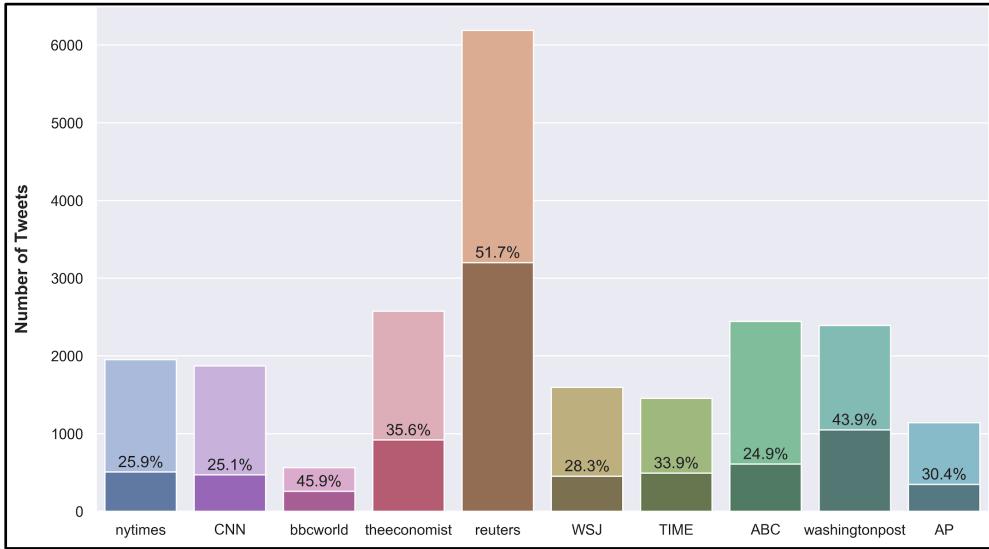


Figure 5a: The proportion of zero- to nonzero-polarity tweets in the non-COVID subset.

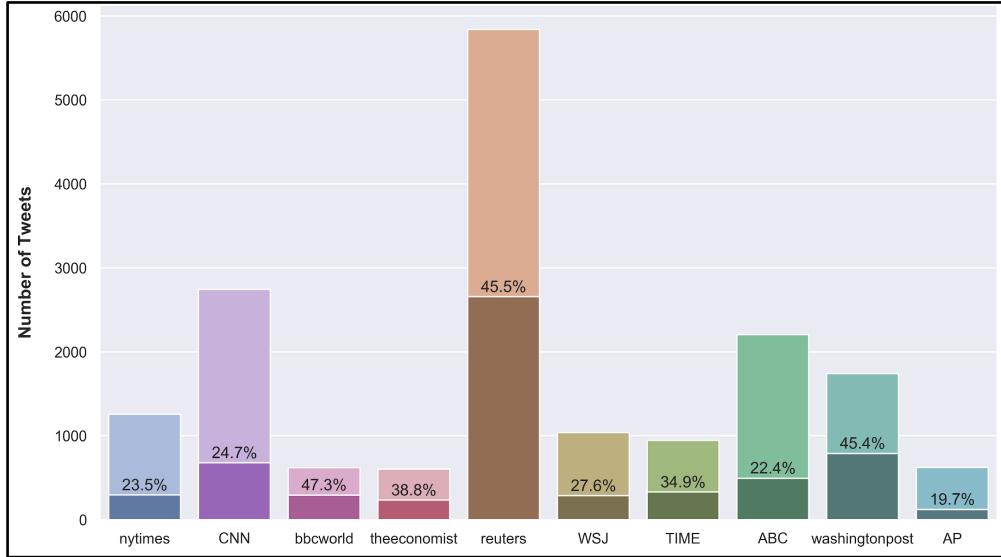


Figure 5b: The proportion of zero- to nonzero-polarity tweets in the COVID subset.

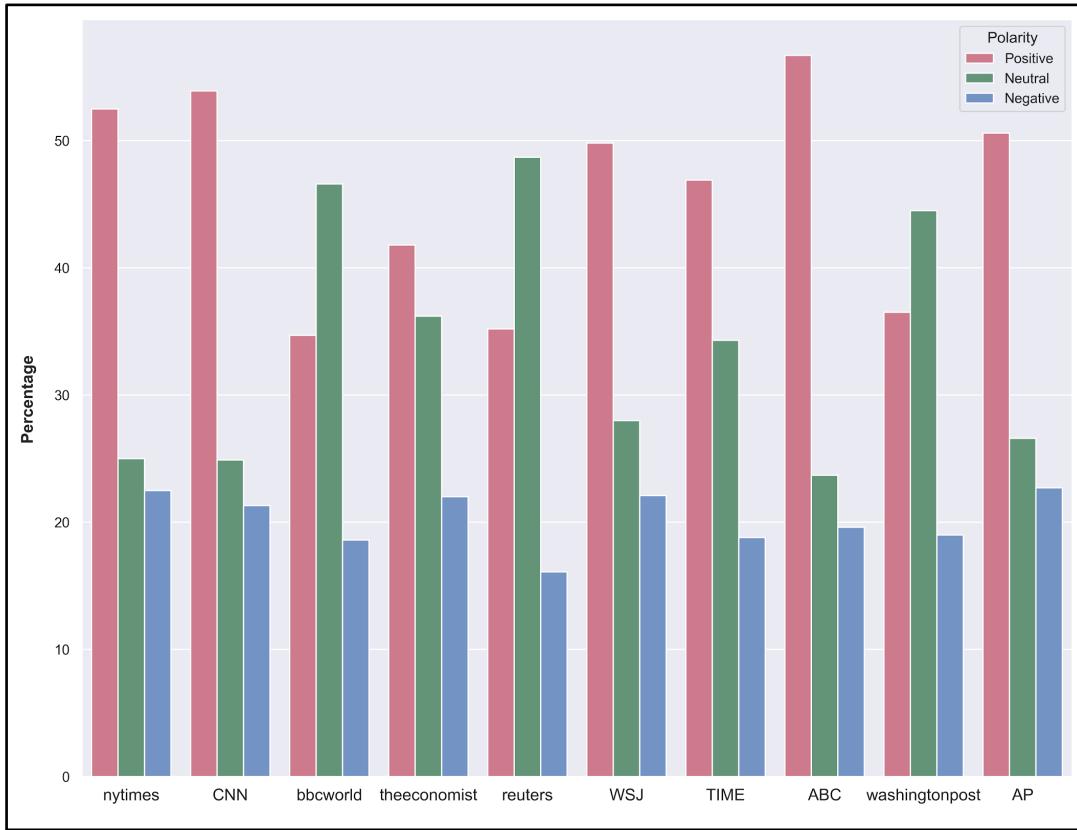


Figure 6: The breakdown of positive, neutral, and negative polarity tweets for each news source.

Next, I examine polarity trends over time in Figure 7. To smooth out the lines, I downsample the data using three-day averages. 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 on average for the entire time period.

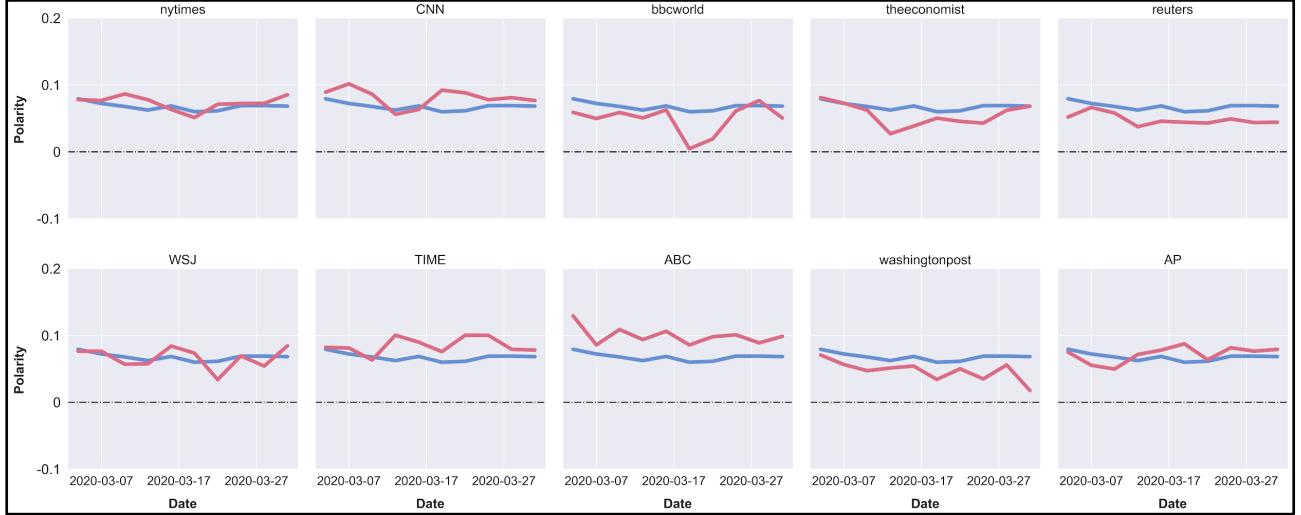


Figure 7: Polarity trends over time for each news source. The red line shows the three-day averages for the given news source, while the blue line shows the unweighted average polarity across all news sources at each time point. The dotted line shows where zero is.

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 somewhat 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 subsets, especially in the first few weeks. Also, the non-COVID subset for BBC (World) briefly dips into the negative polarities halfway through the period, which does not occur in any of the other SA line plots.

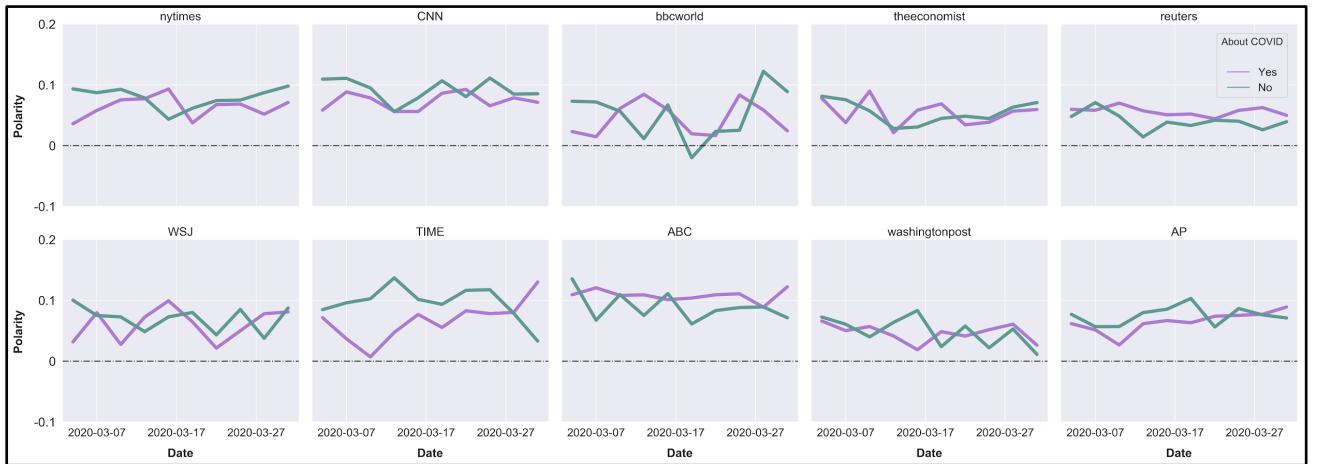


Figure 8: Polarity trends split by COVID/non-COVID. Again, three-day averages are used for both lines. The dotted line shows where zero is.

To better understand how tweet topics for each news source gravitate towards COVID, I plot the percent of tweets that are about COVID across the time period in Figure 9, using the same downsampling method as used in Figures 7 and 8. It is easy to see that certain news sources publish consistently more tweets about COVID than others (CNN and BBC (World) in particular). In contrast, The Economist published significantly less tweets about COVID, never getting above about 30%. Despite these differences, each news source saw increases in the percentage of tweets published about COVID when comparing the beginning of the time period with the end.

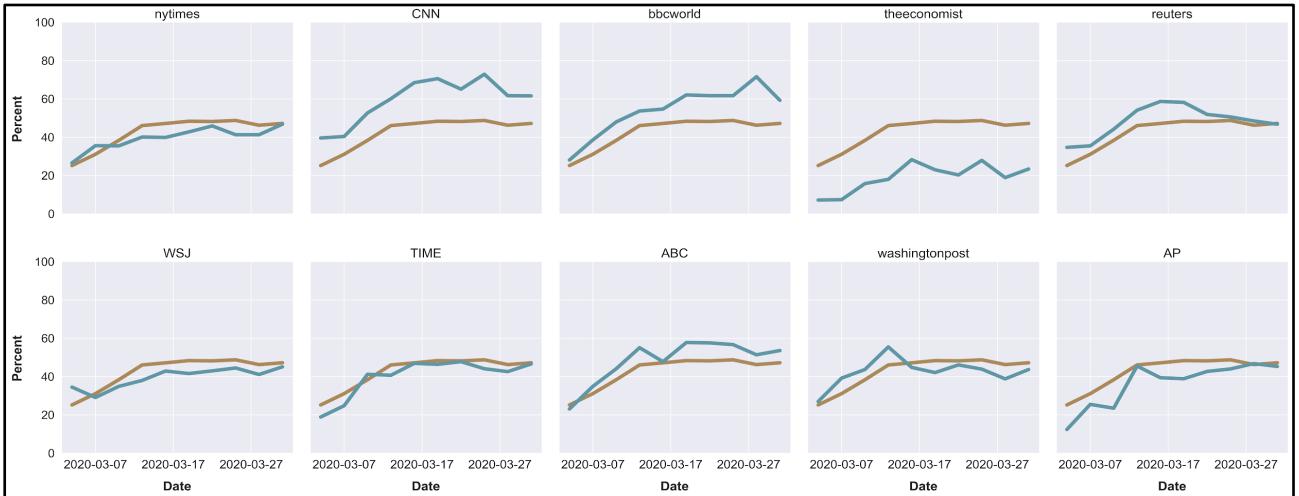


Figure 9: The percent of articles about COVID for each source. Again, three-day averages are used. The blue line shows the trends for each news source, while the brown line shows the unweighted average trend across sources.

3.2 NER Results

Next, I examine the results from the NER. To begin, I explore the tag breakdown for each news source 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” (descriptions for each of these categories are included in the Additional Materials).

While Figure 10 shows overall trends, Figure 11 once again splits by COVID/non-COVID tweets, averaging across news sources. 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%. Otherwise, the breakdowns for the two subsets are quite similar.

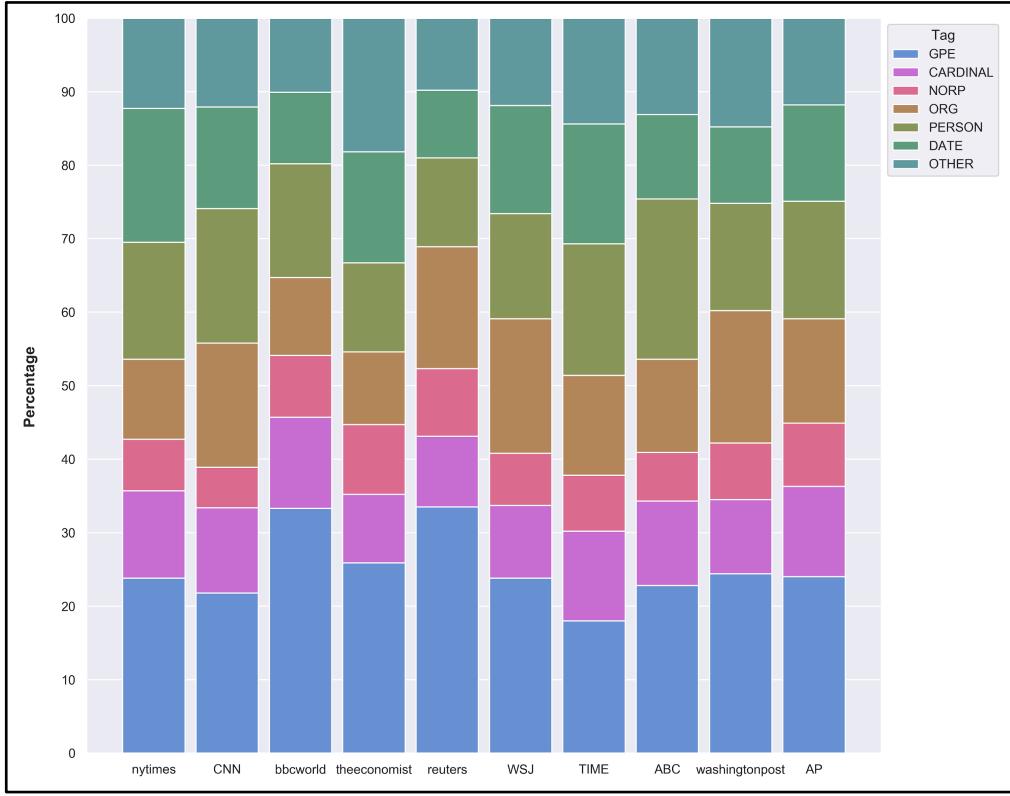


Figure 10: The tag breakdown as categorized by spaCy’s default NER system for each news source. “GPE” is “countries, cities, states”; “CARDINAL” is “numerals that do not fall under another type”; “NORP” is “nationalities or religious or political groups”; “ORG” is “companies, agencies, institutions, 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).

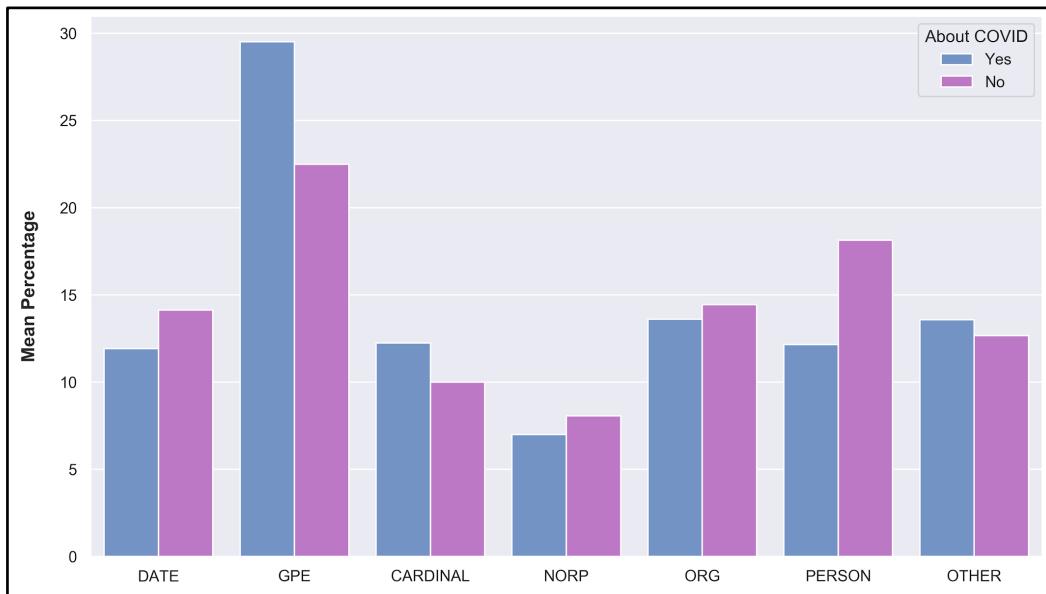


Figure 11: Tag breakdowns in news sources overall, split by COVID/non-COVID tweets.

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, where Joe Biden is the most referenced “PERSON” NE.

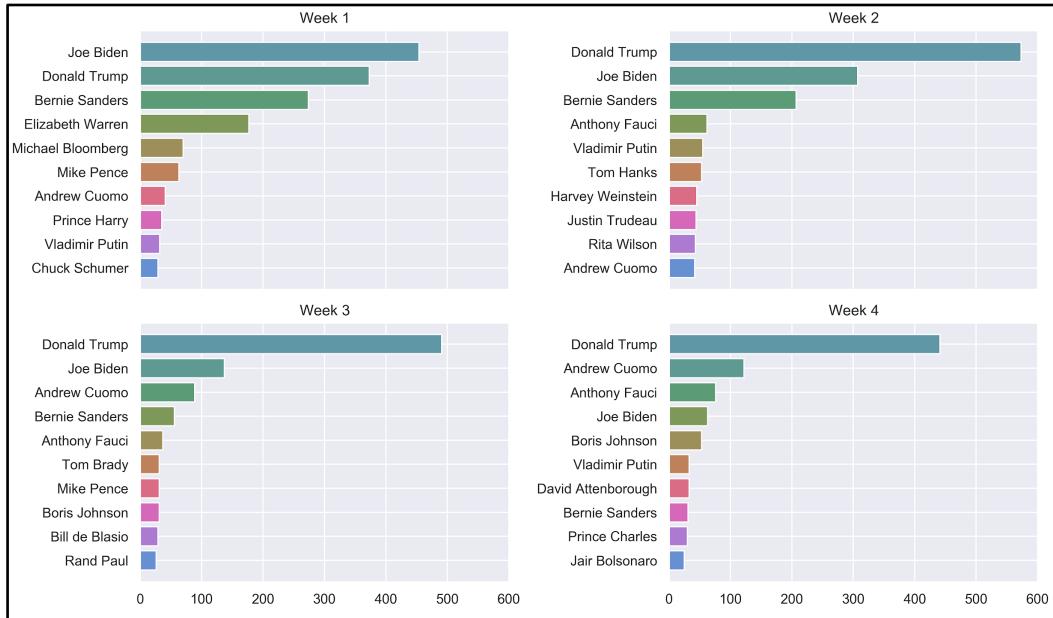


Figure 11a: The top ten most referenced “PEOPLE” NEs for all tweets.

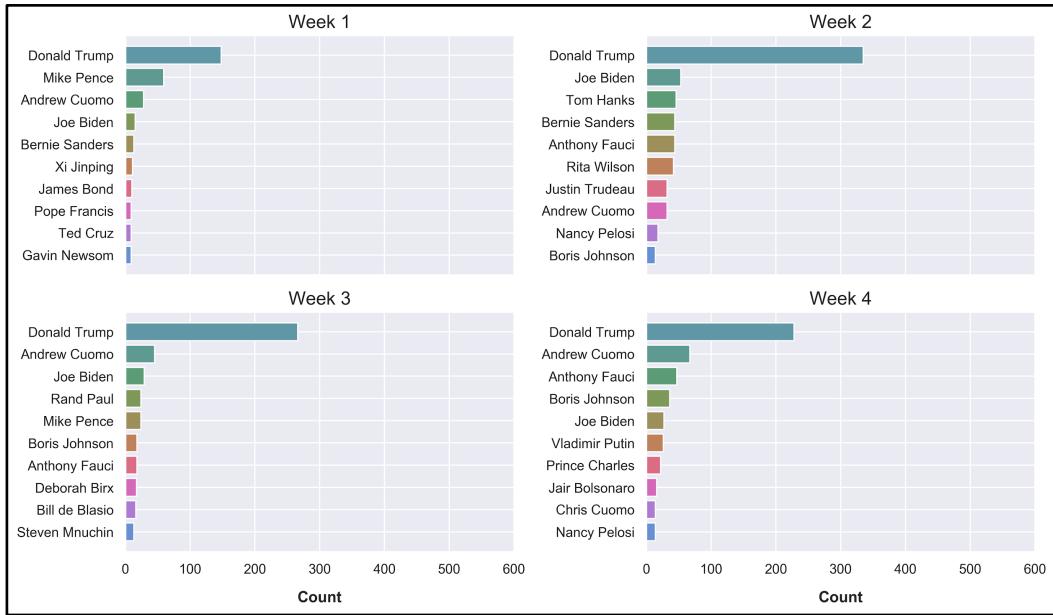


Figure 11b: The top ten most referenced “PEOPLE” NEs for COVID tweets.

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).

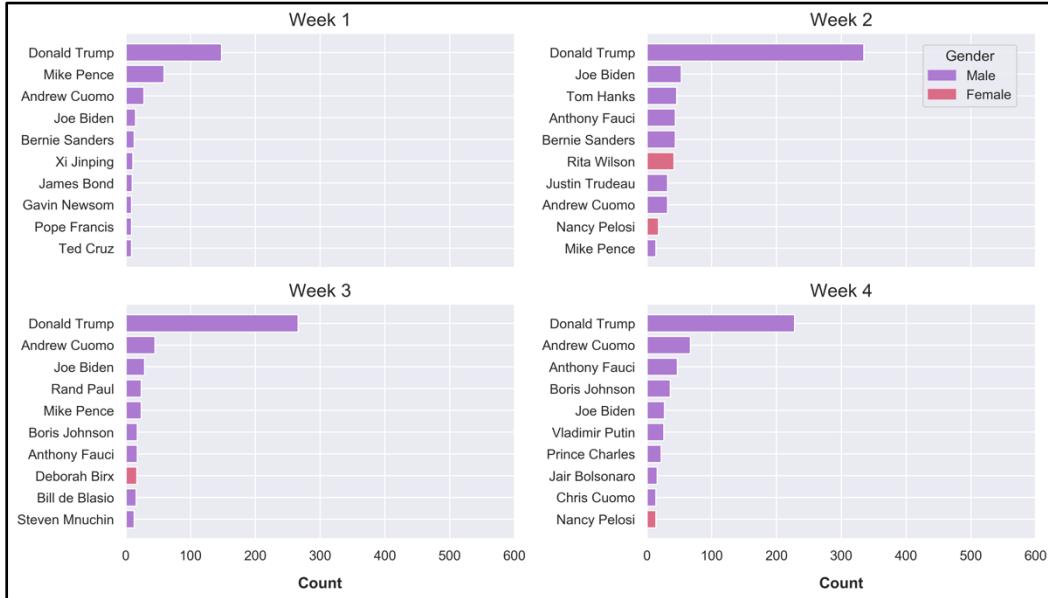


Figure 11c: Figure 11b, split by gender.

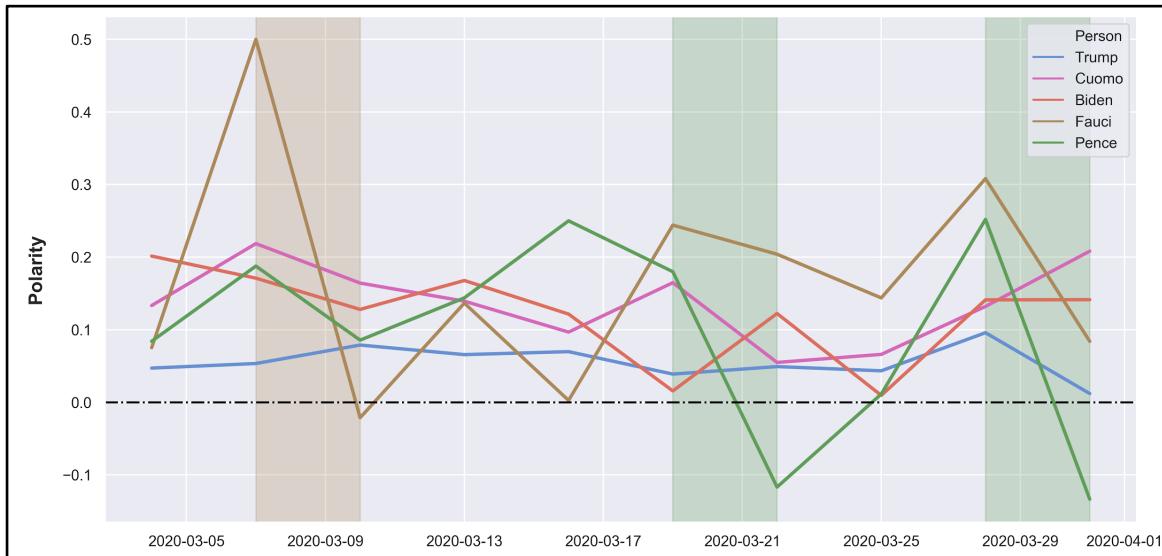


Figure 12: The polarity of the top five NEs. The dotted line shows where zero is. The shaded regions highlight major polarity shifts.

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 concentrate on the polarity of tweets containing these

top five NEs over the entire time period. Again, I use the same downsampling approach as used 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.

4. Discussion

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

4.1 Tendency Towards Positive Polarities in Tweets

One of the most surprising results from the SA is the tendency towards positive polarities (Figures 6 and 7). While these positive polarities do tend to be quite close to neutral, the slight positivity during a somewhat bleak period is unexpected. Even as more tweets are published concerning COVID (Figure 9), the polarities in Figures 7 and 8 tend to stay positive. Collecting additional tweets outside of the time period as a baseline, which is explored further in Section 5.1, would help place this trend into context. Perhaps the trends in Figure 7 actually reflect lower than average polarities, 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 (Figures 5a and 5b) suggests a tendency towards polar coverage when approaching COVID. 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.

4.2 Heterogeneity of Polarity Trends and NE Breakdowns

Figure 7 clearly shows that not all sources maintain the same 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 in sentiment 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 perhaps 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., below -0.5 or above +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. Regardless of the potential limitations associated with the approach, it is again apparent that overall polarity trends do not necessarily line up across sources (i.e., Reuters and ABC News stay consistent in polarity for COVID tweets while TIMES is much less stable). Furthermore, the differences in NE tag breakdowns in Figure

10 (i.e., BBC (World) and Reuters have a tendency towards “GPE” tags), while slight, also begins to reveal the contrasting choices in subject matter across news sources. Demszky et al. (2019) introduce a technique for clustering tweets by topic, which could be applied to this setting to further explore questions of framing and topic choice.

Across many results, trends across news sources tend to vary, sometimes significantly. An emerging theme here is that news sources are not homogenous. News media does not always act as a unified entity and, as such, each news source should likely be considered its own participatory actor within the context of COVID-related interaction on social media. Further research might consider comparing other actors, such as elected officials, public health officials, or crisis relief organizations, with news sources. Each actor plays a role in both the dissemination of information and the framing of issues; recognizing how each actor chooses topics and frames issues related to COVID will be extremely important in understanding how to manage public opinion in the future.

4.3 Inclusion and Portrayal of Public Figures

In the exploration of NE tag breakdowns, the shift from an emphasis on “PERSON” tags to “GPE” tags is unexpected (Figure 11). Perhaps this reflects the broad scope of COVID, which might explain the increase in “GPE” tags. 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 with instead 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 (Figure 11c). Again, a better baseline for comparison would help here, as it is possible that this just picks up on inherent sexism within news media coverage or unbalanced representation of public figures in general. Moreover, many of the most frequently 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.

The results shown in Figure 12 are certainly preliminary, but show some promise. Compared to Figures 7 and 8, the tweets about public figures generally exhibit a larger range of polarity scores. While most news sources stay roughly between 0 and +0.1 in Figures 7 and 8, only Trump stays within that same range in Figure 12. I do worry that the volatility in the polarities for Fauci and Pence may be due to relative data sparseness. 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 few different news sources). On the other hand, Cuomo and Biden do not 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 for the US COVID task force, so polarity shifts pertaining to either of them could be attributable to changing portrayals 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 further in Section 5.4. Studying 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 do not match up.

In periods of crisis, figures of authority must interface with the public to provide guidance and disperse critical information. More research concentrating on how individuals and organizational entities respond to and critique the actions of public figures would help to define the interactions between a broader range of participatory actors during COVID. Analyzing tweets from public figures as well as tweets from individuals could lead well into this line of research.

5. Limitations and Future Directions

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 of tweets that are not during COVID would be needed. Bleich and van der Veen (2018) use a comprehensive 20-year baseline for news articles. 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.

Unexpected data truncation presented another considerable challenge. 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 characters.³ 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 to publish tweets of less than 140 characters. But for many other sources (CNN, The New York Times, The Associated Press, ABC News), the

³ https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html

distribution is much flatter, with a significant number of tweets over 140 characters. 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 prevent 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.

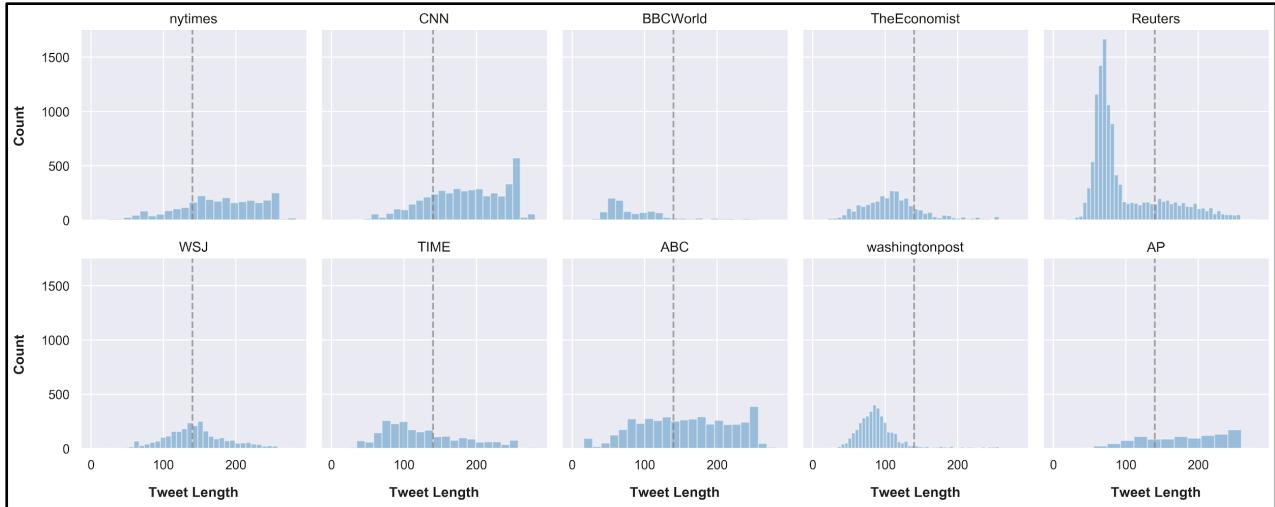


Figure 13: The distribution of character lengths for tweets from each news source. The dotted line shows where 140 characters is on the x-axis.

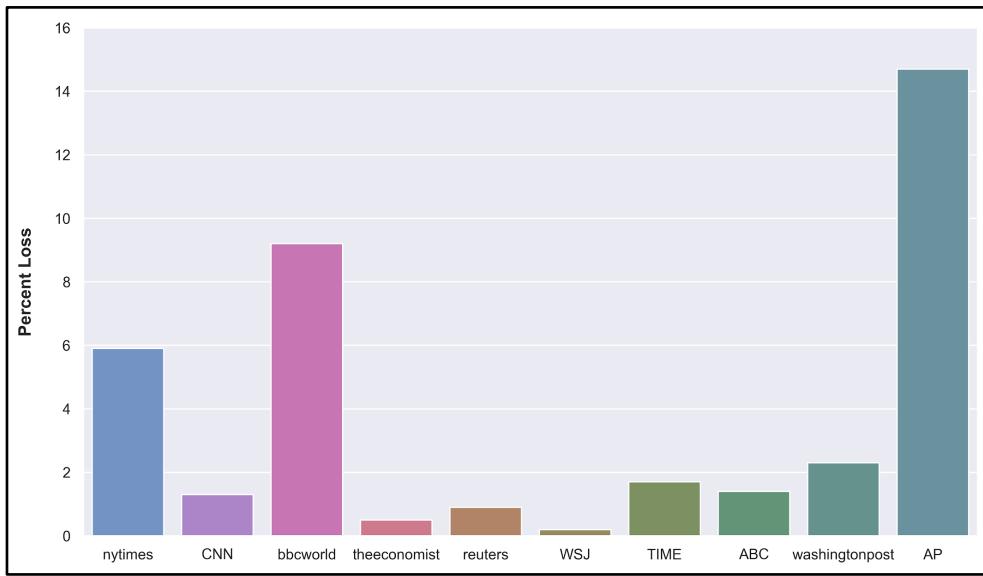


Figure 14: Expected character losses for each source, calculated using the mean character length for non-truncated tweets over 140 characters (using median character lengths produced nearly identical results).

Given this discovery, I then had to decide whether or not it would be appropriate to keep these truncated tweets in the data sets. 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. 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, Table 2), character losses are quite low. Overall, the data set only experienced a 2.8% character loss due to truncation and I decided to keep in the truncated tweets.

5.2 Analysis Methods

My design philosophy for this project focused on using efficient implementations of relatively simple analysis methods. By prioritizing 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 concerned with 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 (perhaps something similar to the methodology of Krouská et al. (2016)). 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. For instance, “@” references are often NEs (i.e., in the tweet excerpt “...@GuillemBalague explains why Lionel Messi is fed up at Barcelona...,” “@GuillemBalague” should be tagged as a “PERSON” NE), but spaCy has trouble properly identifying and tagging those that are. More work towards properly categorizing “@” references would improve NER performance here significantly.

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).⁴ Many of these workshops have included a shared task for NER with NUT, some even using text data from Twitter. Conditional random field (CRF) and long-short term memory approaches (LSTM) found success in the 2015 and 2016 tasks, respectively (Baldwin et al., 2015; Strauss et al., 2016). 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 could improve accuracy with this data, but given the semi-structured nature of the data, 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. Jiang et al. (2016) compare a number of well-established NER systems on structured data from Wikipedia using hand-annotated documents, ranking by F_1 score. A similar approach towards testing systems on semi-structured text corpora would be valuable. Knowing whether Twitter data from news sources is more similar to structured text data or noisier Twitter data would help guide further design choices in this area of research.

⁴ <https://www.aclweb.org/anthology/venues/wnut/>

As with NER, it has been shown that certain SA approaches perform better than others when applied to Twitter data. Thakkar and Patel (2015) show that lexicon-based systems for SA perform significantly worse than machine learning or hybrid approaches. Hamilton et al. (2016) introduce a method for the induction of specialized sentiment 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 to train the models (the approach of Hamilton et al. (2016) 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 semi-structured data would help minimize the amount of work needed to find the most suitable approaches.

5.3 Tweet Classification

The approach that I took for determining whether or not a tweet was about COVID was very simple, which may have affected certain results. To improve classification accuracy, a machine learning approach might be better suited to 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 efficacy of these methods, but I suspect that both would achieve much higher recall (a component of the statistical F_1 measure) than my more simple classification approach.

5.4 Extensions of this Project

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 direct comparison between a news source's presence on Twitter and on their native site. In the end, I chose not to pursue this particular line of research. Using python web scraping libraries like beautifulsoup, this could be an insightful extension to this project. Given the deluge of COVID-related coverage and high uncertainty, differing messages across platforms might exacerbate concerns during an already fraught situation. Additionally, working with full article data would be easier in some respects than working with Twitter data. While the quantity of text data would be much greater, most NLP tools are well suited to article data and would perform much better with little need for additional intervention.

While working on data cleaning and preprocessing, I noticed that some sources tended to repeat tweets often. For TIME and The Economist, only 62.6% and 68.1% of tweets are unique, respectively (Table 2). I decided not to do anything special with duplicated tweets, but taking a closer look specifically at the tweets that are duplicated (identical text) or high in token similarity could both be productive extensions of this project. For calculating token similarity, a simple bag of words approach would likely be sufficient as 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,

closely examining this subset might reveal certain hidden values for a given source (i.e., the stories that they most want to emphasize).

References

- Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.
- Erik Bleich and A. Maurits van der Veen. 2018. Media portrayals of Muslims: a comparative sentiment analysis of American newspapers, 1996–2015. *Politics, Groups, and Identities*, 0(0):1–20, November.
- Claes H. de Vreese, Hajo G. Boomgaarden, and Holli A. Semetko. 2011. (In)direct Framing Effects: The Effects of News Media Framing on Public Support for Turkish Membership in the European Union. *Communication Research*, 38(2):179–205, April.
- Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jeffrey Gottfried, Mason Walker, and Amy Mitchell. 2020. Americans’ Views of the News Media During the Coronavirus Outbreak. Technical report, Pew Research Center, May.
- William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.
- Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology*, 4(1):1–27, January.
- Ridong Jiang, Rafael E. Banchs, and Haizhou Li. 2016. Evaluating and Combining Name Entity Recognition Systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, Berlin, Germany. Association for Computational Linguistics.
- Akrivi Krouskas, Christos Troussas, and Maria Virvou. 2016. The effect of preprocessing techniques on Twitter sentiment analysis. In *2016 7th International Conference on Information, Intelligence, Systems Applications (IISA)*, pages 1–5. July.

Leysia Palen and Kenneth M. Anderson. 2016. Crisis informatics—New data for extraordinary times. *Science*, 353(6296):224–225, July.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. 2018. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human–Computer Interaction*, 34(4):280–294, April.

Elisa Shearer and Katerina Eva Matsa. 2018. News Use Across Social Media Platforms 2018. Technical report, Pew Research Center, September.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 Named Entity Recognition Shared Task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Harsh Thakkar and Dhiren Patel. 2015. Approaches for Sentiment Analysis on Twitter: A State-of-Art study. *arXiv:1512.01043 [cs]*, December. arXiv: 1512.01043.

Additional Materials

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.

TextBlob SA Lexicon

The lexicon used in TextBlob's SA approach can be viewed at the [TextBlob GitHub page](#).