Smith College Communicating COVID-19 (SCC19)

Emily Daubenspeck, Yejin Hwang, Jeny Kwon, Whitney Mutalemwa

Spring 2020 (May 8th)

Abstract

In this project, we investigate Smith College's response to COVID-19 by analyzing community letters sent by President Kathleen McCartney to the entire Smith College community during the 2017-18, 2018-19, and 2019-20 academic years. The community letters are analyzed using three separate sentiment lexicons (NRC, Afinn, and Bing) to compare the letters' content, structure, and tone before and after COVID-19. As the college carefully writes each letter to minimize negative sentiment in the public, it is more difficult to track changes in negative sentiments; however, our analysis shows the usage of positive words decreases over time. The results of our analysis are communicated through graphs, word clouds, as well as tables.

A summary of our findings can also be found on the following webpage: https://sds410-spring2020.github.io/SCC19/.

Introduction

Smith College's main form of communication to its community members is through emails. In fact, the Office of College Relations is tasked with ensuring that each email sent by the college's president Kathleen McCartney "fosters a better understanding of and support for the priorities of Smith in communities."[1] These emails use formal language and language style that closely resemble press releases. Unlike a single tweet with a meaning that is easy to decipher, emails sent by President McCartney require a reader to spend time assessing and evaluating their content. Though many emails are sent under the auspices of the President's Office, some are also released as "Letters to the Community" and permanently hosted on the Smith College webpage. These letters typically address some major event or situation and its relationship to the functioning of the College community. For example, letters have been published to acknowledge the death of former Smith College president Jill Ker Conway (6/2/18), discuss the potential impact of federal tax policy on the College's finances (11/8/17), and to respond to concerns in the wake of the 2016 presidential election (11/15/16). Given their centrality to the entire Smith community, these letters represent a unique window into the opinions and thoughts of the higher-level Smith College administration. The letters are written to convey both sentiment and information, to express heartfelt condolence and jubilance at various times. An analysis of these letters is fundamentally different from an analysis of other college-level communications (i.e. tweets, department emails, etc.), as these letters are intensely curated and vetted for both clarity and feeling. In creating a dataset of the content of these letters, it is possible to analyze both the Smith College administration's thoughts on a situation as well as its thoughts on how precisely to respond. In light of the current COVID-19 pandemic, we set out to conduct text analysis of a number of letters the school has sent throughout the past three academic years (2017-18, 2018-19, 2019-20) to assess its response to the COVID-19 pandemic. Such a momentous event, which has disrupted nearly every aspect of the College's function, represents a unique opportunity to evaluate Smith College's crisis response strategy and perform novel analyses on the information it releases to the community.

Data Description

The data for this analysis comes from publically available collegiate news postings published via the Smith College website. In order to communicate efficiently with students, families, faculty members, and the community, the College frequently opts to publish "letters," or briefings related to current events on campus or on topics that are otherwise relevant to the workings of the College. Most, if not all, of these letters, list Smith College President Kathleen McCartney among their authors. These letters represent valuable data because of their centrality to the Smith Community; the publication of a letter is typically accompanied by an email to students, faculty, staff, and alumnae who elect to receive notifications about the College. Therefore, the letters are not merely passive expressions of the College's thoughts and priorities, but a focal point around which the College community is engaged on particular issues. Because they are designed to communicate the Smith administration's position on current events and situations at the College, these letters are an excellent resource for assessing sentiment in the highest levels of the Smith College administrative hierarchy. Furthermore, due to their narrative structure and regular publication, they form a valuable and ready dataset for a rudimentary introduction to sentiment analysis.

The final dataset includes 77 letters published from August 2017 to April 2020. Though the College maintains published letters dating as far back as 2013, we determined that an abridged range would allow us to complete a more efficient and meaningful analysis. As our focus draws particularly on COVID-19 and the College's response, we feel it would be reasonable to draw only from letters published within the last three academic years. This allows us to distinguish between COVID-19 and non-COVID related sentiment while limiting the size of our dataset and the runtime of our webscraping code. In addition, academic years prior to 2018-19 do not have as many letters available. The letters range in length from 59 words to 788 words long, with 14 letters published in 2017, 25 published in 2018, 24 published in 2019, and 14 published in 2020. For each letter, three pieces of critical information are scraped: the letter's content (i.e. all of the words comprising

[1] "College Relations" Smith College, 05 May 2020, https://www.smith.edu/about-smith/college-relations.

the letter's body), the letter's title (i.e. the primary heading under which the letter is listed, which provides a brief description of the letter's purpose), and the letter's date of publication.

Methods

In order to access the letters and their content, we employ a series of webscraping techniques through R, primarily using the packages rvest and httr. rvest is a staple package for R-based webscraping, as it allows users to create html pages from URLs, select parts of an html document using CSS selectors, and parse html files into organized tables for analysis. In our code, httr is used primarily for its request functionality. This functionality is integrated with functions that allow users to easily check for the existence/content of a webpage. This is a crucial element in helping us determine whether our scraped URLs and webpages are associated with valid web addresses, and allows us to keep our scraping code running efficiently.

Collecting the letter data begins with cataloguing all available links on the landing page for the community letters. Because this page includes links to other pages and documents which are not published letters, we rely on the consistent URL format used to distinguish letters from other, non-letter, links. In cleaning all of the links scraped from the landing page, we employ our knowledge of this consistent format in order to seperate, reconstruct, and organize the URLs for letters we wish to include in our final dataset. After collecting these URLs, we aggregate them into a list. We then filter the list to include only those URLs linking to letters published within the last three academic years; this filtration is possible because text corresponding to the academic year is included in each URL. After filtering, we use the httr function GET() to check that all of our URLs of interest lead to webpages that actually exist on the web. We access individual letters' content with a function designed to iterate over each element in this list of existing URLs and scrape the contents of the associated webpage. Scraping content, as well as each letter's title, requires the use of rvest functions for identifying particular html nodes within an html document. Once these nodes are identified, and the html formatting is cleaned out of the text with the use of text replacement function sub(), one is able to effectively catalogue all of the applicable content into a list. This list, once unlisted, forms the basis for the `content` column in a dataset in which titles, content, and dates of publication for each letter are clearly delineated and associated with one another. Once the contents of each webpage are scraped and cleaned, we can summarise each letter in one row of a dataframe, with a single column containing all of the letter's text. Additionally, we are able to use regular expressions to search for and associate each letter's date of publication with its row in the dataframe. These dates of publication are indicated in the titles for each letter as they are shown on the Letters to the Community landing page.

Table 1: Table continues below

X1	ID	title
1	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
2	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
3	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
4	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
5	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
6	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
7	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
8	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
9	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020
10	1	Join Us In The First-Ever Global Illumination Night, May 4, 2020

Table 2: Table continues below

link
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020
https://www.smith.edu/president-kathleen-mccartney/letters/2019-20/may-4-2020

date	word	wordcount
2020-05-04	dear	1
2020-05-04	students	2
2020-05-04	staff	3
2020-05-04	faculty	4
2020-05-04	families	5
2020-05-04	beloved	6
2020-05-04	campus	7
2020-05-04	commencement	8
2020-05-04	traditions	9
2020-05-04	share	10

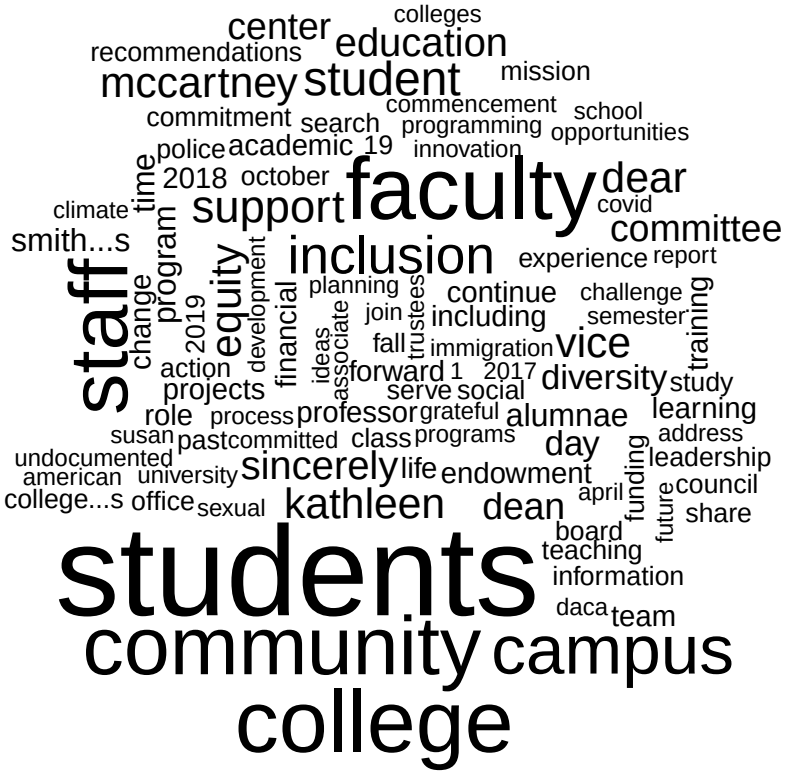
With help from the tidytext package, designed for sentiment analysis and text parsing in R, we separate out each word from each letter into its own row. This leaves us with an expansive dataset explicitly showcasing the entire textual composition of each letter scraped. In order to perform sentiment analysis, we proceed to download and join data sets associated with three separate sentiment lexicons (NRC, Afinn, and Bing) to our own dataset, producing additional dataframes wherein individual words arelinked to their lexicon-specific sentiment rating. These datasets can be manipulated, summarized, and plotted like any dataframe in R, and their content can be used to assess a variety of sentiment constructs within the letters.

The Lexicons

Bing employs a binary categorization of words into either positive or negative sentiment labels. NRC broadens the categorization to include sentiment labels for anger, anticipation, disgust, fear, joy, sadness, surprise and trust, in addition to positive and negative labels. Finally, AFINN uses a scale between -5 and 5 to define sentiment for individual words in the lexicon. Our objective is to analyze the words used in the letters individually; we are not seeking to analyze words as used in relation to their positions in sentences. Contextualizing sentences requires the use of sentiment analysis techniques much more advanced than those we employ for this study.

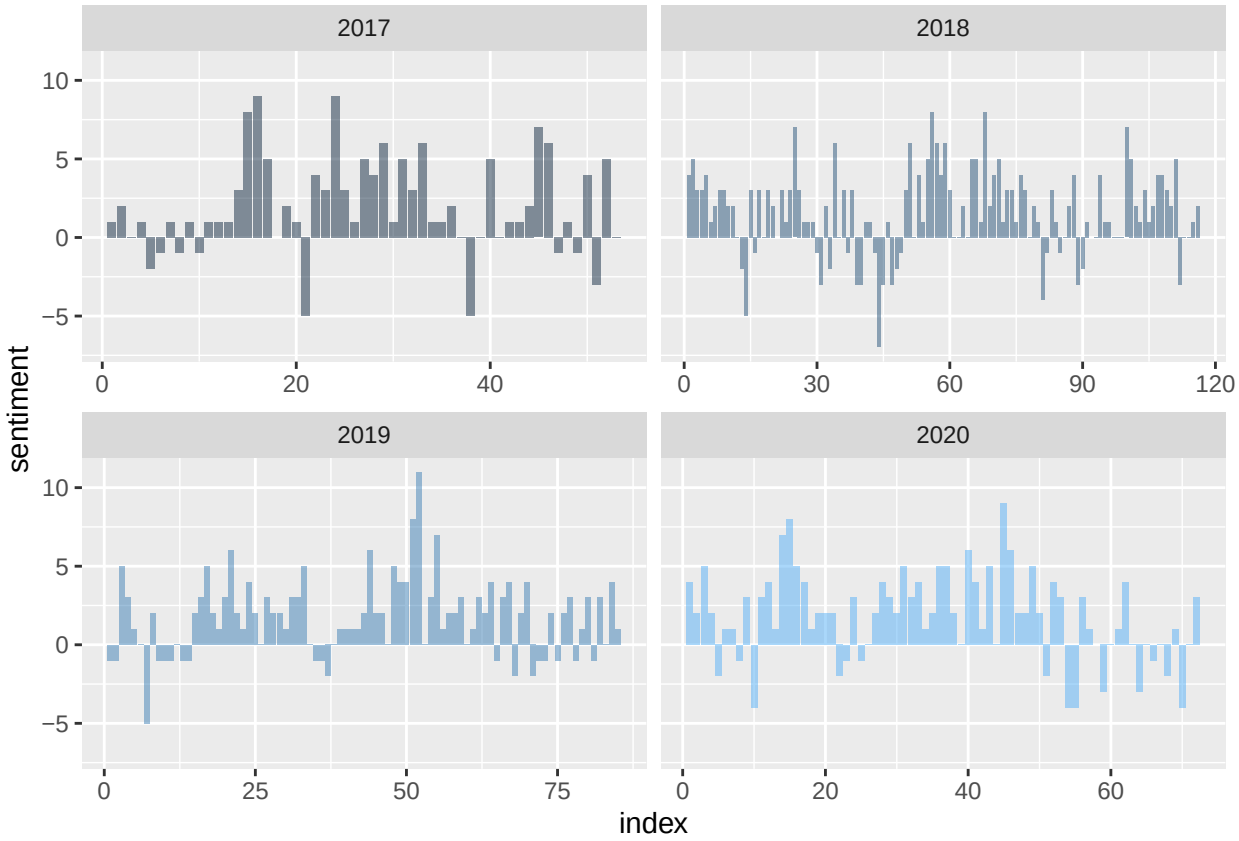
Results

The first exploratory analysis we conduct is a simple frequency counter for finding the most common words that appear in the letters published between 2017 and 2020. Afterwards, the one hundred most frequently used words are displayed in a word cloud. In this format, the pattern of words associated specifically with the Smith College campus was quite apparent. Among the frequently used words were “smith”, “students”, “community”, “faculty”, “campus”, and “staff”. Despite the insights this kind of frequency counting can provide, there is a need to introduce lexicons that qualitatively and quantitatively assess the usage of particular words throughout the letters’ content. These lexicons allow for a more nuanced analysis of the sentiment and structure used throughout the letters.



The first lexicon we apply in analyzing the letters is Bing. Like the frequency word counter, Bing is incorporated to identify the most commonly used words that connote a positive or negative meaning. From this, a second word cloud is generated, but this output is very different. This word cloud assigns positive meaning to “support”, “commitment”, “innovation”, and “recommendations” while negative meaning was assigned to “vice”, “fall”, “undocumented”, and “bias”. In applying sentiment labels to the letters’ content, this word cloud delves deeper into what a frequency analysis can reveal about our dataset. Now, we have access not only to the words used most frequently, but to those used most frequently within a particular sentiment category.

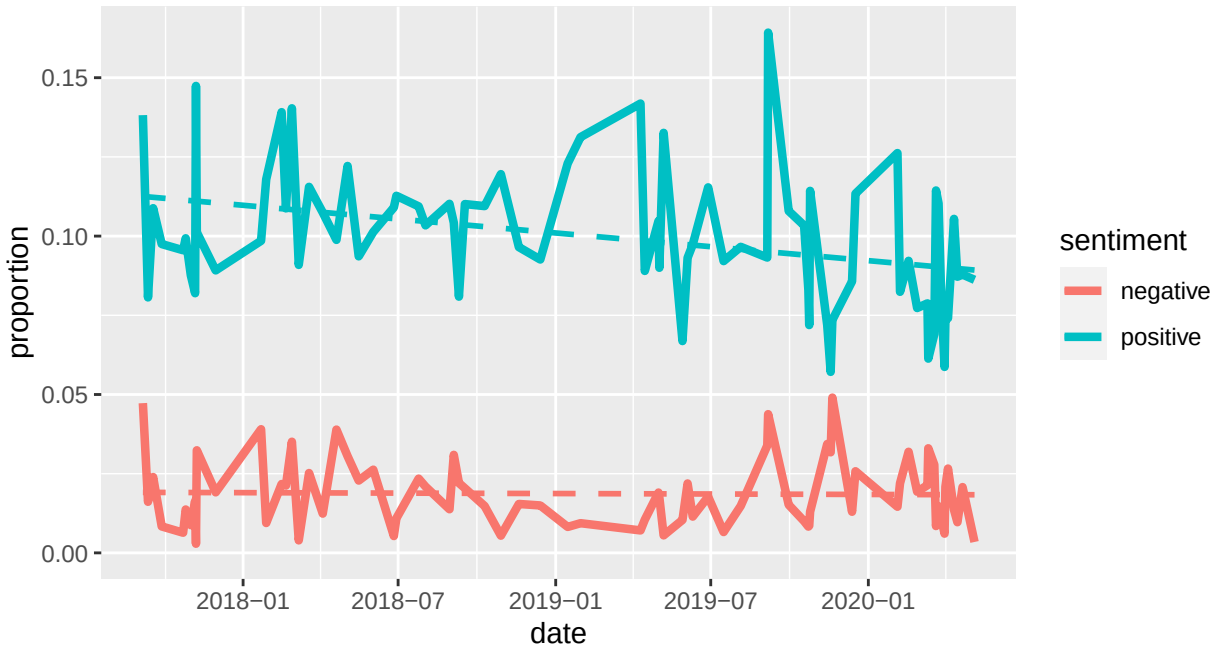
Instead of restricting our analysis with Bing to an exploration of high frequency unigrams, we decide to split the emails into chunks consisting of 50 words each to measure total sentiment scores for the chunks. Something to keep in mind is that while Bing does not have a numerical scale, theoretically, positive words are assigned +1 and negative words are assigned -1. Moreover, not all words will be given sentiment values such as names or words not in the lexicon. This more complex and quantitative analysis is shown in the bar charts faceted by calendar year. This sentiment graph shows a large amount of negative sentiment in both 2020 and 2018, as well as a clear trend in 2020 towards more negative sentiments in word chunks published later in the year.

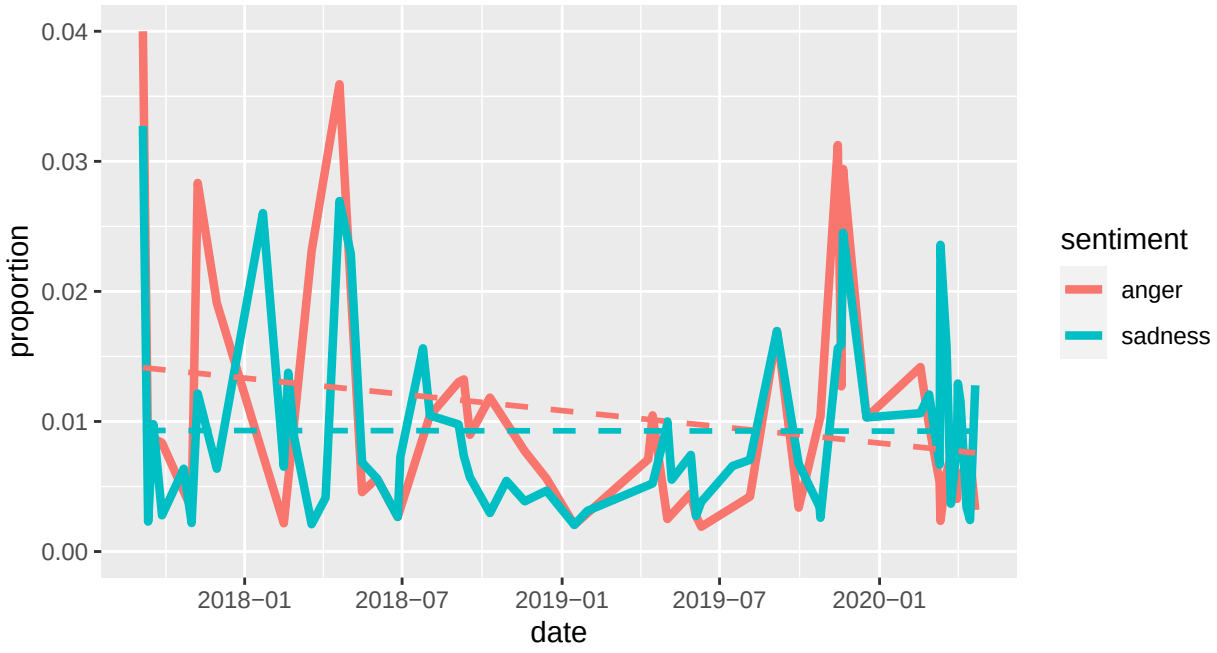


The next step in our analysis was using the NRC lexicon to expand our understanding of this dataset. With NRC, it is possible to track the evolution of particular sentiment categories across letters and time. Based on our analysis with NRC and its ten sentiment categories, the only sentiment that shifts substantially is anger. Anger shifts from 7th place in Pre-COVID-19 content to 8th in Post-COVID-19 content (sentiments with the most observations are at 1st, those with the least observations are 10th). Continuing the analysis using the NRC lexicon, we produce a word cloud to compare sentiments before and after the start of COVID-19. Observing which words are highlighted as frequently used in the word cloud, and going back to the original letters to explore Pre-COVID-19 topics, this shift in anger's prevalence is most likely associated with the previous hate issues that were on campus preceding COVID-19. Words identified as angry were being used to decry the acts of hate committed on campus, and their usage diminished as these acts were addressed less and less in the letters. Though COVID-19 represents a stressful and difficult time, as students reading these letters when they are released, we have come to presume that anger is likely not considered a productive sentiment to include in letters designed to either reassure, commiserate, or inform. Therefore, it makes sense that the use of angry words would decrease in Post-COVID-19 letters.

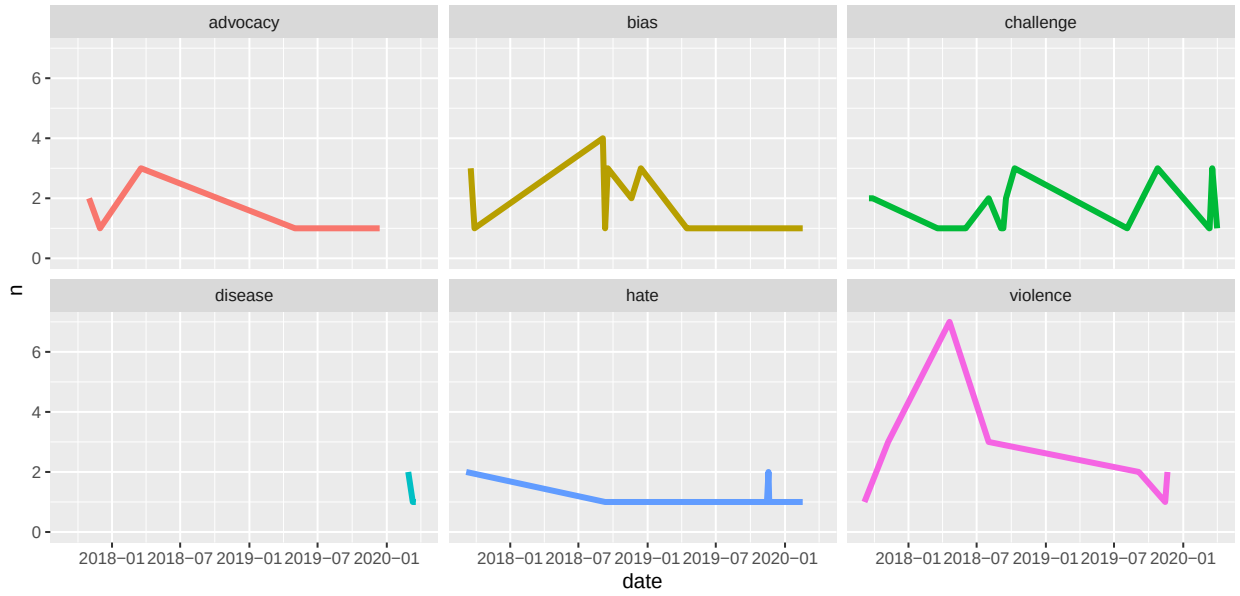
Ranking	Pre COVID-19	Post COVID-19
1	positive	positive
2	trust	trust
3	anticipation	anticipation
4	negative	negative
5	joy	joy
6	fear	fear
7	anger	sadness
8	sadness	surprise
9	surprise	anger
10	disgust	disgust
11	positive	positive
12	trust	trust
13	anticipation	anticipation
14	negative	negative
15	joy	joy

Finally, we approach a sentiment analysis with NRC over time. It is important to note a critical limitation of using sentiment analysis with the "Letters to the Community" - when addressing the public, negative words are intentionally minimized. Thus, the overall use of negative words is always less than the usage of positive words. Still, we notice a decrease in positive words over time, suggesting that while an increase in negative sentiment is not immediately evident, the President acknowledged the worsening COVID-19 situation with an adjustment of the letters' general sentiment. Using similar methods, when comparing anger and sadness, the trends for anger and sadness over time are almost identical in shape. This suggests that the issues addressed in the "Letters to the Community" were issues that caused both sad and angry emotional responses (such as hate crimes and COVID-19). To look at specific words, the five most commonly used angry words and the word "disease" are tracked over time. As expected, "disease" is only used after COVID-19. "Violence" seemed to have peaked in early 2018.

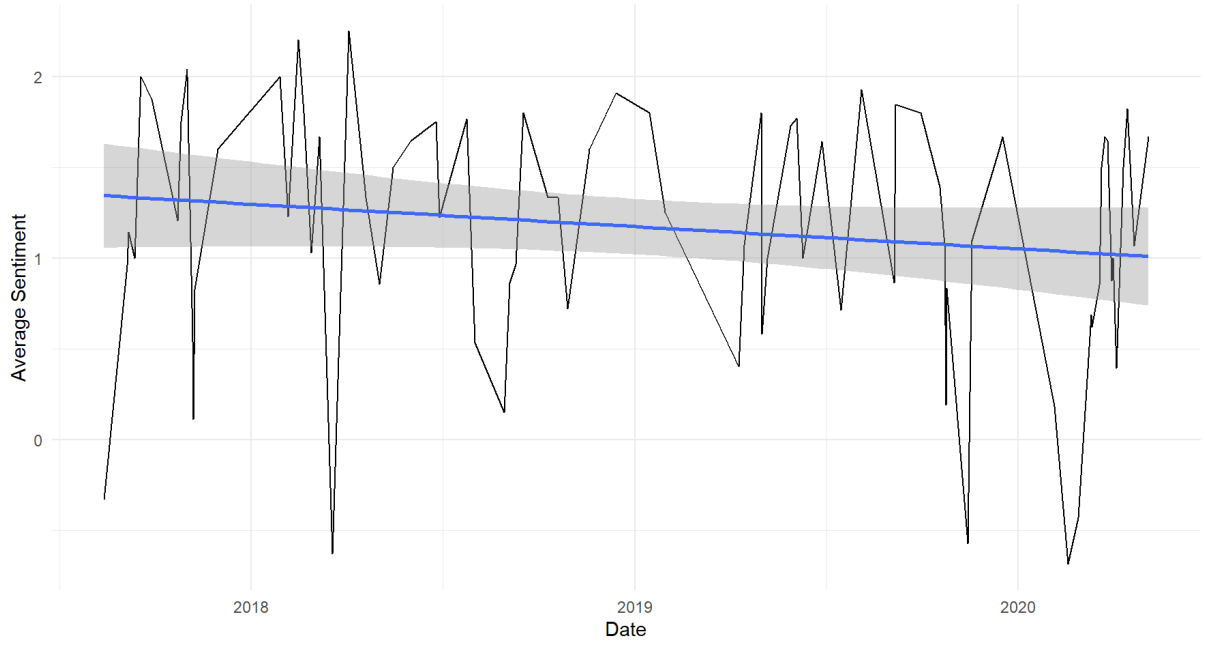

0.15
0.10
0.05
0.00
proportion
2018−01
2018−07
2019−01
2019−07
2020−01
date
sentiment
negative
positive


0.04
0.03
0.02
0.01
0.00
proportion
2018−01
2018−07
2019−01
2019−07
2020−01
date
sentiment
anger
sadness

Anger Words	Frequency
challenge	25
bias	20
violence	19
advocacy	8
hate	8
court	5
gun	5
threat	5
disease	4
fear	4
hatred	4
limited	4
painful	4
powerful	4
cross	3

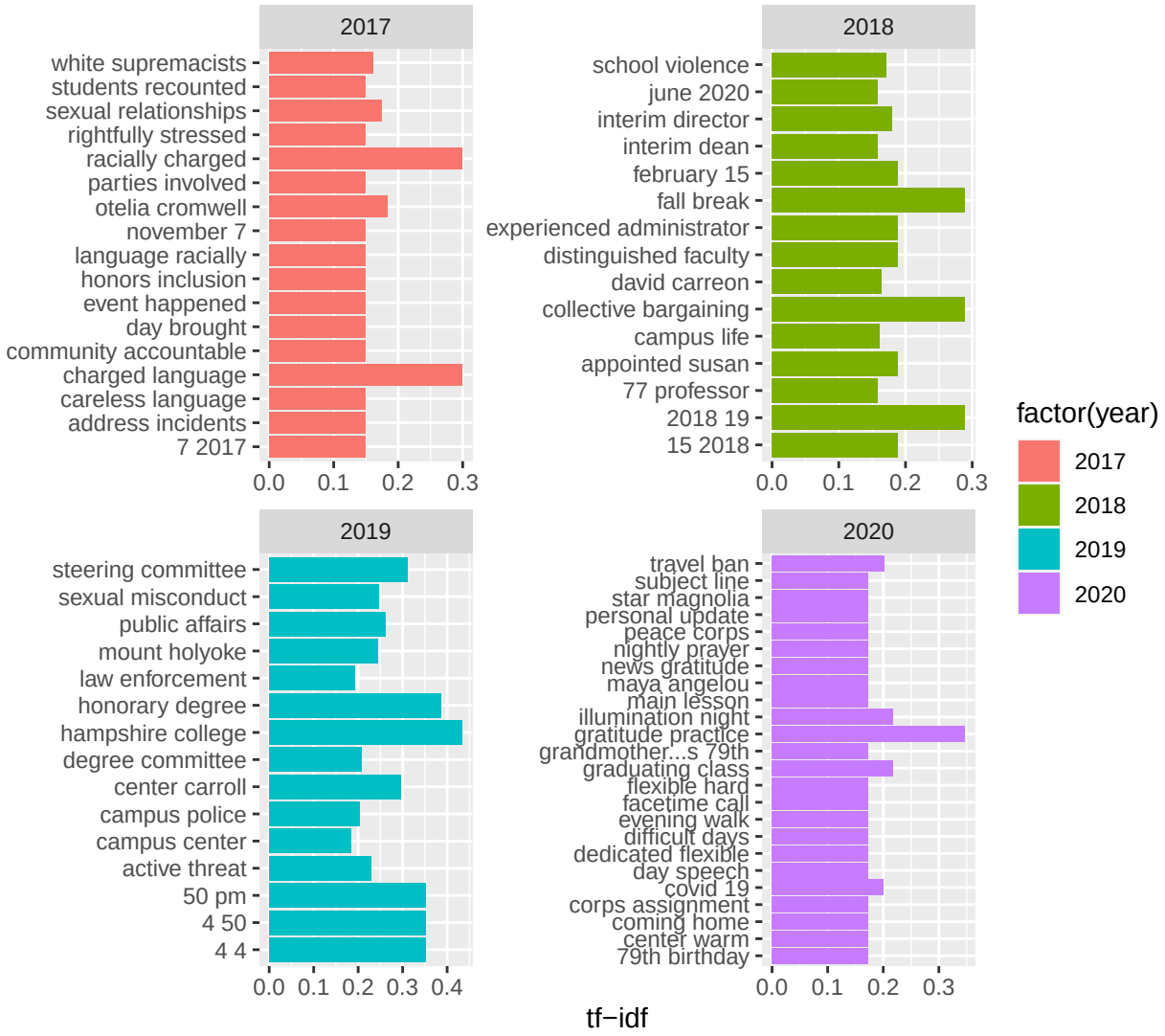


To conclude our formal lexicon-based analysis, we turn to AFINN. AFINN's unique categorization system allows us to perform more purely quantitative analysis on the Smith College "Letters to the Community". By assigning an actual number value to each word in the lexicon, we can calculate the average sentiment in each letter from the year 2017 to present day. In the plot above, we can observe that average sentiment is generally decreasing through time, partially due to an increased frequency in extreme sentiment dips at the end of 2019 and in early 2020. This dip in 2020 may be primarily attributed to news surrounding the COVID-19 pandemic, but the origin of the dips in late 2019 is less clear. This graph is also useful in illustrating the frequency with which "Letters to the Community" were published; we can see that for a period in early 2019, letters were published much less frequently than in prior months, possibly due to the winter vacation from classes. There is a similar pattern in early 2018 and early 2020.



Additional Text Analysis with Bing

Earlier, while we were conducting sentiment analysis using Bing, we encountered inconvenient situations. For example, the words "fall" and "vice" are categorized as negative by this lexicon. In reality, with additional context provided by adding a neighboring word, "fall" ceases to be negative when presented as "fall semester," and "vice" ceases to look malicious when configured as "vice president". Further sentiment analysis of this text can be done on these bigrams, or pairings of words. Once we are able to tokenize the letters' content into pairs of words, we then seek to find the most commonly written bigrams from President McCartney by year. The results are given in terms of how often a bigram appears in a particular letter relative to other community letters by year. This can be calculated as the number of times a bigram appears in a letter divided by the total number of bigrams in the letter. For the purposes of this paper, we are concerned with the output for the year 2020. The three most frequently written bigrams are "gratitude practice", "travel ban", and "covid-19". Clearly, communication of the pandemic as well as "travel ban" dominate the emails sent to the Smith community during this time.



Discussion

The analysis described here emerged out of a collective desire to better understand a few key aspects of data science: web scraping, sentiment analysis, and the applications of data analytics to a novel dataset. Though our analyses are mostly exploratory and involve little, if any, modeling, our work here is valuable in the sense that it has elevated our understanding of these three concepts and allowed us to engage with a current and real-world topic of interest. Additionally, this work has allowed us to explore administrative communications from a quantitative perspective. Though the letters comprising our dataset rely greatly on the context surrounding their topics, assessing the word usage, sentiment, and structure of the letters has allowed us to gain a richer understanding of the factors at play in crafting public institutional communications. Data science allows us to make the common strange; it enables us to peer through a new lens at the letters we receive each day or week, and to apply new skills in unpacking their greater meaning. In its many forms, our analysis has driven us to examine these letters anew, and with an eye to the potential for future analyses of this sort.

Initially, our text analysis focuses on the frequency of words used throughout the emails. We also use the Bing lexicon to expand on the word frequency by categorizing the unigrams as either positive or negative. In using this lexicon, our concern soon became that some of the words assigned as negative may not in fact be

negative given some context. For instance, President McCartney frequently writes about the "fall semester" and administrative positions with the title "vice" in her emails yet "fall" and "vice" were deemed negative by Bing. Still, we were interested to see how the sentiment score of word groups (per 50 words), rather than word frequency, changes as time passes within the four calendar years. For the most part, letters communicated in 2020 are positive overall. However, over time, the letters begin to have more content that the Bing lexicon determines as negative. In our raw data, COVID-19 related letters began on February 7th, 2020. Since this date, the sentiments of the letters have been more negative. There have also been more words in the 2020 letters in comparison to the other three years. The only other year that has a similar negative sentiment as 2020 is 2018, and a significant number of 2018 emails dealt with campus climate and issues of inclusion and equity. Clearly, according to the Bing lexicon, communication between the administration and the Smith community during 2020 is more likely to be negative than in the other three years.

To conclude our analysis, we turn to the AFINN lexicon. AFINN's unique categorization system distinguishes it from NRC and Bing, and makes it a useful part of the tidytext package. By incorporating this lexicon into our analysis, we transition entirely away from a qualitative assessment and enter an environment wherein quantitative sentiment labeling is the default. Of course, there is always the potential for the application of qualitative labels to the numerical AFINN scale, but for the sake of time and organization, we determined that this would be better explored in a future analysis. With AFINN, we are able to seamlessly assess the average sentiment per letter over time and between calendar years. In 2020, for example, we see the kind of dip in average sentiment which may be attributed to the start of the COVID-19 pandemic. This dip occurs near the beginning of the 2020 calendar year, which aligns with our knowledge of when the COVID-19 pandemic began entering Smith College's communications. With AFINN, we are also able to observe the temporality of letter publication, with letters coming more sporadically during the beginning of each calendar year. This could possibly correspond to the period of winter recess, during which communications may be less critical and frequent. Any future analysis using AFINN should consider the potential for applying qualitative labels to the numerical scale, as well as advancing the use of the quantitative sentiment labels. Models could be constructed to predict the month during which negative letters are most likely to be published. Additionally, a more detailed analysis of the temporality of letter publication could help determine the real-world context behind some of the major dips or peaks in sentiment value over time.

A note about mean calculations with the AFINN lexicon: though expansive, AFINN does not include every word in the English language, and through our analysis we have intentionally excluded "filler" words like "the" and "to". In order to calculate the mean sentiment per letter, we are using only the words present in the AFINN lexicon. This means that we are certainly underestimating the number of words per letter, potentially skewing the mean sentiment calculations for longer letters with fewer words present in the AFINN lexicon. The mean calculations should be taken as solely reflective of the letters within the context of this particular lexicon.

Ethical Statement

When scraping data available online, there are ethical considerations to keep in mind. Although the community letters from President Kathleen McCartney and Smith College are publicly available on the College's webpage, the authors might not be aware or comfortable with a mass sentiment analysis of the letters. President McCartney likely did not anticipate that her words would be more closely analyzed; therefore, we need to be considerate of this and certain that we are mindful of her rights and interests while conducting our analysis and sharing the results. This would most likely take the form of an assurance that our results should in no way be assumed to reflect the actual sentiments or feelings of any one member of the Smith community. In addition, when scraping data from the web, the terms of use on any website should carefully be read and understood before any scraping takes place. Such terms of use may be put in place to discourage the aggregation of sensitive data, so a careful assessment of the kinds of data available and its potential for harm is critical before beginning any analysis of this sort. In this case, the "Letters to the Community" are exceedingly public, not merely published on the Smith College website, but also emailed out to the greater Smith community. As letters designed to communicate the position of the greater Smith administration, they are not especially inflammatory, and they do not include any personally identifiable information. In

aggregating their content, we are perhaps making an implicit statement regarding the letters' construction and their curation (i.e. that the letters are expertly crafted, not candid expressions of President McCartney's thoughts and feelings), but this is a known quality of the letters.

We also need to be mindful of not defining Smith College's response to COVID-19 simply based on an analysis of the community letters sent by the President. A more holistic review of how Smith College approached the situation is needed to determine if the college was able to respond adequately. This analysis takes into account only a single form of communication from a single administrative vantage point. Individual departments, professors, faculty, and staff have released their own communications expressing their own views of the current situation. Furthermore, communication from collegiate entities to students and the community does not encapsulate the full breadth of Smith College's actual, or potential, response to the COVID-19 pandemic. The College's actual initiatives, timing, and follow-through on plans to address the situation are not easily assessed through communication and require additional scrutiny. This analysis is, fundamentally, a small window into the changes in a particular form of communication over time. Further analyses of the College's response during this time will likely take myriad forms and delve deeper into the reality of Smith's actions.

Lastly, this analysis follows the communication around COVID-19, which can be a sensitive topic for everyone impacted by this pandemic. In selecting this topic for our analysis, we considered the potential for harm any presentation of our findings might have on our peers or on our own mental health. This situation is uniquely stressful and all-consuming; many students found themselves abruptly uprooted from their lives, homes, communities, and families over the course of the last few months as a direct result of the pandemic. Our own work, in further addressing the topic, has the potential to upset or strain the emotions of anyone impacted by this situation. Therefore, we have sought to keep our communication on the subject of this project as sensitive and tactful as possible. Our GitHub repository does not include any direct reference to COVID-19 in either its name or the associated README. Our communications in the class's public Slack channels have been intentionally vague, and we have kept discussions of our specific topic relegated to our private channel. These efforts have all been implemented in the hope that our work, while stimulating for us, will not negatively impact the mental health of any of our peers as they strive to complete their own projects.