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." 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 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.

```
## # A tibble: 10 x 6
##       ID title                link                 date       word   word_number
##    <int> <chr>                <chr>                <date>     <chr>        <int>
##  1     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 dear             1
##  2     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 stude~           2
##  3     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 staff            3
##  4     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 facul~           4
##  5     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 famil~           5
##  6     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 belov~           6
##  7     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 campus           7
##  8     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 comme~           8
##  9     1 "Join Us In The Fir~ https://www.smith.e~ 2020-05-04 tradi~           9
## 10     1 "Join Us In The Fir~ https://www.smith.e~ 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

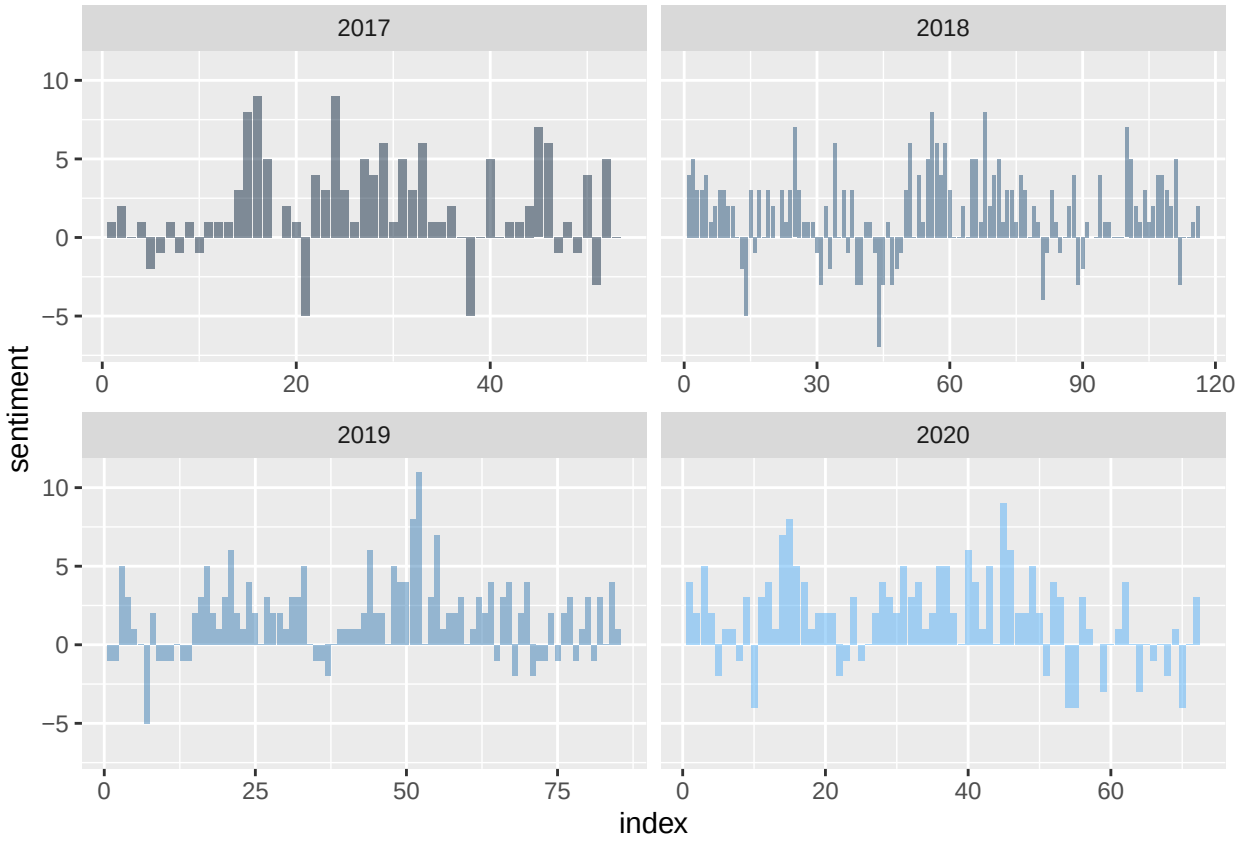
Bing employs a binary categorization of words into either positive or negative. NRC broadens the categorization to include anger, anticipation, disgust, fear, joy, sadness, surprise and trust, in addition to positive and negative. Finally, AFINN uses a scale, between -5 and 5, to assign to words. Our objective is to analyze the words used in the emails rather than seeking to analyze words as used in relation to their positions in sentences because contextualizing sentences takes much more advanced techniques than those we use for this study.

Results

The first exploratory analysis we conduct is a simple frequency counter of finding the most common words that appear in the emails sent between 2017 and 2020. Afterwards, the hundred most frequently communicated words are displayed in a word cloud. From this, the pattern of words related to the Smith College was expected. Among the frequently used words were "smith", "students", "community", "faculty", "campus", and "staff". However, the word cloud created can not be used further, and there is a need to introduce lexicons that qualitatively and quantitatively assess their usage.

Hence, the first lexicon we use is Bing. In addition to the frequency word counter, Bing is incorporated to identify the most used words that connote a positive or negative meaning. From this, a second word cloud is generated with the output being very different. This word cloud assigned positive meaning to "support", "commitment", "innovation", and "recommendations" while negative meaning was assigned to "vice", "fall", "undocumented", and "bias".

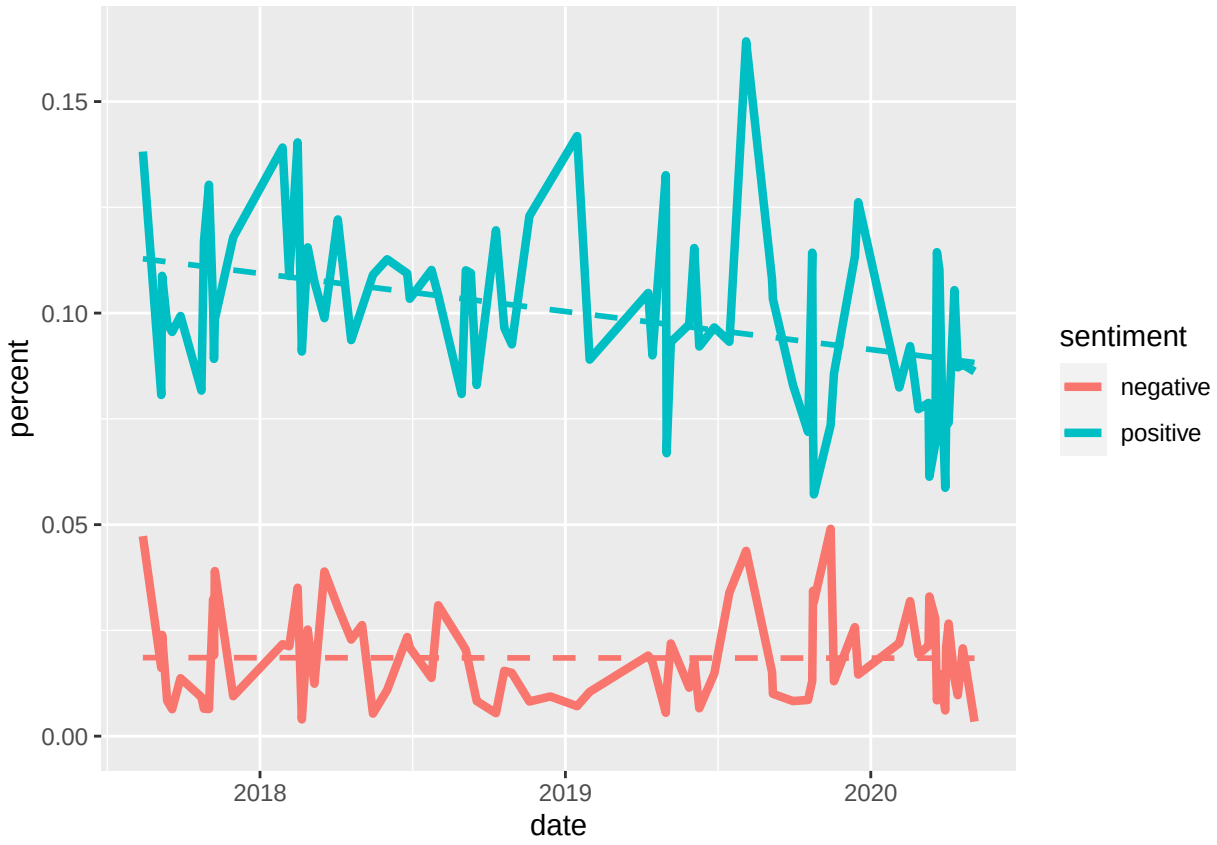
Instead of putting emphasis on Bing's analysis of high frequency unigrams, we decide to split the emails into chunks consisting of 50 words each to measure sentiment scores for the chunks. Something we keep in mind is that while Bing does not have a numerical scale, theoretically, positive words are 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. Over time, the sentiment graph shows negative sentiment for 2020 and 2018.



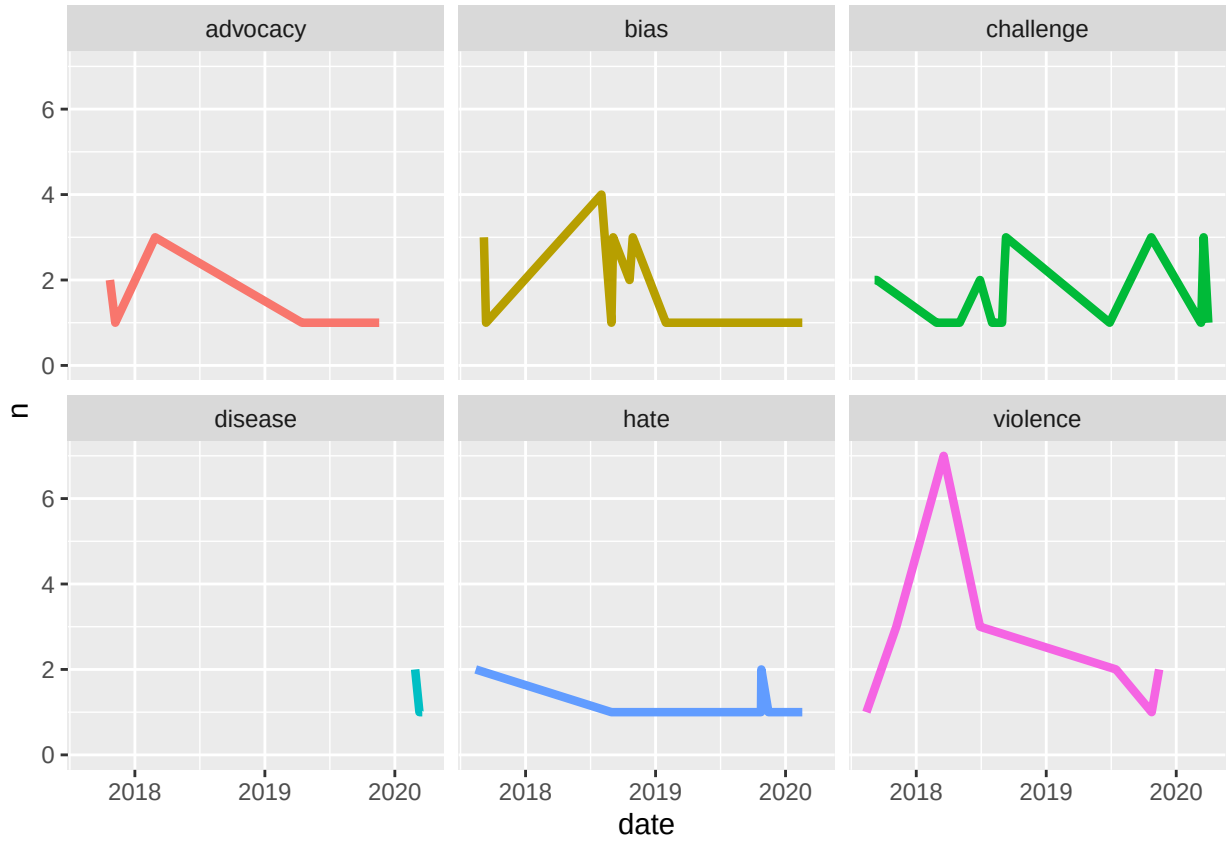
Based on the analysis using the NRC lexicon and its ten sentiment categories, the only substantial sentiment that shifts is anger (from 7th in Pre-COVID-19 to 8th in Post-COVID-19) when ranked from most observations at 1st and least observations as 10th. Continuing the analysis using the NRC lexicon, a word cloud comparing before and after the start of COVID-19, and going back to the original letters, the shift is most likely to do with the previous hate issues that were on campus preceding COVID-19.

[table does not appear :(]

Finally, sentiment analysis based on NRC is analyzed over time. It is important to note a 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 shown, the president verbalised the worsening situation. 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 that the "Letters to the Community" addressed were issues that caused a sad emotional response (such as hate crime and COVID-19). To look at specific words, the five most used anger 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
percent
2018
2019
2020
date
sentiment
negative
positive


0.04
0.03
0.02
0.01
0.00
percent
2018
2019
2020
date
sentiment
anger
sadness


advocacy
bias
challenge
disease
hate
violence
6
4
2
0
n
2018
2019
2020
date

[Need AFINN Results]

Discussion

Initially, our text analysis focuses on the frequency of words used throughout the emails. We also use the Bing lexicon to add to the word frequency by categorizing the unigrams as either positive or negative. 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 the grouped content rather than word frequency, consisting of 50 words each, changes as time passes within the four years. For the most part, emails communicated in 2020 are positive overall, however, over time the emails begin to have content that the bing lexicon determines as negative. In our raw data, COVID-19 related emails began February 7th, 2020, and since then the sentiments of the emails have been negative. There have also been more words contained in the 2020 emails 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 is more likely to be negative than in the other three years.

Ethical Statement

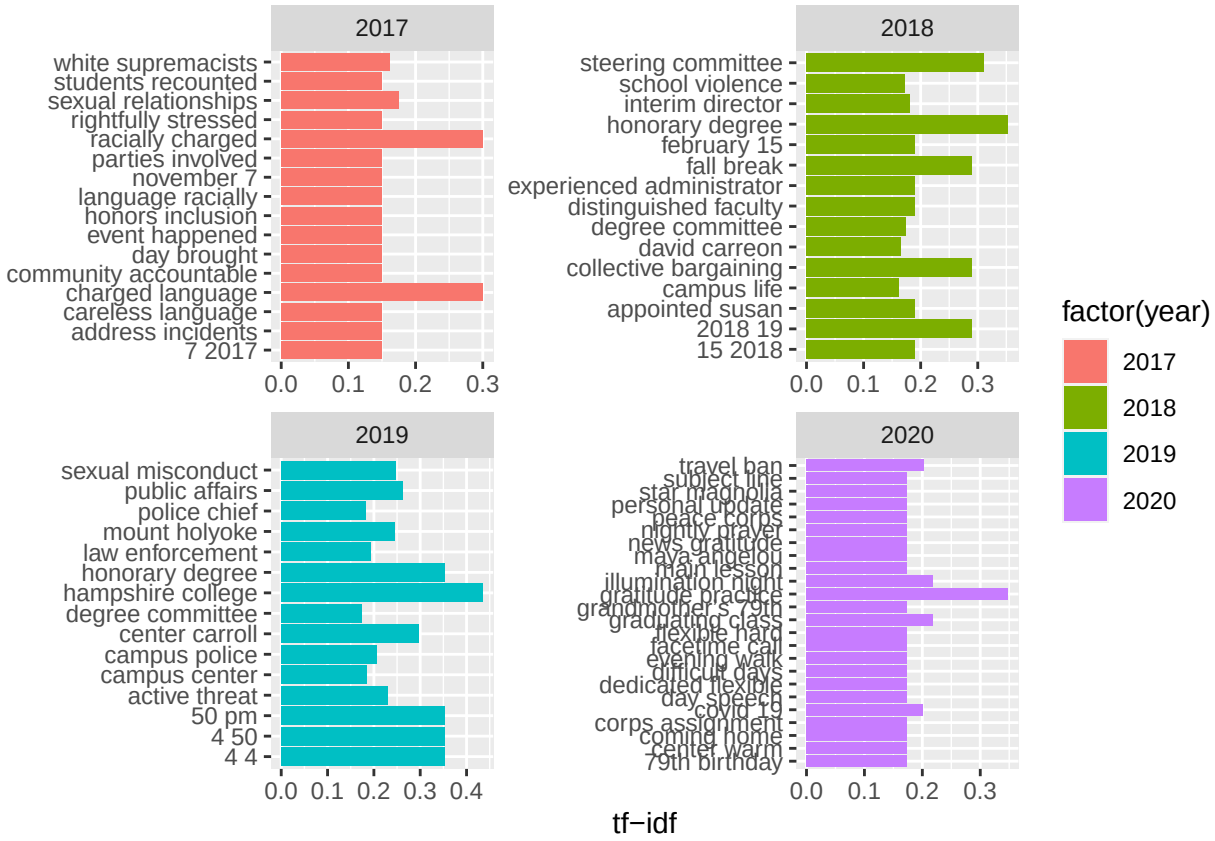
When scraping data available online, there are ethical considerations to keep in mind. Although the community letters from President Kathleen McCartney and Smith College are publicly available on the college's webpage, they might not be aware or comfortable with a mass sentiment analysis of the letters. President McCartney likely did not anticipate her words to be more closely analyzed; therefore, we need to be considerate of this and assure we are mindful of her rights and interests while conducting our analysis and sharing the results. In addition, when scraping data from the web, the terms of use should carefully be read and understood before any scraping takes place.

We also need to be mindful of not defining Smith College's response to COVID-19 just based on the 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.

Lastly, this analysis follows the communication around COVID-19, which can be a sensitive topic for everyone. Therefore, our communication on this topic should be as sensitive and as tactful as possible.

Additional Text Analysis

Earlier, while we were conducting sentiment analysis using Bing, we encountered inconvenient situations such as the words "fall" and "vice" being categorized as negative by the lexicon. In reality, with more context in form of adding a neighboring word to the right, "fall" ceases to be negative as "fall semester" and "vice" ceases to look less malicious as "vice president". Further additional text analysis is done on what could be inferred with the introduction of bigrams, that is pairing of words. Once we are able to tokenize the emails' content into pairs of words, we then seek to find the most commonly written bigrams from President McCartney by year. The criteria for the results is given how often a bigram appears in relation to other community letters by year. This can be calculated as the number of times a bigram appears in an email divided by the total number of bigrams in the email. For 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.


2017
white supremacists
students recounted
sexual relationships
rightfully stressed
racially charged
parties involved
november 7
language racially
honors inclusion
event happened
day brought
community accountable
charged language
careless language
address incidents
7 2017
0.0 0.1 0.2 0.3
2018
steering committee
school violence
interim director
honorary degree
february 15
fall break
experienced administrator
distinguished faculty
degree committee
david carreon
collective bargaining
campus life
appointed susan
2018 19
15 2018
0.0 0.1 0.2 0.3
2019
sexual misconduct
public affairs
police chief
mount holyoke
law enforcement
honorary degree
hampshire college
degree committee
center carroll
campus police
campus center
active threat
50 pm
4 50
4 4
0.0 0.1 0.2 0.3 0.4
2020
travel ban
subject line
star magnolia
personal update
peace corps
nightly prayer
news gratitude
maya angelou
main lesson
illumination night
gratitude practice
grandmother's 79th
graduating class
flexible hard
facetime call
evening walk
difficult days
dedicated flexible
day speech
covid 19
corps assignment
coming home
center warm
79th birthday
0.0 0.1 0.2 0.3
tf−idf
factor(year)
2017
2018
2019
2020