– Independent Work Report Fall 2018 –

"don't ignore this:" Automating the Collection and Analysis of Campaign Emails

Maia Hamin
Adviser: Arvind Narayanan

Abstract

Even as political marketing has become more sophisticated, email has remained one of the most popular and effective ways for campaigns to conduct fundraising and outreach. Emails are a rich source of data about the marketing approaches that each candidate has chosen, from their marketing vendor of choice to the rhetorical approaches they use to persuade their consitutents to donate. Studies in this area, however, are limited by the difficulty of manually accruing emails from political campaigns, especially since the lifecycle of an election – and, therefore, the contemporaneity of an email corpus — can often be shorter than the amount of time required to do the data collection. To address the challenge of email collection at scale, this work adapts a webcrawler to scrape candidate websites and demonstrates its efficacy by filling out email sign-up forms on over 1,700 campaign websites. It also introduces a data pipeline which cleans and stores campaign email data for easy integration with several scripts developed for analysis of emails. Here, the developed analytic tools are used to compare the presence of third-party tracking pixels between different types of campaign, as well as to perform exploratory analysis of the language use and rhetorical strategies employed by different senders.

1. Introduction

When political scientists discuss the importance of email marketing in campaign fundraising, they almost always begin by talking about the 2012 Obama campaign. That campaign raised $690 million through grassroots donations solicited through emails with subject lines like "Hey" — one of the campaign's most successful — that were the result of rigorous A/B testing. Since 2012, driven by success stories like the Obama campaign and the increasing availability of email marketing and

testing services, email has become a mainstay of all major (and most minor) campaigns. And it has paid dividends: the majority of individual-donor donations to the 2016 Clinton campaign campaign were converted through email.

A/B testing for political campaigns typically eschews the focus groups in favor of testing in real time on real constituents by sending different emails to different study groups and monitoring the rates at which each group opens the email, or clicks through to the site and makes a donation. To figure out which of their emails are opened, political campaigns resort to the same tracking techniques employed by advertisers who want to track users across websites. Emails, rendered in HTML, embed "tracking pixels", imperceptible images hosted on the sender's site. When a user opens an email, the image must be loaded from the hosting site, and that load request allows the hosting site to track when a page or email was first opened. Often these images have unique urls containing identifying strings that allow the host to detect precisely which user opened the page. Campaign emails employ this method to determine which of their subject lines result in a high rate of email-opens, as well as tracking which of these email-opens lead to a successful conversion to donation through similarly identifying information in the url of the helpful "Donate Now!" button.

So, what kinds of appeals result in the most clickable emails? It may well be that some campaigns have an empirical answer to that question: given the feasibility of A/B testing, sophisticated political operations are likely to have honed in on the kinds of subject lines and bodies that generate user responses. The question for a researcher, then, becomes whether it is possible to empirically describe the kinds of rhetorical strategies at which campaigns have arrived.

Political rhetoric analysis has traditionally been completed by hand, with researchers coding statements for features of interest and then generalizing about the nature of the text on this basis. Increasingly, however, computer-assisted language processing is useful for cases where hand-coding text might be prohibitive, or where computational pattern-matching and prediction can answer questions that human intuition cannot. The central challenge of processing natural language, in this task and in almost all others, is to represent the nuances of word meaning and sentence structure in vectors or other inputs that a computer is capable of processing. In this task, the

challenge is compounded by the fact that there is no "ground truth;" identifying and categorizing rhetorical approaches is itself a subjective question, rather than a straightforward case of finding the objectively-correct labels to feed into a classifier.

In this case, then, the advantage of the computational approach is that it allows standard analytical frameworks and approaches, already vetted by political scientists, to be translated into the digital domain so that they can be applied at scale. One such framework, and the one chosen as a case study for this work, is a language-based attempt to identifying demagoguery, the manipulation of the polis' emotions or prejudices to advance a political agenda. Campaign emails are an interesting testing ground for the identification of demagogic rhetoric because they are one of the ways in which politicians connect most directly to their supporters and one of the most obvious examples of an attempt to rouse their supporters to action (or donation). Translating this framework from the field of traditional political science, then, is an interesting chance both to experiment with the efficacy of automated rhetoric analysis and to quantify how the metric-driven approach to fundraising might shape its form, for better or for worse.

2. Problem Background & Related Work

2.1. Crawler & Data

The methodological inspiration for this work comes from a 2015 study called "I never signed up for this! Privacy implications of email tracking."[5] This study used a crawler built on the OpenWPM Framework[1], a tool for deploying automated crawls for privacy measurements, in conjunction with a custom form-filler in order to sign up for mailing lists on ecommerce sites in order to analyze the presence of trackers and address-leaking URLs in the resulting emails. The crawler and the form-filling scripts used in this work are adapted from the version built for that study, while the data analysis tools are novel but inspired by the questions of the original work.

The crawler takes as input a list of campaign websites. Critically, these websites must belong to political campaigns likely to maintain an active mailing list, which is almost always synonymous with those campaigns in the midst of an election cycle. Many previous studies have grappled with

the challenge of acquiring data for “live” studies on political campaign emailing tactics, since signing up for many mailing lists can be so time-intensive as to take the entire election cycle. For example, in the 2014 study “I Get By With a Little Help from My Friends: Leveraging Campaign Resources to Maximize Congressional Power,”[2] the authors required nine months, up until and through the election, to sign up for the 2,903 email lists they eventually subscribed to, even with the aid of "a large team of dedicated undergaduates" on the job.

2.2. Rhetoric Analysis

The rhetorical analysis component of this work drew heavily from both theoretical work in political rhetoric analysis and quantitative work in text classification. Its staunchest underpinning in political theory comes from the text Rhetoric and Demagoguery[4] , which argues for a rhetoric-based approach to identifying demagoguery contemporaneously (rather than only with the benefit of historical perspective, as has been the prevailing method). The single factor the author identifies as most indicative of demagogic rhetoric is a speaker’s reliance on in- and out-group identity. Demagogues, she argues, are those speakers who attribute the woes of some ingroup to the actions or mere existence of an identified outgroup. Then, demagogues can appeal to their own in-group identity as a source of credibility, or use an opponent’s outgroup status to discredit them, as well as relying on the use of fear and exaggeration of the threat posed by the outgroup in order to paint action as imperative and hesitation as disloyalty. This framework was the inspiration for many of the linguistic patterns searched for in the emails. Additional guidance on the question of translating higher-level notions of group-constructing rhetoric into specific and detectable linguistic patterns came from the paper “Seeking influence through characterizing self-categories: An analysis of anti-abortionist rhetoric,”[3] in which the author hand-coded a speech to identify rhetorical arguments that employed self-categorizing rhetoric.

Work from the realm of computer science was helpful in validating some of the specific methodological translations of these frameworks. One specific linguistic construction indicative of othering language was suggested by the 2018 paper ‘The Enemy Among Us’: Detecting Hate Speech with

Threats-Based Othering Language Embeddings.'[6] The work defines two-sided pronoun patterns, or constructions of the form "They [....] us," "We [....] them," and similar, and shows that inclusion of a feature quantifying the existence of such patterns improved the performance of a hate-speech classifier, suggesting that these constructions occur frequently in language that fosters or appeals to prejudice. The effectiveness of traditional sentiment analysis in order to categorize the sentiment of political text was established in the paper 'Affective News: The Automated Coding of Sentiment in Political Texts,'[7] This study attempted this coding with multiple sentiment dictionaries and showed that predicted labels for political texts based on the Lexicoder Sentiment Dictionary were the closest to human-predicted labels, thereby giving guidance on the most promising sentiment dictionary to use for the task.

3. Approach

The aim of constructing a web-crawler to sign up for political mailing lists is twofold: first, to validate the idea that a crawler is a feasible way to automate the oft-repeated task of signing up for emails in order to create a corpora; and second, to acquire and analyze an interesting corpora of data to which automated methods are not frequently applied. To this end, the crawler was developed concurrently with the analysis methods, so that one could benefit from the other. As the crawler and the custom form-filler were adapted and brought up to date with current libraries and a new OS, analysis scripts were developed and tested on a separate corpora of emails accumulated through manual sign-ups on the websites of the 2020 Democratic candidates for President.

Since the data analysis of the emails is more exploratory than hypothesis-driven at this stage, the data pipeline and analysis scripts were developed to allow for maximum flexibility in cross-factor comparison. In particular, the data pipeline was constructed to allow a researcher to filter by election type — federal, state, or local — and other factors in order to understand how tracker prevalence and rhetorical strategies differed between campaigns with differential resources and scopes. The analysis, built on the manually-created inbox and then validated on the crawler-generated inbox, explored many different properties of the data, including categorizing rates of email sending, the

presence of third-party trackers, and the rhetorical approaches employed in each. These explored factors offered the chance to better understand the landscape of canvassing emails in politics, as well as to begin to understand how the presence of open-tracking emails might have shaped the development of that landscape and the approaches employed within.

4. Methods

4.1. Crawler

In service of the ultimate aim of making a tool that would be useful for this work and for possibly for other researchers in the future, the formats for tools and scripts were chosen to offer maximum flexibility and ease of use for both the crawler and the data pipeline. As discussed above, the crawler is a modified version of the one designed for the paper "I Never Signed Up For This," which was based on the OpenWPM framework. The OpenWPM version is designed to run on Ubuntu, but this work modified the launcher to ensure compatibility with Mac OSX.

During its operation, the crawler uses the Selenium library to launch a headless Firefox instance and then navigate through a list of websites, calling a custom form filler on each. The custom form filler looks for page elements with features that correspond to email sign-ups: forms with titles like "Subscribe," "Email Updates," or "Stay Connected," and without words like "Make an Account," or "Log in." If such a form is identified, the script fills the fields with the appropriate information, including an email address generated with a custom post-fix that identifies the site. Fortunately, postfixing can be easily accomplished with use of the "+" character because GMail routes all emails addressed to "address+anything@gmail.com" to "address@gmail.com". Such tagging is desirable because it allows for the detection of email-list-sharing between senders, as well as ensuring that a strange "From" address can always be linked back to the corresponding website where the mailing list was subscribed to. If no such form is identified, the crawler navigates to other internal links found on the page, iterating through other pages on the site until a successful signup is made or until a timeout occurs when 20 seconds have passed.

4.2. Data Acquisition

The only input data required for the crawler is, essentially, a list of urls. Perhaps predictably, given this work's goal of overcoming the temporal challenge of political email collection, one of the biggest stumbling blocks was finding large lists of live campaign websites for bulk sign-up. For an automated approach, an approach that involved manually assembling a list by Googling every race of interest would slow the process significantly and require an investment of manual labor that the method attempts to evade in the first place. Therefore, this work sought out existing lists of campaign emails from different locations, including considering previous academic studies, marketing company websites, and election data clearinghouses like Ballotpedia as potential sources of data.

The first and largest single list of campaign websites located was the directory available on ActBlue. ActBlue is the grassroots fundraising platform of choice for most major Democratic candidates — in 2018, 55% of all Democratic donations from individual donors passed through ActBlue — and the company maintains a full directory of the candidates who use its service on its site. While the list isn't available for download, the format of the site rendered the data easy to scrape for the relevant information about each candidate. Because not all listed candidates had links to their campaign website, scraping the 7,590 candidate entries led to 4,094 websites. This rate roughly aligns with the one estimated manually for the proportion of candidates who provided a link, and since the format for presenting the elements was totally standard between pages, so it seems safe to assume that this represents all candidates whose website was listed. It bears keeping in mind that the sample of candidates who utilize ActBlue is already likely to overrepresent more sophisticated and organized campaigns, a fact only amplified by the selective bias of which campaigns choose to upload a link and which do not. Unfortunately, this type of sampling bias is somewhat unavoidable because more sophisticated campaigns are also more likely to have an easy-to-access mailing list signup and more likely to send frequent mail. Since this study does not purport to draw conclusions about the nature of every campaign in the U.S., such a sampling bias seems, in this case, permissible.

The second source of data was an existing .csv file accumulated by researchers at Comparitech

who undertook a survey of the security (specifically the http / https status) of campaign websites in multiple countries. The .csv file of American candidates had 596 entries of Republican and Democratic candidates for the House, Senate, and Governorship. After cross-referencing these entries with the list of successful sign-ups, the list was reduced to 393 candidates whose email lists had not already been subscribed to. For both sources of data, the urls were fed to the crawler, which attempted to sign up for an email list on each and recorded the email address used for successful sign-ups along with the url.

4.3. Email Data Pipeline

Once email handles had been saved for each successful sign-up, each handle was matched with correlating candidate information and inserted into a MySQL database. The information for each candidate included their name, the position for which and the state in which they were running, and their party. This information was obtained along with the urls from the data sources listed above and packaged with the email handles to allow the analysis scripts to filter the emails and retrieve subsets corresponding to some feature of interest.

In the case of ActBlue, the additional candidate data was scraped from the page along with the website. ActBlue provided all three factors (candidate name, race, and location). However, the information was scraped about two weeks after the initial list of websites had been generated, and turnover was so high that nearly 1,182 of the websites could not be matched with complete information on their candidate. These emails are still included in the corpus for full-text statistics but are omitted from those queries which filter based on candidate information.

For the Comparitech data, however, the .csv included the name of the candidate, but lacked information about their position or the location. Sometimes this information could be reconstructed from their title (those whose names appeared preceded by “Senator” were labelled Senate candidates) or the url of their site (candidates with sites like “candidatenameforcongress.com” were labelled congressional candidates), but some had to be entered into the database with location or race “unknown.” For the majority of the analyses, which examined statistics based on these factors, such

candidates therefore had to be discarded, but for all-inbox statistics these emails could remain in the corpus.

Definitionally, the crawler running successfully meant that an email address began to receive emails from multiple campaigns. The use of a GMail address was chosen partially because it allows the export of emails en masse in the format of an .mbox file, a platform-independent directory of emails. This means that the data pipeline constructed to analyze the emails is compatible with any email provider that allows for exports in .mbox format.

The aim of the aforementioned email pipeline was to make it easy to compare the presence of patterns in tracking or language use across the corpora or filtered by a feature of interest. To that end, the initial steps required cleaning the data and outputting it into useful formats. The first useful output file format is as a .html file which preserves the invisible elements of interest, like tracking pixels. Reading the emails into HTML format was relatively simple using Jinja template matching. The second useful output format was as a .txt file of raw text, which allows for natural language analysis on the subject line and body of the email. Parsing the raw text required resolving encoding issues and removing special characters and artefacts of HTML tags. Both the HTML and the text files are named with a prefix specific to the sender, and the prefix corresponding to each sender's files is stored in the database along with the candidate information, making it easy to return the HTML or text files corresponding to one or many candidates of interest.

4.4. Data Analysis

In order to analyze the presence of trackers in the emails by certain senders, the HTML is read from the files corresponding to each sender and then parsed into an XML tree. To find tracking pixels, the XML is searched for image elements whose width and height are both 1px. Upon detection, the image URL is saved, since the site from which the image is loaded is the one aggregating the tracking information based on the calls to load the image. URLs are then stripped down to their base so that the frequency of a single hosting site across all emails can be monitored.

The methods used to investigate the text data were largely exploratory. Since there were no "true

labels" against which to train a classifier, supervised learning was not a good fit for the task. Instead, several types of unsupervised learning were experimented with, including topic modelling and k-means clustering with bag-of-words features. Neither of these models were ultimately revealing, since both detected patterns in the data that were not ultimately of interest, like the the frequency of phrases like "Thank you," and "Welcome to the Team."

This similarity was investigated by way of n-gram matching, wherein the literal sequences of words were compared between emails. This was part of an attempt to identify the use of email templates by analyzing long shared n-grams that ultimately proved unfruitful. Many of a candidate's emails has a lot of linguistic similarity to the other emails sent by their account, but that similarity diminished drastically for emails from all other senders. This would seem to indicate that most candidates do not rely on templates to draft their emails, but there are other possible explanations for the lack of n-gram matching as well.

Since the questions of interest in this domain are so specific, it made sense to move towards manual feature selection informed by both the language of the frameworks of interest and by manual observation and tagging of emails of interest. After examining the linguistic regularities of interest, a method of counting the occurrences of categorized regular expressions was implemented. When used correctly, regular expressions can capture a variety of convergent linguistic patterns, and the grouping of thematically-related regular expressions provided more flexibility in the kind and number of patterns that could be matched. Regular expressions were constructed with an approach that mixed theory and experimentation: as discussed above, they were informed by the literature and by observed patterns in the data; but each category of expressions were tested and amended manually to ensure that it had a high rate of true positives and a low rate of false positives for the argumentative structure of interest. 1 gives an example of how some of the features of interest might have been coded in a sample email.

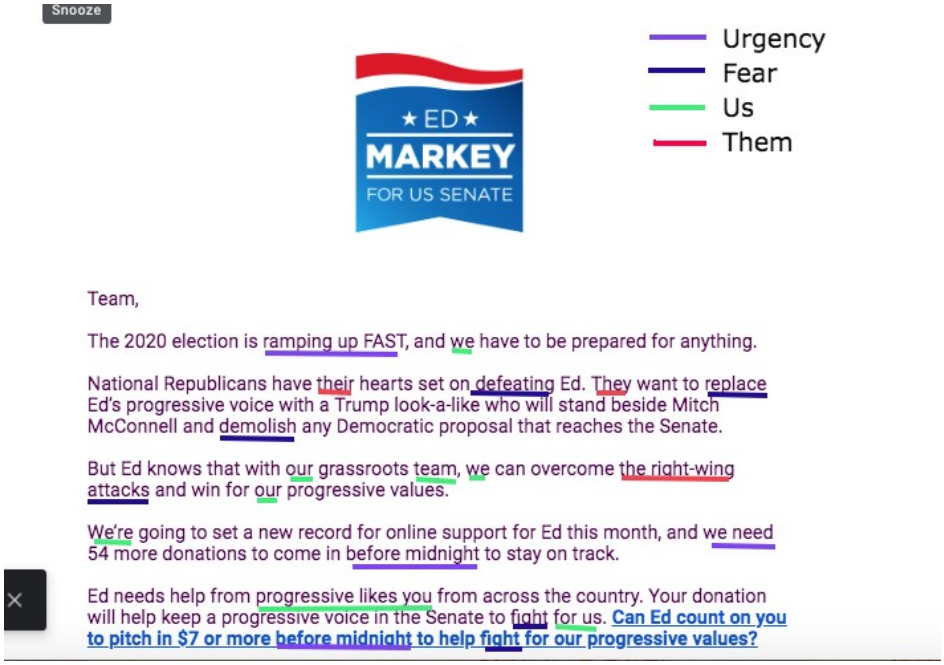

Figure 1: Manual coding of constructions of interest in a sample email.

5. Results & Evaluation

5.1. Crawler

There are a few possible benchmarks by which to evaluate the success of the crawler. Of the 4,094 websites aggregated from the ActBlue directory, the crawler successfully submitted a form on 1,482. This 36% conversion rate seems, on its face, troubling. However, of the 393 sites from the Comparitech study, 296 of them led to successful form submissions, a conversion rate of 75%. The Comparitech list is made up of the websites of candidates for federal office, while the ActBlue list is made up of candidates from the Presidential to the hyper-local. It seems highly likely that more local candidates are less likely to have a sophisticated campaign apparatus that includes a mailing list. This conversion rate was validated through a controlled sample of ten presidential campaigns whose websites were known to have mailing lists available. Of these ten, the crawler successfully converted eight, suggesting that a figure in the ballpark of 70 - 80% is the likeliest value for its success rate. Both of the two failed sites had a donation popup immediately on the landing page, meaning that this is the likely source of the crawler error. Still, even at 70% accuracy, the crawler offers a significant time benefit since it takes, on average, just 28 seconds per site.

The next metric of import is the conversion rate from each of these sign-ups to the actual reception of emails. Total sign-ups from each of the above lists yielded 1,779 potential senders. At present, the mailbox contains emails from 252 unique senders whose information is known, and an additional 199 who have sent emails but whose information is no longer available on ActBlue. Table 1 shows the percentage of sign-ups that have led to emails, as well as the number of emails received, by type of election. Overall, the inbox has accumulated over 1,200 emails, and averages 7 emails per hour during business hours.

Election Type	No. Sign-Ups	No. Senders
President	17	8
U.S. Senate	84	37
U.S. House	248	65
Governor	13	2
State House	278	68

Table 1: Sign-Ups and Senders by Election Type.

5.2. Tracker Prevalence

The analysis of the trackers present on the emails revealed that tracking is common, especially among more sophisticated campaigns. Chart 2 shows the proportion of emails that contained a one-pixel by one-pixel tracking image for each type of election race in the crawler corpus. Table **??** shows the origins of the most-common non-proprietary origins of the trackers. Interestingly, most presidential campaigns used tracking pixels whose origin was their own domain. It stands to reason that larger campaigns, with the resources to analyze and make use of the email-open data, might have more incentive to deploy such tracking techniques. This indicates that there may be a divide between those candidates who rely on an external service to manage their email canvassing and those whose campaign email apparatus is built by an in-house team.

5.3. Rhetoric Analysis

By the nature of the exploratory nature of this analysis, there is no neat way to benchmark its performance or success except to the degree to which it reveals an interesting pattern in the

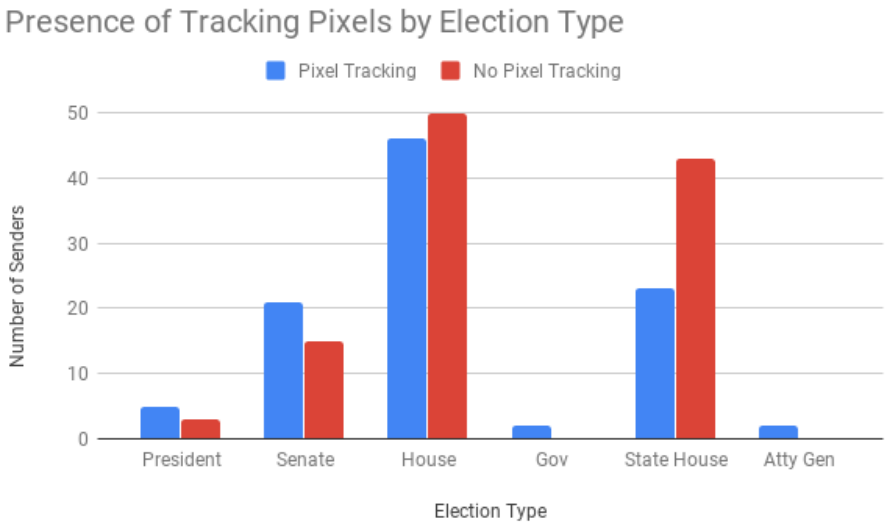

Figure 2: Presence of tracking pixels in email for senders of each election type in crawler data.

Tracker Origin	Prevalence
list-manage.com	0.08723747981
nationbuilder.com	0.05977382876
sendgrid.net	0.02907915994
doubleclick.net	0.02100161551
nationsend15.com	0.01938610662
wix.com	0.01938610662
actionkit.com	0.01938610662

Table 2: Tracker frequency by domain origin.

data. Some of these interesting patterns are laid out below. The rhetoric-analysis code has been demonstrated on the large corpus of emails from the 2020 Democratic Presidential candidates, since these figures offer an interesting insight into the rhetoric used by the most well-oiled (and well-known) campaign machines, as well as providing a means for narrowing an otherwise-broad dataset.

The metric ultimately arrived at to serve as a preliminary approximation of the demagoguery score of a given sender was to analyze the presence of three patterns in their emails. The first pattern is the presence of two-sided pronoun constructions. These are defined as single sentences which use some both variant of the word "us" and of the word "them." The second is the use of fear-words, or words that convey a sense of existential threat. The third is based on the constructions that the candidate uses in which they self-identify as an underdog or a candidate unlike all the others. This

metric simply measures the relative frequency of these constructions in the average of all emails the candidate has sent.

Another area of investigation was the relative frequency with which candidates discussed categories of issues. Regular expressions were constructed to detect common constructions when discussing different political issues. Each section of the bar corresponds to an issue category, broadly constructed based on observational occurrence of the topics and the associated vocabulary as it occurred in a small sample of the email bodies.

5.4. Rhetorical Analysis: 2020 Candidates

The following figures display the prevalence of the aforementioned factors on a subsample of the 2020 candidates for the Democratic nomination for President. Each differently-colored section of the bar corresponds to a different regular expression category. Figure 3 combines three features of interest in the rhetorical definition of demagoguery.

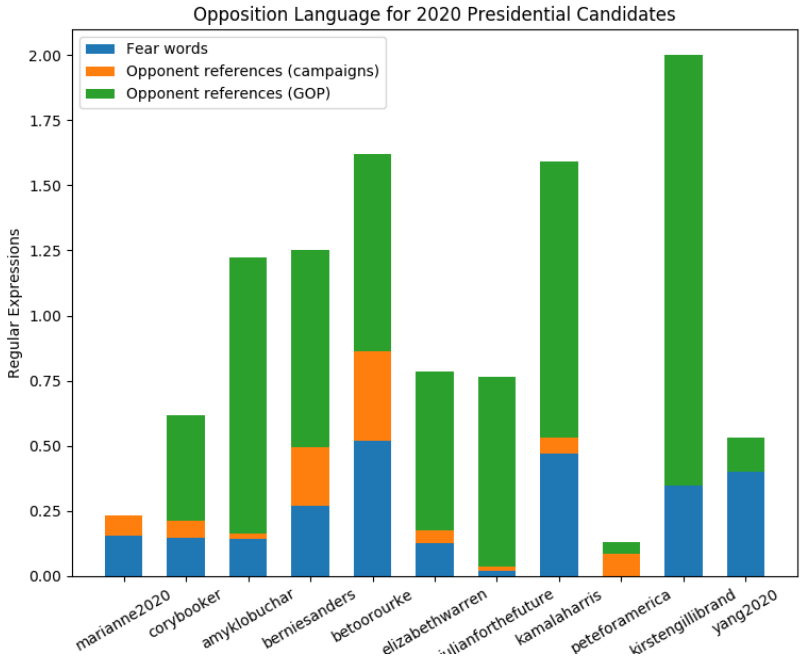

Figure 3: Prevalence of two-sided pronouns and fear words in 2020 Presidential primary candidates.

Figure 4 demonstrates the relative frequency with which candidates used issue-specific words in their emails.

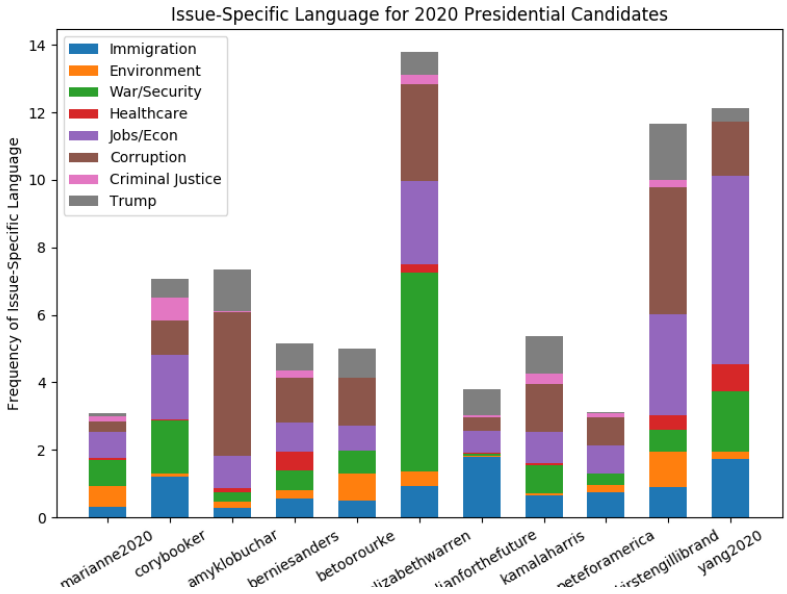

Figure 4: Most-used issue-specific language for in 2020 Presidential primary candidate emails.

6. Conclusions

The major takeaway from this work is that computational methods are, if not genuinely required, at least deeply useful when investigating the patterns present in political communications like emails that themselves take advantage of computational methods. The crawler significantly speeds up the process of signing up for emails. At its current speed of a signup every 28 seconds, it would be able to crawl the 2,903 websites of the study "I Get By With a Little Help From My Friends" in under 23 hours — much better than 8 months — if allowed to run continuously. The crawler's integration with a full data pipeline also offers increased performance with decreased effort on the behalf of a researcher. The integration with the MySQL database, while time-consuming, was ultimately worthwhile for its facilitation of interchangeable scripts for analysis and for the clarity it lends to the structure of the data.

Despite the crawler's promise for making feasible this type of data collection in the future, significant stumbling blocks remain. The imperfect conversion rate for the crawler is frustrating, since it might require researchers to check the negatives (cases where the crawler did not detect a signup form) in order to ensure that they are true negatives rather than a failure of the crawler,

mitigating some of the time savings it promises. Acquiring appropriate data is also a nontrivial difficulty. Writing a scraper to retrieve data from online directories could easily be prohibitive for many political scientists, and, while data can sometimes be appropriated from other studies in the realm of political science, the existence of such studies, as well as their relevance to the question at hand, cannot be guaranteed. In addition, the difficulties encountered in real data — for example, the fact that it changes week-to-week on ActBlue — mean that some amount of manual verification is likely inevitable.

The fact that the business of tracking email-opening in politics is relatively evenly split between large companies and in-house teams is another interesting finding. It seems likely that such tracking pixels would be embedded by the party responsible for, or at least involved in, the creation of the email, and so tracing the origin of single-pixel trackers may allow us to guess at the layout of the email marketing landscape in politics at large. This model is also interesting in that it is significantly different from that of the private sector, where the most common third parties are large advertising firms, and many fewer sites handle tracking in-house.

7. Future Work

One of the most obvious directions for future work stems from the fact that even after the cessation of the project, the inbox will continue to accrue emails. Having more emails would lend further credence to all of the discovered results, especially those that make comparisons about a category based on a relatively smaller number of senders. In addition, it would be interesting to diversify the sources of data to include an more equivalent mass of candidates from the Republican party. There are major differences in the ways that Republicans and Democrats fundraise — as evidenced by the lack of a Republican equivalent to ActBlue — which likely leads to differential patterns of email-open tracking that ought to be catalogued in any sort of complete taxonomy of the space. Rhetoric is also likely to vary significantly between the two parties, which is both worthy of study in its own right, and might provide an interesting anchoring point to better understand intra-party rhetorical differences as well as inter-party ones.

There are also improvements to the crawler that might make it even more effective for reducing manual labor and improving the speed of sign-up. Right now, the crawler struggles with a few specific scenarios, like landing-page pop-ups or the presence of depreciated elements. By fixing these errors, the success rate might be raised to a level that would make it feasible for the crawler to operate with less supervision. It would also be beneficial to spend more time attempting to reconstruct the information lost with the updates to the list of ActBlue sites. Simple scraping might be sufficient to link the unlabelled candidates from the Comparitech data with the appropriate position, perhaps through web scraping.

The analysis of tracking could be made more robust in several ways. Currently, the focus is on email-open tracking. The paper "I never signed up for this!," however, also studied the leakage of email addresses through third-party links which contain hashed identifiers in the url itself. It would be interesting to compare the identifiers from different emails with trackers placed by the same company, to figure out whether signups using a shared email are connected across different campaigns by the marketing provider.

The area of the rhetorical analysis is perhaps the one with the widest-ranging possibilities for future work. This task, lacking clear labels, is an imperfect fit for many of the performance-oriented techniques in machine learning, but the application of newer unsupervised models might allow for better extraction of meaning and sentiment from the text, which would in turn facilitate the application of more richer theories of rhetorical analysis. It would be interesting to take a more nuanced approach to issue detection — the current system is primarily keyword-based, but a more sophisticated model could conceivably identify subtle references that omit such keywords, or perhaps detect the sentiment of the context in which references to a specific issue occur.

The demagoguery score could also be made more robust by incorporating other frameworks for the automatic detection of rhetorical persuasion. More exposure to data might also help with the manual selection of features to ensure the rate of true-positives is maximized without incurring too many false positives. Having a fully-integrated demagoguery score would allow researchers to identify demagoguery in individual candidates as well as comparing the levels of demagoguery

between types of election, locations, or parties. Such a model could potentially be validated on speeches by historical demagogues, because a performance that aligns with our intuitions in a known domain would add credence to the metric's judgements in the more unknown realm of contemporary politics.

In general, political emails are a rich grounds for study because they are one of the clearest examples available of a politician (or their marketing team) trying to incite direct action from their supporters. They very well may be, then, a place where politicians are incentivized to be at their most demagogic: they want supporters to be moved by irrepressible passion towards the donation link, and they don't have to worry about alienating non-supporters who are unlikely to be on their email list in the first place. Emails, then, are an important piece of the puzzle in understanding the political climate of the United States and the way that new technologies intersect with human instincts to create a confusing environment rich for further study.

References

[1] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of ACM CCS 2016*, 2016.

[2] J. M. B.-S. et. al., "I get by with a little help from my friends: Leveraging campaign resources to maximize congressional power," in *114th Annual Meeting of the American Political Science Association*, 2019.

[3] S. Reicher and N. Hopkins, "Seeking influence through characterizing self-categories: An analysis of anti-abortionist rhetoric," in *British Journal of Social Psychology*, 1996.

[4] P. Roberts-Miler, *Rhetoric and Demagoguery*, 1st ed. Southern Illinois University Press, 2019.

[5] J. H. Steven Englehardt and A. Narayanan, "I never signed up for this! privacy implications of email tracking," in *Proceedings on Privacy Enhancing Technologies*, 2018.

[6] H. L. Wafa Alorainy, Pete Burnap and M. Williams, "The enemy among us: Detecting hate speech with threats based 'othering' language embeddings," in *arXiv.org*, 2018.

[7] L. Young and S. Soroka, "Affective news: The automated coding of sentiment in political texts," 2012.