Using Text Analysis and Machine Learning to Classify U.S. Governors' COVID-19 Press Briefings

Authors

Jason Marshall, Binghamton U.-State U. of New York, jmarsha7@binghamton.edu
Francis J. Yammarino, Binghamton U.-State U. of New York, fjyammo@binghamton.edu
Srikanth Parameswaran, Binghamton U.-State U. of New York, sparames@binghamton.edu
Minyoung Cheong, Pennsylvania State U., Great Valley, mxc1016@psu.edu

Jason Marshall, Binghamton U.-State U. of New York, jmarsha7@binghamton.edu
Francis J. Yammarino, Srikanth Parameswaran, Minyoung Cheong

Using Text Analysis and Machine Learning to Classify U.S. Governors' COVID-19 Press Briefings

ABSTRACT

Increased computing power and greater access to online data have led to rapid growth in the use of computer-aided text analysis (CATA) and machine learning methods. Using "Big Data", researchers have not only advanced new streams of research, but also new research methodologies. Noting this trend, while simultaneously recognizing the value of traditional research methods, we lay out a methodology to bridge the gap between old and new approaches. With a combination of web scraping, CATA, and supervised machine learning, using ground truth data, we train a model to predict CIP (Charismatic-Ideological-Pragmatic) leadership styles from running text. To illustrate this method, we apply the model to classify U.S. state governors' COVID-19 press briefings according to their CIP leadership style. In addition, we demonstrate content and convergent validity of the method.

Keywords:

Text analysis; web scraping; machine learning; leadership; COVID-19

Jason Marshall, Binghamton U.-State U. of New York, jmarsha7@binghamton.edu
Francis J. Yammarino, Srikanth Parameswaran, Minyoung Cheong

Using Text Analysis and Machine Learning to Classify U.S. Governors' COVID-19 Press Briefings

As the internet is crawling with data (pun intended), researchers have begun to leverage this "Big Data" source (Braun, Kuljanin, & DeShon, 2018) to advance theoretical understanding in a variety of research contexts. Advanced computing capabilities and modern analytic techniques have made this possible. For example, with web scraping techniques, researchers can access data directly from a web page to be analyzed on their personal computer. In addition to numerical data, there is a plethora of textual data available on the internet. Again, accessing this data is as easy as executing a few lines of code in your favorite programing language (e.g., *Python* or *R*). This ease of access has corresponded with an increase in text analysis studies in recent years (Banks, Woznyj, Wesslen, & Ross, 2018). Specifically, computer-aided text analysis (CATA; McKenny, Aguinis, Short, & Anglin, 2018; McKenny, Short, & Payne, 2012; Short, Broberg, Cogliser, & Brigham, 2010) has grown more popular. In brief, CATA provides rich insights into individual cognitions, values, and identities in ways that cannot be replicated via traditional research methods, such as self-report surveys (Pollach, 2012).

Beyond traditional CATA, which typically utilizes deductive dictionary-based approaches to text analysis, newer machine learning techniques, including text mining (Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018), natural language processing (NLP; Pandey & Pandey, 2019), and neural network models (DeTienne, DeTienne, & Joshi, 2003; Minbashian, Bright, & Bird, 2010), are gaining traction. There are two basic forms of machine learning: supervised and unsupervised. In unsupervised machine learning, the machine is not aware of any preexisting categories, dictionaries, or classifications, but merely uses a form of cluster analysis, based on co-occurrence matrices to establish which groups of words cluster

together. In supervised machine learning, the researcher provides an input and output, then allows the machine to learn how to best fit a model to match the two (Janasik, Honkela, & Bruun, 2009).

With all the new advanced data analysis tools and techniques, and the relatively easy access to Big Data, one might be compelled to forget all about traditional research methods and move on to the shiny new toys. However, we offer an alternative perspective. Rather than take an "out with the old and in with the new" approach, we believe a better method is to leverage existing ground truth datasets (i.e., the old) and apply web scraping, text analysis, and supervised machine learning techniques to advance new streams of research. To do so, the current study illustrates how to use a combination of web scraping, text analysis, and machine learning techniques to classify categorical outcomes.

CURRENT STUDY: CIP STYLE IN U.S. GOVERNORS' COVID-19 PRESS BRIEFINGS

The particular categorical outcome we have chosen to model is the CIP (Charismatic-Ideological-Pragmatic) outstanding leadership style approaches (Lovelace, Neely, Allen, & Hunter, 2019; Mumford, 2006) of U.S. state governors. With the current COVID-19 pandemic, the effects of leadership (or lack thereof) have been magnified. As Bligh and colleagues noted, "People become increasingly susceptible to the leader and his or her vision in the wake of a crisis" (Bligh, Kohles, & Meindl, 2004: 215). The pandemic is certainly a crisis and leaders (e.g., U.S. state governors) have been frequently communicating their visions via press briefings. Accordingly, in this study, we seek to categorize the U.S. state governors' press briefings according to their CIP leadership style in hopes of not only advancing a new methodology, but also providing data analysis that may be of interest/value to the general public. The latter

aspiration is a direct response to a call from Aguinis and colleagues (Aguinis, Suarez-Gonzalez, Lannelongue, & Joo, in press) to reach an audience outside of academia.

CIP MODEL OF OUTSTANDING LEADERSHIP

Mumford (2006) developed the CIP model of outstanding leadership based on early work from Max Weber (1924). According to the model, there are three different leadership styles that are all equally likely to produce outstanding leadership. The "C" in the model stands for *charismatic*. Charismatic leaders are future-oriented, use positive emotional imagery, and are concerned with motivating followers. Regarding problems solving, charismatics are best during the middle problem-solving phase when their ability to generate possibilities is most valuable (Lovelace et al., 2019). The "I" in the model stands for *ideological*. Ideological leaders (ideologues) are past-oriented, values-driven, and use negative emotion in their communication (Hunter, Cushenbery, Thoroughgood, Johnson, & Ligon, 2011). During times of crisis, ideologues perceive situations as the causal mechanism, focus on changing the system, and may rally their base constituency (i.e., like-minded individuals), rather than attempt to appeal to the masses (Griffith et al., 2018). With regard to their problem-solving value, ideologues tend to be best during late problem-solving phase when their ability to implement solutions is most needed (Lovelace et al., 2019). Finally, the "P" in the model stands for *pragmatic*. Pragmatic leaders are present-focused, communicate using rational appeals, and are best during the early phase of problem-solving when data are being gathered and evaluated. During times of crisis, pragmatic leaders adopt an interactionist approach, focusing how the situation affects people and their behavior (Griffith et al., 2018; Lovelace et al., 2019). For each of these styles, leaders can be either personalized or socialized (Yammarino, Mumford, Serban, & Shirreffs, 2013); however, for simplicity, this study focuses only on the CIP aspects of this leadership approach.

Prior CIP studies have primarily used historiometric methods (Ligon, Harris, & Hunter, 2012) to assess the CIP style in outstanding historical leaders such as presidents (Yammarino et al., 2013), world leaders (Serban et al., 2018), and college/NFL football coaches (Hunter et al., 2011). In other words, these researchers have used historical documents, such as biographies, to code whether an individual is C, I, or P. Recently, Lovelace and colleagues (2019) stated that research on the CIP model has been over-reliant on the historiometric approach and claimed this may be limiting the potential applications of the model.

There are a few reasons why the reliance on historiometric methods is limiting the study of CIP in organizations. First, the need for historical documents (e.g., biographies) to code means that a researcher must wait until after the leadership has already happened, often years after, to assess which style the leader employed. Second, this method is not flexible in its ability to detect situational variation in C, I, or P style. Theory (i.e., Mumford, 2006) suggests that the style remains relatively stable; however, as Yammarino, Sotak, and Serban (2020) noted, there may be some situational constraints on this stability. For example, Yammarino et al. (2020) stated that Barrack Obama was a C (i.e., campaigning as a C) prior to becoming President, at which point he became a P (i.e., governing as a P). Third, Yammarino et al. (2020) also called attention to the idea that leaders may not necessarily be purely C, I, or P, but rather a combination of two or three of the styles. Again, the historiometric method is limited in its ability to detect these nuances. Fourth, historiometric methods are labor-intensive to employ (Lovelace et al., 2019). Researchers spend dozens of hours learning how to code the documents. Then, they spend countless hours reading over and manually coding documents. Then, the documents are checked for interrater reliability. Put simply, there must be another way to assess CIP leadership styles.

Thus, to address the limitations of the historiometric approach, which the CIP model has heavily relied upon, we advance a more granular, flexible, and timely operationalization of the CIP model. Using computer-aided text analysis (CATA; Banks et al., 2018; McKenny et al., 2012; Short et al., 2010) will not only allow researchers to code large volumes of text in a fraction of the time, but it will also allow us to study running-text (e.g., speeches, interviews, meeting transcripts, tweets) in current leaders, without waiting for biographies to be written after the fact. As a side benefit to this study, we might also see which CIP leadership style is most frequently used, and perhaps most effective (pending data collection), in the current pandemic situation.

METHODS

Sample and Selection

Yammarino and colleagues (2013) conducted a historiometric study of United States presidents and classified each president as either C, I, or P. We used Yammarino et al.'s (2013) classifications to build our CIP ground truth dataset for all U.S. presidents (i.e., the entire population of U.S. presidents).

In terms of our model training and testing, the dataset consisted of transcripts from 844 U.S. presidential addresses. Through web scraping techniques, we collected the addresses from "The American Presidency Project", hosted by the University of California (UC) Santa Barbara at https://www.presidency.ucsb.edu/index.php. The addresses included inaugural addresses, state of the union addresses, commencement addresses, remarks at the White House correspondents' dinner, holiday addresses, remarks to the United Nations, remarks to U.S. congress, remarks to foreign governments, and remarks on foreign affairs. The addresses spanned from George Washington, in 1789, to Barrack Obama, in 2016. Data were also collected from Donald Trump;

however, they were not used in this dataset because he was not included in the Yammarino et al. (2013) CIP classifications. As each president was classified by a single CIP style, we applied that same style to each presidential address. For example, Barrack Obama was classified as "P", therefore, each of his presidential addresses was labeled as "P".

Our model was applied to a dataset containing 730 COVID-19 press briefing transcripts, representing 44 United States governors. Through web scraping techniques, we collected the press briefing transcripts from https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts. The press briefings spanned from February 27, 2020 to December 18, 2020, when the data were collected. Regarding political affiliation, 481 of the briefings were from Democrat governors (65.89%) and 249 (34.11%) were from Republican governors. Females delivered 75 of the briefings (10.27%) and males delivered the other 655 (89.73%). It is important to note that the Rev.com transcripts are not all inclusive of every press briefing that occurred during this time frame.

Web Scraping

Web scraping refers to the process of obtaining data directly from a webpage. While some companies develop specific APIs (application programming interfaces) that allow researchers to quickly and easily access data (Braun et al., 2018), many times APIs are not available. For websites without an API, Kobayashi and colleagues recommend using web scraping techniques to acquire the data (Kobayashi et al., 2018). Prior to beginning the web scraping process, it is important to consult the terms of use for the website for which you wish to acquire the data. For example, the terms of use for the Rev website, where we scraped the governors' press briefings, stated restrictions such that a user may not, "take any action that imposes an unreasonable or disproportionately heavy load on the Platform or its infrastructure or

that negatively affects the ability of others to access or use the Platform" (Rev, 2020). This does not explicitly forbid web scraping per se, but rather necessitates the use of responsible web scraping practices, which we will describe in detail below. In addition to reviewing the terms of use, in this particular case, we also received written consent stating, "You're welcome to use the content as you please" (Lawson, 2020).

For this study, we followed Braun and colleagues' (2018) general web scraping process: a) identify the website address(es), b) identify the data on the website to be extracted, c) write a script to extract the data, and d) execute the script to download the data onto a computer.

Identify the website address(es). After you have identified the home page of the website, you must identify the specific pages that contain the data you wish to extract. For example, the website for the governors' COVID-19 press briefings is https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts. Many times, the data are not all available on this single page, but rather several pages. For example, at the time of this writing, there are 103 pages of COVID-19 press briefings on the Rev.com site. Fortunately, for web scraping purposes, most websites establish a stable pattern for labeling these pages (e.g., https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts/page/2; https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts/page/3; https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts/page/4). These patterns enable you to write a programming script, using your preferred programming language (e.g., *R* or *Python*), to access the web pages for scraping.

Identify the data on the website to be extracted. Most web pages consist of some combination of text, images, tables, and/or hyperlinks to other websites or pages. On the front-end, the contents are often displayed in some aesthetically pleasing manner. On the backend, however, it is merely a combination of HTML (HyperText Markup Language) and CSS

(Cascading Style Sheets). A full review of HTML and CSS is beyond the scope of this manuscript (for a detailed overview, see Mitchell, 2018), thus, we simply want to note that most websites use patterns, in the HTML and CSS, that enable researchers to write a programming script to extract the data from the web pages. There is an easy way to access the HTML and CSS to identify these patterns in the web pages: While viewing a web page on a web browser (e.g., Google Chrome), you can highlight a portion of the page you wish to extract, right click, and click "inspect". This will give you access to the HTML and CSS for that particular section of the web page. There you will find the tags and classes needed to write the programming script.

Write a script to extract the data. After you have identified the web page(s) you want to visit and identified patterns in the HTML/CSS, you can write a script to extract the data. We used *BeautifulSoup* (Richardson, 2020) in *Python*; however, there are options in *R* (R Core Team, 2020) as well (for a summary of *R* packages for web scraping, see Braun et al., 2018). Writing the script follows the same general steps as the overall scraping process. After loading your packages (e.g., *BeautifulSoup*, *time*, *requests*, *pandas*) you create a list of the web pages you wish to visit. If the website has multiple pages, like our Rev.com example, you may have to create an empty list and use a for loop to recreate the pattern of the website's pagination to populate your website list. If the pages you visit merely hold preview content with links to the actual data (like our Rev.com example), you must create another empty list, write a for loop to visit each summary page, retrieve the hyperlinks to the pages with the actual data you wish to extract, and store the hyperlinks in the new list. Next, you write a for loop to visit each hyperlink that contains your desired data, retrieve the data from the specified HTML tags/classes, and store the data.

This is the phase of the process where responsible web scraping is critical. As noted in the Rev.com terms of service, websites do not want you to overload their servers with rapid requests. Doing so may not only violate the terms of use, but it may also result in your IP (internet protocol) address being blocked from accessing the website. This would hamper your data collection process. To avoid overloading the server, it is important to incorporate time delays in your web scraping script. The *time* package in *Python* allows you to put your script to "sleep" for a specified number of seconds before resuming the scraping process. This allows you to avoid overloading the server and, more pragmatically, avoid getting blocked from the website. Of course, this was a cursory overview of the script writing process. For a detailed overview of the entire process, using *BeautifulSoup*, see Mitchell's (2018) book.

Execute the script to download the data onto a computer. The final step in the web scraping process is executing the script and downloading the data onto your computer. This phase of the process will be largely dependent on the type of data you are scraping and the format in which the data is received from the website. Using a combination of lists to store the data and post processing techniques (e.g., strip text, split text, relabel variables), you can prepare your data in *Python* (or *R*) to be downloaded/exported in a common data file format such as CSV (comma-separated values) or XLSX (i.e., Microsoft Excel spreadsheet). In *Python*, the *pandas* and *numpy* packages can be used in conjunction to create a data frame and export the data in your desired format. A similar process could be followed in *R*, storing the data in data frames and exporting to CSV or XLSX via the *openxlsx* package.

Computer-Aided Text Analysis (CATA)

CATA enables the operationalization of constructs by transforming textual data into quantitative data, based on word frequency in specific dictionaries or dictionary categories

(McKenny et al., 2018). There are several CATA software programs available, including The General Inquirer (GI; Stone, Bales, Namenwirth, & Ogilvie, 1962), DICTION (Hart & Carroll, 2012), and Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, Boyd, & Francis, 2015). Additionally, there are several CATA packages in *R*, such as *sentimentr, tm, stm, syuhzet,* and *quanteda*. For this study, we chose LIWC for two reasons: 1) there are specific dictionary categories that map onto our CIP construct very well (e.g., focuspresent, focusfuture, focuspast) and 2) LIWC has been used in high-quality journals to operationalize language in the political leadership context (e.g., Sergent & Stajkovic, 2020).

LIWC. We used LIWC2015 (Pennebaker et al., 2015) to analyze both the presidential addresses (for model training) and the governors' COVID-19 press briefings. LIWC2015 is capable of analyzing text in a variety of formats, including plain text, PDF, RTF, .csv, and .xlsx. The program processes words sequentially throughout the text, adding incremental value to the dictionary category (or categories) to which the word belongs. If a word is present in more than one dictionary category, it is counted in all applicable categories. There are approximately 90 variables calculated throughout this process, including 41 dictionary categories designed to tap psychological constructs, grammar/punctuation variables, and other descriptive variables (e.g., word count, words per sentence, percentage of words longer than six letters). For reference, in Table 1, we provide a list of the variables we used in this study, along with sample words and the total number of words in each dictionary category.

Insert Table 1 About Here

Establishing validity. While CATA offers many psychometric advantages over traditional measures, such as self-report surveys and archival data (McKenny et al., 2018), the

validity of the measure must be established (Short et al., 2010). To establish content validity, after analyzing the presidential addresses via LIWC2015, we assessed the differences among CIP style, based on theoretically relevant dictionary categories. For example, we conducted a one-way ANOVA and subsequent post-hoc Tukey test (Tukey, 1977) to assess whether C, I, and P differed, as theory suggests, in time orientation (Mumford, 2006). As you can see in Figure 1, P is significantly higher than I and C in the focuspresent variable, I is significantly higher than C and P in the focuspast variable, and C is significantly higher than I and visually higher, although not significantly, than P in the focusfuture variable. As time orientation is one of the key differentiating aspects among CIP styles, the fact that these LIWC2015 variables also captured these differences suggests that the text-based measure is appropriately differentiating among the CIP styles, thus, confirming the content validity. Additionally, after applying the model to the governors' press briefings, we consulted established CIP scholars[1] (i.e., subject matter experts), who were also familiar with some of the governors' press briefings, to establish face validity of the measure of CIP in governors. The CIP scholars indicated that the measure was assessing CIP in the governors with relative accuracy. As such, this step provided further evidence of content validity (Short et al., 2010).

Insert Figure 1 About Here

As a reminder, the CIP styles of the presidents were established in a human-coded historiometric study (Yammarino et al., 2013). Thus, in addition to establishing support for content validity, these results also suggest convergent validity (Campbell & Fiske, 1959). Regarding external validity, we chose presidential addresses to develop our text-based model

[1] Information related to the consultation with CIP scholars is available from the corresponding author upon request.

because of the similarity in context (i.e., political leadership) and content (addresses and briefings) to our test sample of U.S. state governors' COVID-19 press briefings.

Multinomial Logistic Regression (MLR)

Commonly referred to as a *discrete choice model*, multinomial logistic regression (MLR) is similar to the more oft-used logistic regression, with a primary difference being that it is used when there are more than two factors in the dependent variable. In general, the goal of MLR is to model the odds of a particular categorical outcome, as a function of the covariates (i.e., independent variables) in the model (Hosmer & Lemeshow, 2015). More precisely, we are using MLR, in this study, to construct a model capable of explaining the relationship between the independent variables (LIWC categories) and the known outcome (CIP), so that the model can accurately predict CIP in the governors' COVID-19 press briefings, for which we do not currently have a CIP classification. This process is commonly referred to as supervised machine learning, as both the input and desired output are given to the model (Janasik et al., 2009; Welbers, Van Atteveldt, & Benoit, 2017). While there are multiple *R* packages available to conduct MLR (e.g., *mnlogit*, *VGAM*, *maxent*), we used *nnet* (Ripley & Venables, 2020) for this study.

Model specification. Prior to running an MLR, the researcher has to establish which variables to include in the model. There are multiple ways to select variables and, ultimately, the researcher uses a combination of theory, statistics, and discretion to choose the final model. One recommended approach is to first analyze the effect of each possible covariate on the dependent variable (Hosmer & Lemeshow, 2015). For example, we conducted a one-way ANOVA and post hoc Tukey test (Tukey, 1977) on each of the possible variables to see which variables significantly varied among CIP factors. Then, we created an initial model including all variables

that demonstrated significant variation among CIP factors. After the initial model was created, we used fit statistics (e.g., AIC, likelihood ratio) and theory to determine which covariates ultimately remained in the model (see the appendix for an example of this alternate approach to model building).

Another approach, which we ultimately ended up using to create our final model, is referred to as stepwise model selection (Hosmer & Lemeshow, 2015). As the name suggests, the model is stepped one variable at a time in order to maximize an established outcome criterion. A model can be stepped forward (i.e., adding a variable), backward (i.e., removing a variable), or both forward and backward. Using the *stepAIC* function of the *MASS* package (Ripley et al., 2020) in *R*, we chose to select our variables based on a forward and backward step, which resulted in an optimized AIC (Akaike Information Criterion). Following best practice (Hosmer & Lemeshow, 2015), we examined the recommended variables for their theoretical relevance to our study. After noting that two relevant variables (focusfuture and focuspast) were missing from the optimized model, we conducted a series of tests (using the training and testing methods described below) to determine whether or not to include the variables in the final model. Ultimately, including either focusfuture or focuspast in the model not only resulted in a less parsimonious model (i.e., higher AIC), but also reduced model accuracy. Thus, we chose the model that maximized parsimony (mean AIC = 659.79) and accuracy (mean = 80.44%), rather than including the dictionary categories we presumed, *a priori*, to be relevant. In addition to AIC, we calculated a Hosmer-Lemeshow test (HL test; Hosmer & Lemeshow, 2015) using the *logitof* function of the *generalhoslem* package (Matthew, 2019) in *R*. The HL test is a common model fitness test in multinomial logistic regression that is actually a measure of lack of fit; thus, a non-significant *p*-value indicates a good fitting model (Fagerland & Hosmer, 2017). The HL

test results for our model showed a non-significant (p = .98) chi-squared ($\chi^2 = 6.77$) with 16 degrees of freedom, which indicates a good fit. Following notation from Hastie and colleagues (2009), the resulting model can be expressed in the following two equations:

$$\log\left(\frac{Pr(CIP = I)}{Pr(CIP = C)}\right) = \beta_{10} + \beta_{11}(achieve) + \beta_{12}(adj) \ldots \beta_{154}(you) + \varepsilon \quad (1)$$

$$\log\left(\frac{Pr(CIP = P)}{Pr(CIP = C)}\right) = \beta_{20} + \beta_{21}(achieve) + \beta_{22}(adj) \ldots \beta_{254}(you) + \varepsilon \quad (2)$$

Equation one represents the logit, where a one-unit increase (or decrease) in the independent variable(s) is associated with an increase (or decrease) of the log odds for the model selecting "I" vs. "C". Similarly, equation two represents the logit, where a one-unit increase (or decrease) in the independent variable(s) is associated with an increase (or decrease) of the log odds for the model selecting "P" vs. "C". The factor, "C", was arbitrarily chosen as the baseline for the model. To view the full model, including the coefficients, standard errors, and Wald statistics for each independent variable, see Table 2. Note there are 54 independent variables included in the final model. Sample size guidelines for multinomial logistic regression recommend a minimum of 10 cases per predictor variable (Starkweather & Moske, 2011), thus our sample of 844 presidential addresses is sufficient (i.e., > 540 cases).

Insert Table 2 About Here

Training. After establishing which variables to include in the model, the model must be trained to select the appropriate outcome. To do this, we selected a random subset of the presidential dataset, representing 70% of the total dataset. Using the *multinom* function of the *nnet* package in *R*, we trained the above specified model, using maximum likelihood, on the randomly selected training dataset. In this phase of the process, the objective is to maximize the

accuracy of the model (i.e., how well it predicts the correct known outcome or ground truth), while avoiding overfitting the model. Overfitting, in prediction models, occurs when the researcher overutilizes the available predictor variables when building the model based on the sample/training data (Putka, Beatty, & Reeder, 2018). Unfortunately, it is impossible to know whether a model is overfit until it is applied to the test dataset, which, in this case, was the 30% holdout data from the full presidential address dataset.

Testing. Next, the trained model was applied to the unseen test dataset (i.e., the remaining 30% of the presidential address dataset). After applying the model, a probability score for each outcome factor was generated. Next, we populated a new variable in the dataset where the factor with the highest probability, for each observation, was entered. An accuracy percentage was then calculated based on a comparison of the predicted outcomes to the known outcomes. In Table 3, we show a sample of ten iterations of the training and testing process, including the AIC statistic, accuracy percentage for the training dataset, and accuracy percentage for the testing dataset. The comparison between the accuracy of the model applied to the training data versus the testing data is where evidence of overfitting can be found. A noticeable drop in accuracy from training to testing can be attributed to overfitting the model to the training data. On average, our final model predicted CIP in the test data with 80.44% accuracy.

Insert Table 3 About Here

Applying to new data. After systematically building, training, and testing the model, we finally applied the model to the governors' COVID-19 press briefing dataset. Unlike the presidential dataset, there are no known CIP outcomes in the governor dataset. Hence, there are no accuracy scores to be calculated. Instead, we relied upon the training/testing procedure in the

presidential address dataset to build our confidence in the model's ability to accurately assess CIP in the governor dataset. Ultimately, the model classified the press briefing based on the CIP category with the highest probability of being correct. A subset of the modeled probabilities is provided in Table 4. As an additional robustness check, as mentioned in the validity section above, we shared the results with subject matter experts (i.e., established CIP scholars) to assess whether the model was accurately capturing CIP styles in the governors' press briefings. Again, the scholars agreed that the model appeared to be relatively accurate in its predictions.

Insert Table 4 About Here

RESULTS

A summary of the results is provided in Table 5. In brief, the breakdown of the press briefing CIP categorization was as follows: 7 charismatic, 468 ideological, and 255 pragmatic. Figure 2 shows the within-person variability of the CIP classifications over time. Although there was variability, most governors tended to have a dominant style. For example, Governor Cooper (North Carolina) had 14 briefings classified as P, but also had 4 classified as I. Perhaps the lack of granularity in the historiometric methodology, which has dominated the study of CIP to date (Lovelace et al., 2019), led researchers to perceive and theorize that CIP style is relatively stable in individuals. We attribute the relative lack of charismatic press briefings to the fact that these briefings all took place during a global pandemic. The positive emotional imagery and focus on the future that characterizes the charismatic style of leadership may be difficult to portray, and likely ill-advised, when thousands of individuals are dying on a daily basis.

Insert Table 5 About Here

Insert Figure 2 About Here

DISCUSSION

Theoretical Implications

While the primary focus of this manuscript was to advance a research method for using publicly available textual data to classify categorical outcomes, there are some theoretical implications as well. Similar to the way in which experience sampling methods (ESM; Gabriel et al., 2019) allowed researchers to capture the within-person variation of personality over time (Fleeson, 2001, 2004), our text-based operationalization of CIP will allow researchers to further examine the stability assumptions of CIP that were previously untestable using historiometric methods (Lovelace et al., 2019). Stated differently, our method of assessing CIP from running text allows researchers to examine CIP from multiple levels of analysis (within and between), over time, and across situations, rather than being limited to merely aggregating CIP across time and situations, assuming any within-person fluctuations to be error.

Practical Implications

From a practical standpoint, this method could be extremely valuable for upper echelon leaders. For example, a CEO could use this methodology to classify his/her communication (speeches, interviews, annual reports, etc.) and compare the classification against objective and subjective outcome criteria. With outcome data, a CEO could modify his/her communication style to garner more favorable results. Similarly, politicians (and their speech writers) could use this method to adjust their messaging to win support for political initiatives. We are not advocating for the use of this methodology as a tool for manipulation. Although we acknowledge

this as a possibility, we hope that the method can be used in a positive manner to promote improved communication and connection with a target audience including various stakeholders.

Limitations and Future Research

There are certainly limitations to our study and this method. First, although 80% seems to be high accuracy percentage for predicting CIP, based on running text, there is still a 20% chance that the model predicted the CIP style incorrectly. When a press briefing falls on the margin between two categories, this 20% could lead to unstable classifications among multiple runs of the model. For example, Governor Cuomo seemed to have several press briefings that fell along the margin between "I" and "P" classification. Thus, each iteration of the model would elicit slightly different totals for "I" and "P" classifications for Governor Cuomo. Many of the other governors, however, were more stable in their classifications. For example, Governor Beshear's press briefings in the "P" category consistently ranged between 10 and 12 (out of 12), with very little fluctuation. Nonetheless, this is a limitation of this method (and/or this model). A second limitation to this particular study is the fact that all of the press briefings took place during a global pandemic. The gravity of the situation likely overpowered an individual leader's natural style and necessitated a certain style of communication. While this provides additional theoretical considerations for the study of CIP, it places a constraint on our method's ability to generalize to less situationally constrained times (i.e., normal, non-crisis life).

We put forth this method in hopes of stimulating and generating new streams of future research. By combining web scraping, text analysis, and machine learning techniques there are many possibilities. For example, researchers can use other ground truth datasets in which individuals are classified by certain categorical variables and apply the approach demonstrated in this study. Similarly, researchers could identify ground truth datasets in which individuals are

scored on continuous scales of other psychological constructs (e.g., personality, goal orientation) and apply a similar, modified approach. With a continuous variable as an outcome, rather than using multinomial logistic regression, the researcher would simply train a linear model. Beyond the use of ground truth datasets and archival textual data, new online communication platforms (e.g., Zoom) provide another avenue for utilizing the methods we put forth in this manuscript. For example, researchers could collect psychometric data from participants and have the participants engage in some sort of online Zoom activity (or perhaps use an already recurring Zoom activity, such as a college course), save the transcripts from the Zoom session, and proceed with the above-mentioned methods.

CONCLUSION

There is a blue ocean of text available on the internet. There is also a plethora of research studies conducted on public figures. We put forth a method that combines web scraping, text analysis, and machine learning techniques to classify categorical outcomes. Through the implementation of this methodology, we were able to classify U.S. governors' COVID-19 press briefings according to their CIP leadership styles. Additionally, we opened the door for other researchers to build their research streams using this methodology.

REFERENCES

Aguinis, H., Suarez-Gonzalez, I., Lannelongue, G., & Joo, H. In press. Scholarly impact revisted. ***Academy of Management Perspectives***.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. 2018. A review of best practice recommendations for text analysis in R (and a user-friendly app). ***Journal of Business and Psychology***, 33(4): 445–459.

Bligh, M. C., Kohles, J. C., & Meindl, J. R. 2004. Charisma under crisis: Presidential leadership, rhetoric, and media responses before and after the September 11th terrorist attacks. ***Leadership Quarterly***, 15(2): 211–239.

Braun, M. T., Kuljanin, G., & DeShon, R. P. 2018. Special considerations for the acquisition and wrangling of big data. ***Organizational Research Methods***, 21(3): 633–659.

Campbell, D. T., & Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. ***Psychological Bulletin***, 56(2): 81–105.

DeTienne, K. B., DeTienne, D. H., & Joshi, S. A. 2003. Neural networks as statistical tools for business researchers. ***Organizational Research Methods***, 6(2): 236–265.

Fagerland, M. W., & Hosmer, D. W. 2017. How to test for goodness of fit in ordinal logistic regression models. ***Stata Journal***, 17(3): 668–686.

Fleeson, W. 2001. Toward a structure- and process-integrated view of personality: Traits as density distributions of states. ***Journal of Personality and Social Psychology***, 80(6): 1011–1027.

Fleeson, W. 2004. Moving personality beyond the person-situation debate. ***Current Directions in Psychological Science***, 13(2): 83–87.

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., et al. 2019. Experience

sampling methods: A discussion of critical trends and considerations for scholarly advancement. ***Organizational Research Methods***, 22(4): 969–1006.

Griffith, J. A., Gibson, C., Medeiros, K., MacDougall, A., Hardy, J., et al. 2018. Are you thinking what I'm thinking?: The influence of leader style, distance, and leader–follower mental model congruence on creative performance. ***Journal of Leadership and Organizational Studies***, 25(2): 153–170.

Hart, R. P., & Carroll, C. E. 2012. ***DICTION: The text analysis program: Help manual***. Digitext, Inc.

Hastie, T., Tibshirani, R. J., & Friedman, J. 2009. ***The elements of statistical learning: Data mining, inference, and prediction*** (2nd ed.). Springer.

Hosmer, D. W., & Lemeshow, S. 2015. ***Applied logistic regression***. (N. A. C. Cressie, N. I. Fisher, I. M. Johnstone, J. B. Kadane, D. W. Scott, et al., Eds.) (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Hunter, S. T., Cushenbery, L., Thoroughgood, C., Johnson, J. E., & Ligon, G. S. 2011. First and ten leadership: A historiometric investigation of the CIP leadership model. ***Leadership Quarterly***, 22(1): 70–91.

Janasik, N., Honkela, T., & Bruun, H. 2009. Text mining in qualitative research. ***Organizational Research Methods***, 12(3): 436–460.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. 2018. Text mining in organizational research. ***Organizational Research Methods***, vol. 21. https://doi.org/10.1177/1094428117722619.

Lawson, C. 2020. ***Personal communication***. Rev.com.

Ligon, G. S., Harris, D. J., & Hunter, S. T. 2012. Quantifying leader lives: What historiometric

approaches can tell us. ***Leadership Quarterly***, 23(6): 1104–1133.

Lovelace, J. B., Neely, B. H., Allen, J. B., & Hunter, S. T. 2019. Charismatic, ideological, & pragmatic (CIP) model of leadership: A critical review and agenda for future research. ***Leadership Quarterly***, 30(1): 96–110.

Matthew, J. 2019. ***Package "generalhoslem": The goodness of fit tests for logistic regression models***. https://cran.r-project.org/web/packages/generalhoslem/generalhoslem.pdf.

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. 2018. What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. ***Journal of Management***, 44(7): 2909–2933.

McKenny, A. F., Short, J. C., & Payne, G. T. 2012. Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. ***Organizational Research Methods***, 16(1): 152–184.

Minbashian, A., Bright, J. E. H., & Bird, K. D. 2010. A comparison of artificial neural networks and multiple regression in the context of research on personality and work performance. ***Organizational Research Methods***, 13(3): 540–561.

Mitchell, R. 2018. ***Web scraping with Python***. (A. MacDonald & J. Billing, Eds.) (2nd ed.). Sebastopol, CA: O'Reilly Media, Inc.

Mumford, M. D. 2006. ***Pathways to outstanding leadership: A comparative analysis of charismatic, ideological, and pragmatic leaders***. Mahwah, NJ: Lawrence Erlbaum Associates.

Pandey, S., & Pandey, S. K. 2019. Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. ***Organizational Research Methods***, 22(3): 765–797.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. 2015. ***Linguistic inquiry and word count: LIWC2015 operator's manual***. Austin, TX: Pennebaker Conglomerates. www.LIWC.net.

Pollach, I. 2012. Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. ***Organizational Research Methods***, 15(2): 263–287.

Putka, D. J., Beatty, A. S., & Reeder, M. C. 2018. Modern prediction methods: New perspectives on a common problem. ***Organizational Research Methods***, 21(3): 689–732.

R Core Team. 2020. ***R: A language and environment for statistical computing***. Vienna, Austria: R Foundation for Statistical Computing.

Rev. 2020. ***Terms of service***. https://www.rev.com/about/terms.

Richardson, L. 2020. ***Beautiful soup documentation***.

Ripley, B., & Venables, W. 2020. ***Package "nnet": Feed-forward neural networks and multinomial log-linear models***. http://www.stats.ox.ac.uk/pub/MASS4/.

Ripley, B., Venables, W., Bates, D. M., Hornik, K., Gebhardt, A., et al. 2020. ***Package "MASS": Support functions and datasets for Venables and Ripley's MASS***.

Serban, A., Yammarino, F. J., Sotak, K. L., Banoeng-Yakubo, J., Mushore, A. B. R., et al. 2018. Assassination of political leaders: The role of social conflict. ***Leadership Quarterly***, 29(4): 457–475.

Sergent, K., & Stajkovic, A. D. 2020. Women's leadership is associated with fewer deaths during the COVID-19 crisis: Quantitative and qualitative analyses of United States governors. ***Journal of Applied Psychology***, 105(8): 771–783.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. 2010. Construct validation using computer-aided text analysis (CATA). ***Organizational Research Methods***, 13(2): 320–347.

Starkweather, J., & Moske, A. K. 2011. ***Multinomial logistic regression***. http://www.unt.edu/rss/class/Jon/Benchmarks/ MLR_JDS_Aug2011.pdf.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence unit of information. ***Syst. Res.***, 7: 484–498.

Tukey, J. W. 1977. ***Exploratory data analysis***. Reading, MA: Addison-Wesley Pub. Co.

Weber, M. 1924. ***The theory of social and economic organizations***. New York, NY: Free Press.

Welbers, K., Van Atteveldt, W., & Benoit, K. 2017. Text analysis in R. ***Communication Methods and Measures***, 11(4): 245–265.

Yammarino, F. J., Mumford, M. D., Serban, A., & Shirreffs, K. 2013. Assassination and leadership: Traditional approaches and historiometric methods. ***Leadership Quarterly***, 24(6): 822–841.

Yammarino, F. J., Sotak, K. L., & Serban, A. 2020. Charismatic, ideological, and pragmatic model with shared and collective leadership: A multi-level integration. In S. T. Hunter & J. B. Lovelace (Eds.), ***Multiple pathways to success: Extending the charismatic, ideological, and pragmatic approach to leadership***: 116–142. New York, NY: Routledge (Taylor & Francis Group).

TABLE 1

Summary of Independent Variables, Respective LIWC Dictionary Categories, and Sample Words

Variable Name (case sensitive)	Dictionary Category	Sample Words	Total Words in Category
achieve	Achievement	Win, success, better	213
adj	Common adjectives	Free, happy, long	764
adverb	Common adverbs	Very, really	140
AllPunc	All Punctuation	NA	NA
Analytic	Analytical thinking	NA (summary category)	NA
Apostro	Apostrophes	NA	NA
assent	Assent	Agree, OK, yes	36
body	Body	Cheek, hands, spit	215
cause	Causation	Because effect	135
certain	Certainty	Always, never	113
Comma	Commas	NA	NA
compare	Comparisons	Greater, best, after	317
conj	Conjunctions	And, but, whereas	43
discrep	Discrepancy	Should, would	83
drives	Drives	NA	1103
female	Female references	Girl, her, mom	124
filler	Fillers	I mean, you know	14
focuspresent	Present focus	Today, is, now	424
health	Health	Clinic, flu, pill	294
hear	Hear	Listen, hearing	93
home	Home	Kitchen, landlord	100
i	1st person singular	I, me, mine	24
ingest	Ingestion	Dish, eat, pizza	184
insight	Insight	Think, know	259
interrog	Interrogatives	How, when, what	48
ipron	Impersonal pronouns	It, it's, those	59
leisure	Leisure	Cook, chat, movie	296
male	Male references	Boy, his, dad	116
money	Money	Audit, cash, owe	226
motion	Motion	Arrive, car, go	325
negate	Negations	No, not, never	62
negemo	Negative emotions	Hurt, ugly, nasty	744
nonflu	Non-fluencies	Er, hm, umm	19
number	Numbers	Second, thousand	36

Parenth	Parentheses	NA	NA
Period	Periods	NA	NA
prep	Prepositions	To, with, above	74
pronoun	Total pronouns	I, them, itself	153
QMark	Question marks	NA	NA
Quote	Quotation marks	NA	NA
relativ	Relativity	Area, bend, exit	974
reward	Reward	Take, prize, benefit	120
risk	Risk	Danger, doubt	103
SemiC	Semi Colons	NA	NA
Sixltr	Words > 6 letters	NA	NA
social	Social processes	Mate, talk, they	756
space	Space	Down, in, thin	360
swear	Swear words	Fuck, damn, shit	131
tentat	Tentative	Maybe, perhaps	178
time	Time	End, until, season	310
verb	Common verbs	Eat, come, carry	1000
you	2nd person pronoun	You, your, thou	30

Note*:* Dictionary categories, sample words, and totals were identified in Pennebaker et al. (2015)

TABLE 2

Multinomial Logistic Regression Coefficients, Standard Errors, and Wald Statistics

Variable	I			P		
Name	Coeff.	Std. Err.	*z*	Coeff.	Std. Err.	*z*
Intercept	-39.10	.33	-120.17**	4.13	.79	5.24**
achieve	1.10	.58	1.90	.99	.42	2.37*
adj	-.23	.55	-.42	-.63	.35	-1.80
adverb	2.31	.54	4.26**	1.50	.39	3.87**
AllPunc	-2.41	.74	-3.25**	.43	.23	1.91
Analytic	.01	.08	.19	-.20	.05	-3.68**
Apostro	4.29	.98	4.37**	.03	.50	.07
assent	11.65	4.50	2.59**	3.35	3.54	.95
body	6.22	1.84	3.38**	1.67	1.31	1.28
cause	2.84	.92	3.10**	-.08	.62	-.13
certain	.57	.79	.72	-.70	.55	-1.28
Clout	.34	.12	2.87**	.23	.07	3.23**
cogproc	-1.60	.74	-2.16*	.26	.46	.56
Comma	2.46	.82	2.98**	-.15	.30	-.50
compare	-.99	.80	-1.25	1.02	.51	2.00*
conj	.17	.37	.46	-.97	.24	-4.06**
discrep	1.61	.82	1.97*	.53	.51	1.05
drives	-.52	.29	-1.77	.14	.19	.77
female	-1.11	1.48	-.75	-2.14	1.00	-2.12*
filler	-24.68	.17	-144.51**	13.09	.76	17.11**
focuspresent	-1.03	.39	-2.61**	-.18	.24	-.76
health	-1.61	.83	-1.93	-.70	.48	-1.46
hear	1.62	1.43	1.13	-1.12	1.07	-1.04
home	-2.50	1.48	-1.69	-.18	.83	-.21
i	1.86	.90	2.08*	2.33	.55	4.23**
ingest	-1.21	2.48	-.49	1.14	1.23	.93
insight	1.50	.95	1.57	-.21	.60	-.35
interrog	-1.39	.81	-1.71	-.22	.52	-.41
ipron	.33	.61	.55	.75	.40	1.86
leisure	2.35	1.34	1.75	2.35	.91	-.09
male	-1.13	.70	-1.60	-.37	.36	-1.02
money	.50	.32	1.57	.17	.21	.80
motion	-2.06	1.52	-1.35	-3.31	1.16	-2.85**
negate	3.56	1.07	3.31**	-.18	.72	-.25
negemo	-.43	.38	-1.15	-.56	.24	-2.37*
nonflu	12.12	3.58	3.39**	-5.63	2.41	-2.34*
number	1.10	.28	2.90**	-.48	.28	-1.71
Parenth	-11.80	5.86	-2.02*	-7.82	2.20	-3.56**
Period	1.73	.80	2.16*	.34	.39	.86
prep	.88	.33	2.68**	.85	.25	3.46**

pronoun	-.02	.63	-.03	-.86	.40	-2.15*
QMark	10.70	2.73	3.91**	5.21	2.25	2.31*
Quote	.46	1.38	.33	-1.79	1.38	-2.43*
relativ	2.43	1.48	1.64	3.49	1.15	3.02**
reward	-2.50	.86	-2.91**	-1.70	.54	-3.10**
risk	-1.71	.86	-1.98*	-.70	.53	-1.33
SemiC	3.86	1.44	2.68**	-1.44	.87	-1.65
Sixltr	.37	.15	2.49*	-.11	.09	-1.16
social	-.63	.41	-1.53	-.51	.28	-1.82
space	-2.48	1.44	-1.72	-2.48	1.16	-3.07**
swear	-53.31	.94	-56.96**	-31.93	6.42	-4.97**
tentat	1.30	.87	1.49	-.51	.58	-.88
time	-3.15	1.47	-2.14*	-3.45	1.16	-2.98**
verb	.56	.36	1.57	-.54	.24	-2.23*
you	1.31	.47	-2.79**	-.28	.29	-.94

Note: C was arbitrarily set as the reference for the model. Coeff. = coefficient, Std. Err. = standard error, and z = Wald statistic.

* (P >|z|) < .05

** (P >|z|) < .01

TABLE 3

Fit and Accuracy Summary of Final Multinomial Logistic Regression Model

Iteration	AIC	Training Data Accuracy (%)	Testing Data Accuracy (%)
1	669.45	83.76	78.66
2	640.45	85.79	81.42
3	650.27	84.26	81.82
4	660.11	85.79	80.63
5	636.00	85.28	83.40
6	676.07	84.09	81.82
7	631.41	85.79	79.05
8	662.71	84.94	77.47
9	676.81	85.45	79.45
10	694.60	83.76	80.63
Mean	659.79	84.89	80.44

Note*:* Each iteration includes resampling of the data with a random 70% of the data used for training the model and the remaining 30% used for testing the model. Accuracy refers to the extent to which the model accurately predicted the known outcomes (C, I, or P).

TABLE 4

Example of Probabilities of CIP Factor for Press Briefings

Press Briefing Number	C	I	P
1	1.32%	0.00%	98.67%
2	41.34%	0.16%	58.51%
3	7.39%	2.38%	90.23%
4	25.26%	2.53%	72.21%
5	95.56%	0.05%	4.39%
6	73.26%	4.94%	21.80%

Note: C, I, and P stand for charismatic, ideological, and pragmatic, respectively. Percentages represent the likelihood that a particular observation (i.e., press briefing) falls into each CIP category, based on the application of a trained multinomial logistic regression model.

TABLE 5

Total Governor Press Briefings Classified by CIP Category

Governor Name	C	I	P
Governor Abbott	0	7	18
Governor Baker	0	22	1
Governor Beshear	0	0	12
Governor Brown	0	5	3
Governor Carney	0	0	1
Governor Cooper	0	4	14
Governor Cuomo	0	91	36
Governor DeSantis	0	35	21
Governor DeWine	0	26	35
Governor Ducey	0	2	4
Governor Edwards	0	15	2
Governor Evers	1	0	1
Governor Gordon	0	1	0
Governor Grisham	0	0	1
Governor Herbert	0	0	2
Governor Hogan	0	10	4
Governor Holcomb	0	4	0
Governor Hutchinson	0	4	0
Governor Inslee	0	11	5
Governor Ivey	0	5	1
Governor Justice	0	1	0
Governor Kelly	0	2	2
Governor Kemp	0	6	1
Governor Lamont	2	8	0
Governor Lee	0	5	2
Governor Little	0	1	1
Governor McMaster	0	5	2
Governor Murphy	0	24	22
Governor Newsom	0	73	10
Governor Noem	0	3	0
Governor Northam	0	23	11
Governor Parson	0	0	2
Governor Polis	1	2	3
Governor Pritzker	1	32	8
Governor Raimondo	0	2	1
Governor Reeves	0	2	1
Governor Reynolds	0	13	1
Governor Sisolak	0	2	0
Governor Stitt	0	1	0
Governor Walz	2	3	1
Governor Whitmer	0	13	23
Governor Wolf	0	5	3

Note*:* C, I, and P stand for charismatic, ideological, and pragmatic, respectively. Press briefings were collected from Rev.com. While we gathered all available data from Rev.com, the data source is not an exhaustive collection of all governors' press briefings during this time frame.

FIGURE 1

Boxplots with Tukey Post Hoc of Time Orientation by CIP Style

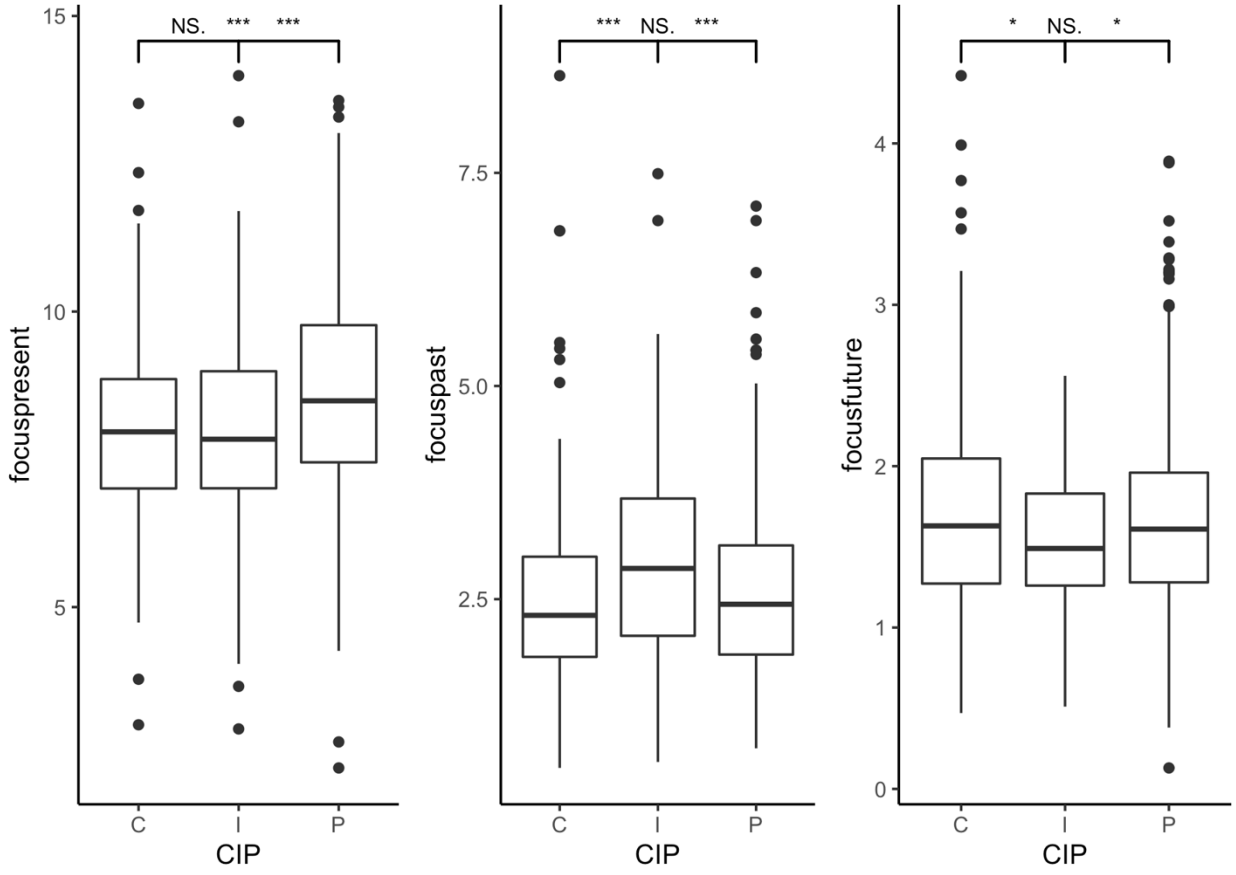

FIGURE 2

Governors' Press Briefings Classified as C, I, or P Over Time

APPENDIX

Box Plots of LIWC Dictionary Categories with Tukey Post Hoc Significance Identified

Note: We used this figure to identify which variables to include in the original version of the model. A significant ANOVA resulted in the variable being included in the model. This method resulted in 68 independent variables used in the model. A summary of the model fit statistics, accuracy statistics, and a summary regression table are presented below.

Alternative Multinomial Logistic Regression Model Equations (Based on ANOVA Selection Criteria)

$$\log\left(\frac{Pr(CIP = I)}{Pr(CIP = C)}\right) = \beta_{10} + \beta_{11}(achieve) + \beta_{12}(adverb) \ldots + \beta_{168}(tentat) + \varepsilon$$

$$\log\left(\frac{Pr(CIP = P)}{Pr(CIP = C)}\right) = \beta_{20} + \beta_{21}(achieve) + \beta_{22}(adverb) \ldots + \beta_{268}(tentat) + \varepsilon$$

Fit and Prediction Summary of Alternate Multinomial Logistic Regression Model

Iteration	AIC	Training Data Accuracy (%)	Testing Data Accuracy (%)
1	757.00	84.43	81.03
2	749.81	82.91	79.05
3	791.57	81.05	77.87
4	796.39	82.40	79.45
5	777.39	83.25	77.08
6	776.07	84.60	80.63
7	756.47	84.09	82.21
8	788.54	82.57	79.45
9	773.44	83.42	77.47
10	741.52	83.93	80.24
Mean	770.82	83.27	79.45

Note*:* Each iteration includes resampling of the data with a random 70% of the data used for training the model and the remaining 30% used for testing the model. Accuracy refers to the extent to which the model accurately predicted the known outcomes (C, I, or P).

Multinomial Logistic Regression Table: Alternate Model

Variable Name	I			P		
	Coeff.	Std. Err.	$\widehat{OR}$	Coeff.	Std. Err.	$\widehat{OR}$
Intercept	-2.34	.54	.01	-11.74	1.00	.00
achieve	-.18	.82	.84	.83	.55	2.29
adverb	2.09	.57	8.07	1.39	.40	4.03
affect	5.48	3.53	2.40e+02	2.14	2.70	8.51
affiliation	-1.13	1.27	.32	-.18	.73	.83
AllPunc	-1.93	.97	.14	-1.50	.66	.22
anger	.86	1.25	2.37	.37	1.50	1.45
anx	1.46	1.50	4.32	.85	.97	2.35
Apostro	4.34	1.21	76.82	1.92	.85	6.84
article	-.51	.46	.60	-.31	.29	.74

assent	12.68	4.88	32.05e+04	4.98	3.42	1.46e+02
Authentic	.05	.09	1.05	.05	.06	1.05
Colon	2.46	2.55	11.66	1.45	1.63	4.26
compare	.30	.87	1.35	.68	.56	1.98
conj	-.29	.41	.75	-.37	.27	.69
death	-1.02	1.53	.36	-1.01	.93	.36
Dash	-.15	1.22	.86	2.01	.74	7.48
Dic	.35	.25	1.42	.13	.17	1.14
drives	-.93	1.08	.39	-.24	.67	.78
family	-1.81	1.76	.16	-2.09	.90	.12
focusfuture	-.49	.57	.61	-.72	.39	.48
focuspast	-.01	.38	.99	-.69	.30	.50
focuspresent	-.89	.40	.41	-.27	.24	.75
health	-6.33	1.62	.00	-4.76	2.18	.00
hear	3.22	1.48	25.09	-1.14	1.01	.32
home	-1.72	1.66	.18	1.34	.93	3.84
i	1.47	.77	4.35	.81	.47	2.27
insight	.04	.74	1.04	-.35	.52	.71
interrog	-2.04	.86	.13	-1.28	.59	.28
ipron	-8.70	.60	.00	1.57	.37	4.82
male	-3.06	1.13	.05	-.53	.73	.59
negate	2.33	1.08	10.24	-.24	.75	.79
negemo	-5.79	3.85	.00	-3.43	2.84	.03
nonflu	7.17	3.19	1.30e+03	-4.07	2.29	.00
number	.82	.43	2.28	-.10	.33	.91
Period	1.83	1.09	6.23	2.39	.71	10.91
posemo	-6.23	3.61	.00	-2.74	2.72	.06
power	.18	1.03	1.20	.33	.62	1.39
ppron	-10.37	.59	.00	.37	.37	1.45
QMark	10.48	3.02	3.55e+04	7.36	2.34	1.57e+03
quant	-2.04	.74	.13	-1.16	.47	.31
relativ	1.33	1.65	3.77	3.45	1.22	31.54
relig	-1.73	1.17	.18	-.07	.66	.93
reward	-1.86	1.12	.16	-.87	.65	.42
risk	-.56	1.23	.57	-.77	.75	.46
SemiC	2.24	1.86	9.42	-.31	1.14	.74
sexual	-16.78	4.80	.00	-10.77	2.81	.00
space	-2.18	1.57	.11	-4.12	1.21	.02
they	2.97	.72	19.57	1.16	.50	3.18
time	-2.82	1.60	.06	-3.80	1.17	.02
Tone	.08	.06	1.08	.01	.03	1.01
we	4.15	.93	63.39	1.17	.47	3.22
work	.45	.36	1.57	-.05	.25	.95
WPS	.01	.10	1.01	-.06	.09	.94
adj	-.29	.60	.74	-.14	.40	.87
bio	5.59	1.40	266.41	4.49	2.15	89.17

body	-2.09	1.96	.12	-3.41	2.38	.03
cause	.02	.68	1.02	.28	.45	1.32
Comma	2.30	1.04	9.96	2.43	.74	11.34
differ	-2.30	.84	.10	-.91	.59	.40
filler	-14.52	.17	.00	20.59	.19	8.71e+08
ingest	-9.71	2.33	.00	-5.21	2.38	.01
motion	-1.26	1.59	.28	-3.41	1.18	.03
prep	.37	.36	1.45	.84	.27	2.32
pronoun	8.72	.56	6.16e+03	-.73	.35	.48
sad	3.04	1.79	20.85	1.37	1.25	3.95
shehe	3.40	1.33	29.88	.27	.91	1.31
tentat	-.88	.71	.41	-.80	.51	.45

Note: C was arbitrarily set as the reference for the model. Coeff. = coefficient, Std. Err. = standard error, and $\widehat{\text{OR}}$ = odds ratio.