DOCUMENTATION FOR: "Getting as good as you give: The effects of the press on politician's post-office earnings"

Joshua Y. Levy [*]

October 8, 2021

---

[*]Joshua Y. Levy (joshua.levy@chicagobooth.edu), Stigler Center, Booth School of Business, University of Chicago

HYPOTHESIS: Corporations have a continued interest in accessing the halls of power to better navigate the law to their own ends. Thus, a so-called "revolving door" effect has emerged whereby government officials begin working for – or at the behest of – private corporations shortly after leaving office to help their new masters/clients navigate, shape, and benefit from rules, regulations, and laws. We suppose however, that a business is willing to retain a former government official or leader, conditional on the leader's competence and/or ideological leanings. Despite often maintaining large government relations departments, we suppose that business leaders are sensitive to the signals of leaders' competence (and pro-business attitudes) issued by elite, business-facing news publications such as *The Economist*. That is, we hypothesize that the more favorably such publications write about a leader, the more they are likely to earn as consultants to, or employees of private corporations after leaving office.

Enclosed is documentation about the methodology and some preliminary results working to verify this hypothesis

1   00 - The Economist Scraper

In order to test this hypothesis, I begin by collecting the corpus of *The Economist*. I use a Python script to scrape all articles stored by ProQuest: ABI/Inform.[1] Comments are included on each script but below is a high-level overview of this step's organization. The bulleted list is in order that each script should be run. The key output file of this folder is "formatted_compiled_articles.csv" which is a three column .csv file with the headers: "date," "link," and "text" containing the date of the issue the article was published in, the link to the ABI/Inform page from which the text was scraped, and the body of the text, respectively. This script collects on the order of 123,000 articles between January 1992 and December 2019.

- **ABI-INFORM Economist Scraper.py**
  This Python script that does the bulk of the article collection. It is broken into two parts: (1) visiting "issue-level" pages to collect links to every article; and (2) visiting every "article-level" page (links to which were collected in part (1)) to collect the body of each article's text. There are a number of parameters that need to be set by the user to run this script correctly describe in the commented header.

  Note that this scraper was written to access a database that sits behind shibboleth-style dual-factor authentication (specifically the University of Chicago's). As a result, it requires some amount of supervision to run properly. If just part (1) is run then the only supervision required is at the time of initiating the scraper as the user must create an authenticated session which a Selenium driver will use to collect links to articles. (Note that Selenium is used in part (1) because the issue-level pages interact with the server using AJAX so the page is dynamic and the scraper must wait for certain information to load into the page before parsing the HTML.)

  If just part (2) is run then the user is required to supervise somewhat more closely. As before, the user will have to authenticate their session to generate a set of authenticated cookies

---

[1]**N.B.** Apparently this violates UChicago library policy. Please continue reading for information on collection.

that will be used by the article-level scraper. This secondary article-scraper runs using the requests library in Python because it is faster than simulating the entire browser experience (as Selenium does). However, dynamically updating the authenticated cookies so that the authenticated session does not time out becomes prohibitively difficult. Instead, a single authenticated cookie is passed to each request which generates a new session. By my experimentation, it is thus best to run the scraper 10,000 requests (i.e 10 "batches" of a thousand requests each) at a time. This maximizes scraping for minimal user re-authentication. Part (2) of the scraper stores links-to and article-text in .csv files of 1,000 rows and saves them to the "sub_csvs" folder as .csv files. Additionally, it generates and updates the "tracker_sheet.csv" file so that successive runs do not generate duplicate article-observations.

N.B. I do NOT recommend running this script again. Bulk collection is in violation of University Library policy and I (Joshua) have already been warned once. We currently have around 1,500 observations that failed and I will collect by hand.

- **csv_compiler.py**
  This Python script reads in each of the .csv files generated by part (2) of the above scraper cleans them. Every quarter *The Economist* produces a "special report" consisting of several articles. However, instead of creating separate article-pages for each of these, ABI/Inform collates all of them and saves them on a single page. This results in a very long string that breaks the formatting of the above scraper. This script reformats all of the sub_csv files and compiles them into a single .csv file that can be used for subsequent analysis. The output of this script is called "formatted_compiled_articles.csv"

2 01 - Leader Selection

In order to better understand the role that the (business-facing) press has in mediating the sums of money that business pay former leaders for advice or access, we additional narrow the scope of nations under consideration to substantive democracies. I do so for two reasons: first, substantive democracies tend to be more transparent – suggesting that the press has a bigger role in those nations relative to more autocratic ones. Indeed, in many settings, freedom of the press is considered one of the necessary constituent parts of substantive democracy. Additionally, when considering former leaders of autocratic nations, businesses may be inclined to access or retain former leaders not because of the supposed competence of the leader in question as conveyed by the press, but because of the perception that doing so is either necessary or advantageous for practicing some kind of (illicit) corruption. This thus circumvents the mechanism of interest.

What follows is a description of how nations were selected and nations identified.

- **LeaderSelection.R**
  This R script identifies nations of interest using the Polity V dataset. Polity constructs a "score" at the nation-year level ranging from -10 (most autocratic) to 10 (most democratic). To identify as large a sample of substantive democracies as possible, we identify nation-years that receive a Polity score greater than or equal to 8.[2] In order to keep the sample of nations

---

[2]Note that this level was selected arbitrarily and can be amended.

as large as possible, if a nation receives a Polity score of 7 at any point in the sample, but maintains a score of at least 8 for at least 10 years, the country is retained in the sample. This dataset of sample-years is then matched with the REIGN dataset of leader observations.

I additionally use this dataset of leader-years to output two .csv files. The first, "leaders_names_pre_styling.csv" is a list of leaders and nations with their REIGN stylings.[3] This however, is not necessarily how *The Economist* styles the leaders names. I modify this list, identifying *The Economist's* editorial style saving it "leaders_names_econ_styling.csv". A similar process is conducted for "countries_pre_details.csv". This .csv simply contains a list of countries of interest. I then add nation-level information such as the leader's title (e.g. prime minister, president, chancellor, etc.) national denonyms (e.g. Chileans, Japanese, etc.) and other information. This is stored in "national_titles_adjectives.csv". These files are important for cleanly identifying leaders within the body of text of each article.

	country	polity	year
	Length:1952	Min. : 7.00	Min. :1992
	Class :character	1st Qu.: 8.00	1st Qu.:2001
Table 1: Summary Statistics	Mode :character	Median : 9.00	Median :2007
		Mean : 9.19	Mean :2007
		3rd Qu.:10.00	3rd Qu.:2013
		Max. :10.00	Max. :2019

3   02 - Coreference Resolution

Coreference resolution is a natural language process (NLP) technique by which ambiguous noun-phrases or "entities" within a body of text are identified and "resolved" to be unambiguous. For instance, the following text remains semantically ambiguous to a computer. "The quick brown box jumped over the lazy dog. It flew through the air." The pronoun, "it," at the beginning of the sentence clearly refers to the "fox" in the first sentence, but a computer is regularly unable to discriminate between this potential relationship and a potential relationship between "it" and the "dog"

To mitigate this problem I use a combination of NLP tools: Allen NLP's pre-trained SpanBERT Coreference resolution model and spacy, a general purpose NLP tool (Lee, He, & Zettlemoyer, 2018). Trained on hundreds of thousands of sentences that have been marked for their semantic and lexicographic features such as part of speech, position relative to punctuation, subject-object relationship, number consistency, etc. SpanBERT can resolve the above example to: "The quick brown fox jumped over the lazy dog. The fox flew through the air."

This process is important because SentimentR (described in greater detail below) assesses sentiment at the sentence-level. Thus, it is important for us to maximize the sample of sentences that semantically refer to our entities of interest: former leaders. For example on January 23rd, 1991, *The Economist* published an article that began as follows:

---

[3]Note that because the REIGN dataset does not include leaders from Taiwan or Montenegro, these are added manually.

"The one thing his aides had told him to do was to keep it brief and for once he listened. On January 20th, moments after being sworn in as America's 42nd president, Bill Clinton gave an inaugural address that clocked in at under 15 minutes."

Naively passing these two sentences for sentiment analysis would yield only one observation of a Clinton-sentence. However, using coreference resolution, I resolve this text to:

"The one thing Bill Clinton's aides had told Bill Clinton to do was to keep it brief, and for once Bill Clinton listened. On January 20th, moments after being sworn in as America's 42nd president, Bill Clinton gave an inaugural address that clocked in at under 15 minutes."

Because the corpus of *The Economist* contains many more non-leader sentences than it does leader sentences, maximizing the number of semantic signals for sentiment analysis is of the utmost importance.

- **coref_resolver.py**
  This Python script completes several tasks. First it creates an index file, "date_index.csv" so that it is not necessary to keep all 123,000+ articles in memory at any given time. This script additionally reads in leader-level information from "leader_meta_matched.csv" so that it is possible to create a subset of potential articles to search on. It then reads in all articles published six months before, during, and six months after the term of the leader of interest. The script then tags the articles which contain a mention of a leader so that coreferences can be resolved. This script generates a series of .csv files as outputs, one for each leader and stores them in the "leader_tagged_resolved" folder with the format "[LEADER-LASTNAME]_[COUNTRYCODE]_resolved.csv" **NOTE:** Coreference is a computationally expensive task. DO NOT RUN THIS FILE WITHOUT THINKING ABOUT SAMPLE SIZES. Each such .csv file contains article meta-data, the original article text, and the resolved text.

4 03 - Sentiment Analysis

To test our hypothesis, it is necessary to have a systematic way by which to evaluate the way *The Economist* writes about (various) leaders, and indeed if there exists considerable hetereogeneity in its editorial position on each leader's performance. That is, we need a way to consistently measure if *The Economists* writes more favorably of some leaders than it does others.

The tool that we use, an R package called SentimentR, attempts to do this by using an "augmented dictionary approach." (Rinker, 2019) The most simple way to measure sentiment is to first construct a dictionary of "positive" and "negative" words, and then simply tally up the number of times such words appear in a body of text. One could additionally, weight some words more heavily than others — "exemplary," for example might receive a stronger positive weighting than "fair."

SentimentR introduces another layer of complexity by identifying "valence shifters," that may amplify, mitigate, or negate the sentiment of a cluster of positive or negative words. After identifying "polarity clusters," the SenitmentR algorithm alters a sentence's score based on the presence or

absence of these shifters. Words like "however," "but," and "although," for example, are adversative conjunctions that can amplify or de-amplify a cluster's polarity based on their position in the sentence relative to the cluster.

SentimentR conducts analysis at the sentence-level. Thus, as noted above, maximizing the number of sentences that mention a leader is important because an article's latent information about a leader's competence can easily be identified by a human reader but not by a computer. Once articles that mention a leader of interest is identified, it is split on its sentence boundaries, and then each sentence is identified as either a "leader-sentence" — that is, the sentence itself mentions the leader — or an "ambient-sentence" which does not. A polarity score is then generated for each sentence. We segregate leader-sentences from ambient-sentences because the sentiment of an article at large may be contrary to the sentiment expressed about the leader of interest. Consider, for example, a hypothetical article that has a negative sentiment due to an ongoing political or economic crisis. Despite the dour mood expressed by such an article, *The Economist* can still be laudatory of the leader managing the crisis. We would be interested in the positive sentiment expressed about the leader despite the circumstances he or she may face. We additionally aggregate these sentence-level sentiment scores to the "entity-level" — an average of all leader-sentence or ambient-sentence scores in each article. [4]

To test the validity of the SentimentR results, we also employ VADER, a similar sentiment analysis packaged developed for Python. In much the same way as its R counterpart, VADER uses valence shifters and a dictionary of polarity scores to gauge the sentiment expressed by a given sentence.

4.1 Preliminary Sentiment Results: Figures

Below is an example of SentimentR output based on 233 articles of text for Jose Maria Aznar, former prime minister of Spain. The imposed lines are lowess-smoothed lines of entity-level sentiment at each date of observation (issue publication date). Each point in the scatter plot is a sentence level observation. SentimentR normalizes all composite scores to an interval that ranges from -1, the most negative, to 1 the most positive.[5] This process is replicated for every leader of Spain in our time series (from 1992 to 2020) and the results are concatenated as presented in Figure 2.[6]

As detailed above, we segregate leader-sentences from ambient-sentences in order to separate evaluations of the leader of interest from the ambient sentiment that may be more reflective of a crisis, a particular event, or conditions outside the control of the leader. In Figure 3, we plot *The Economist's* sentiment about a leader against ambient sentiment expressed in the same article. The 45-degree line in black indicates perfect correlation between ambient- and leader-sentiment.[7] This exercise serves as a heuristic test of whether segregating leader sentiment from ambient sentiment is worthwhile. That is, if leader and ambient sentiment are highly correlated, the segregation step

---

[4] When aggregating sentence-level sentiment to the entity-level, SentimentR uses a weighted average that down-weights the scores of neutral sentences (on the assumption that users are interested in identifying the positive-/negative- leanings of the document at large. We replicate this weighted average when aggregating VADER scores detailed below.)

[5] Note that observations for a single sentence can exceed this range due to the effect of valence shifters. We subsequently refer to this type of figure as a "leader figure"

[6] We subsequently refer to this type of figures as a "timeline figure."

[7] We subsequently refer to this type of figure as a "entity correlation plot." In some instances many leaders are plotted "together" or are "faceted."
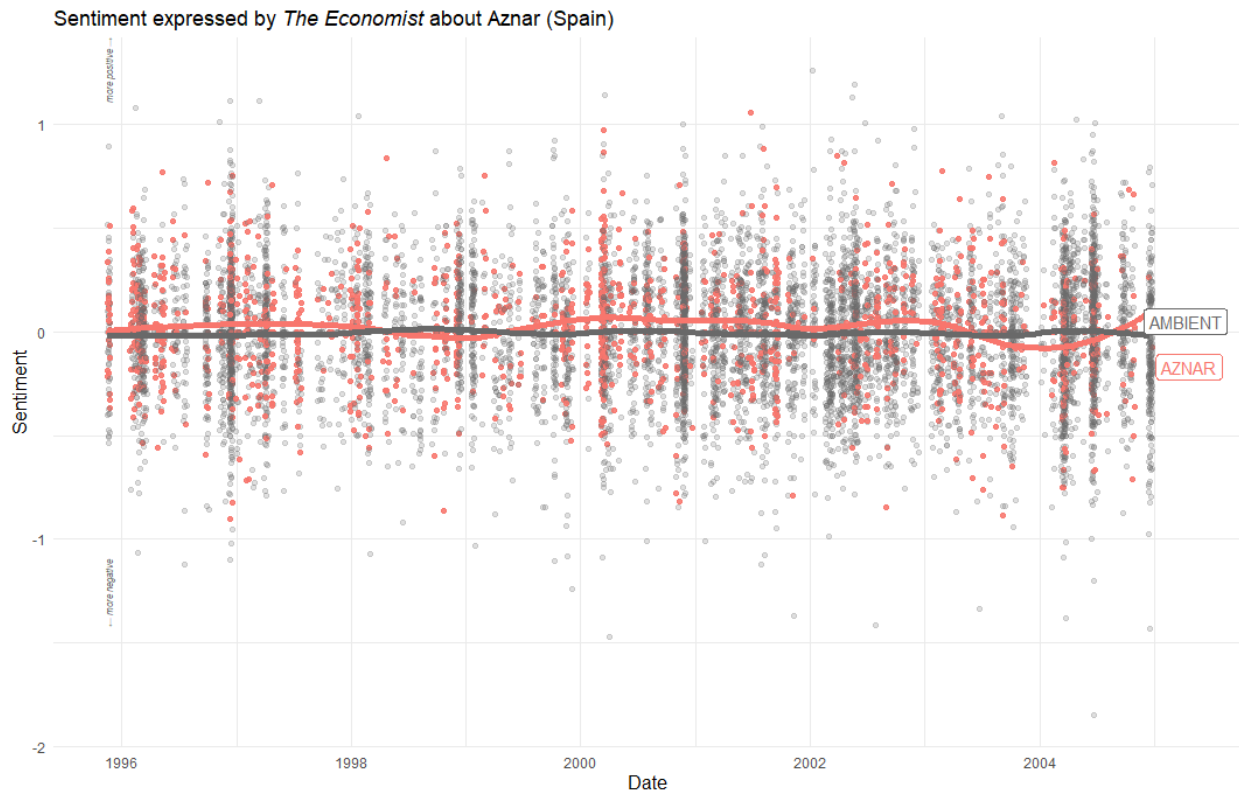
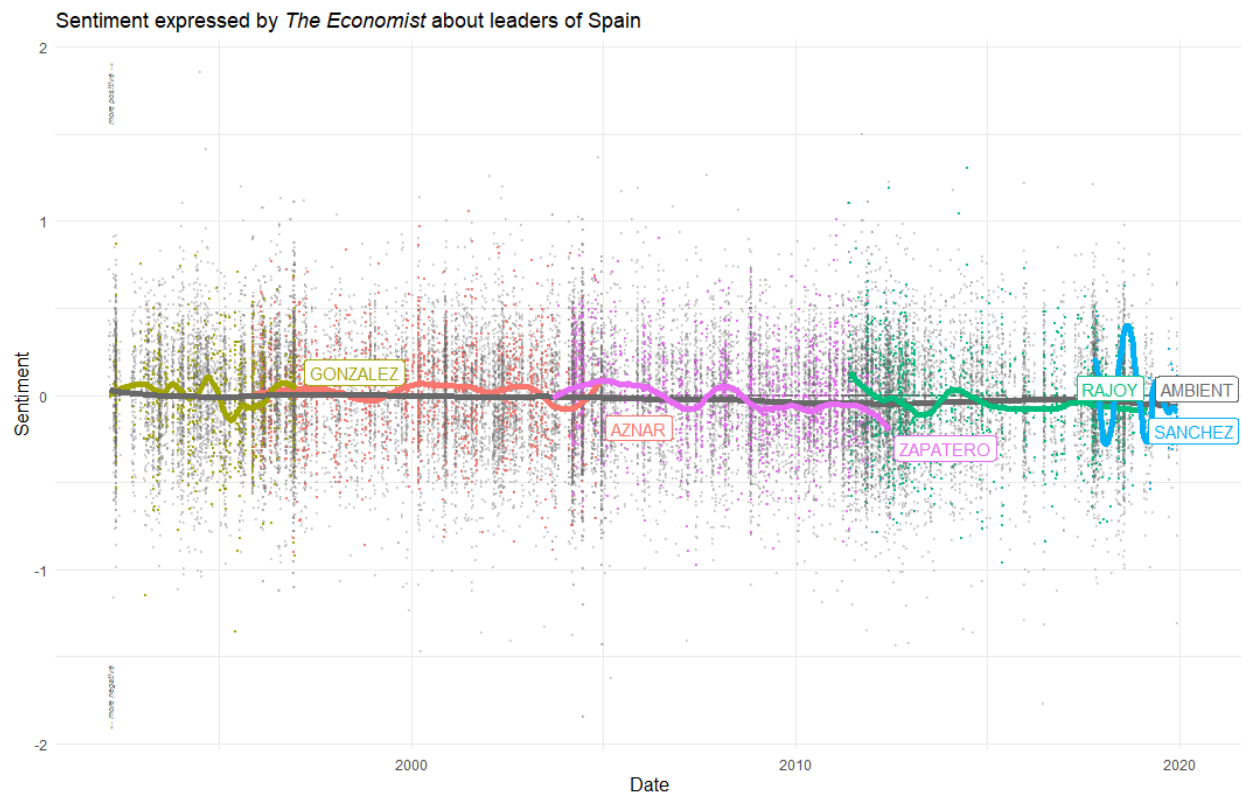Figure 1: Sentiment expressed by *The Economist* about Aznar (Spain)

Figure 2: Sentiment expressed by *The Economist* about Spanish leaders
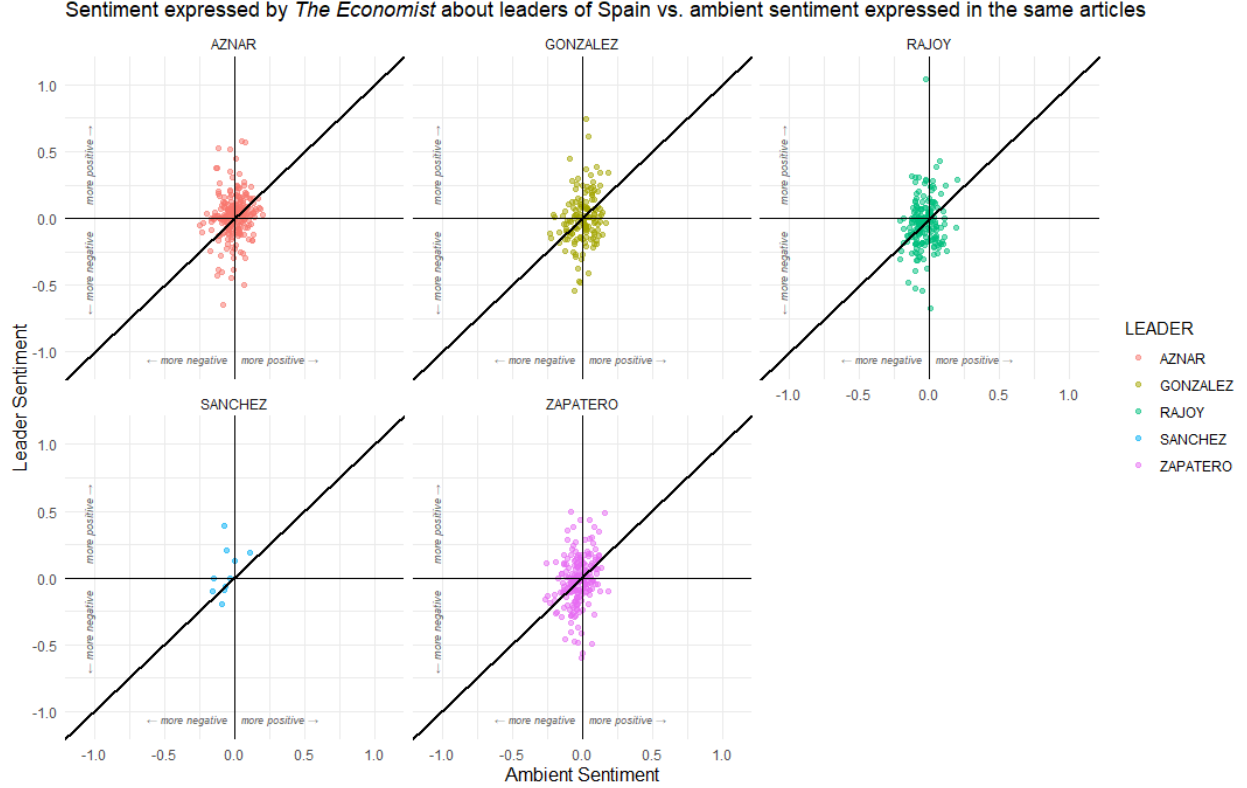
Figure 3: Sentiment expressed by *The Economist* vs. ambient sentiment expressed in the same articles

could be bypassed and sentiment could simply be evaluated at the article level (perhaps even without coreference resolution.) Note in Figure 3 that there is no clear corelation between leader- and ambient-sentiment, suggesting that the segregation process is an important step in identifying *The Economist's* evaluation of a given leader.

We also conduct a similar heuristic test of whether VADER and SentimentR agree on the sentiment expressed in a given article. In Figure 4 below, we plot the R^2 value of a simple OLS regression of SentimentR score on VADER score.[8] Note that despite its simplicity (the only regressor is VADER score), most models exhibit a very high goodness-of-fit. An interactive version of this heatmap can be accessed online at this link. Plots of individual leader-level correlations can be found in this files repo.[9] Again, note that VADER and SentimentR scores are tightly correlated.

4.2 Preliminary Sentiment Results: Discussion

There are three important results worth noting. Firstly, VADER and SentimentR both return results that suggest that the ambient sentiment expressed by *The Economist* is very neutral (scores

---

[8]We subsequently refer to this type of figure as a "sentiment score heatmap." The OLS model computed can be expressed by SentimentR $= \alpha + \beta$VADER $+ \epsilon$.

[9]Navigate to 00_README_DOCUMENTATION/figures/plotly_VADER_SENTIMENTR_corel_facet.png
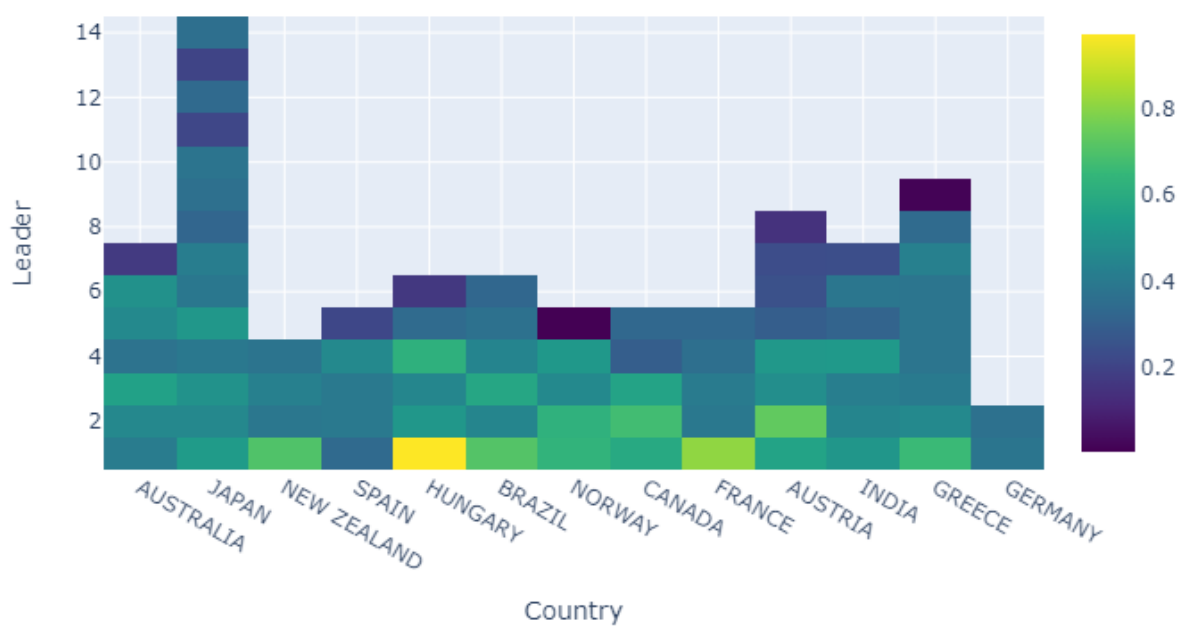
Figure 4: Goodness-of-fit of OLS Model fitting SentimenR scores to VADER scores

| Country | CCODE | Leader figures | Timeline Figure | Entity Correlation | | Present in Heatmap |
				Faceted	Together	
Australia	900		✓	✓	✓	✓
Austria	305		✓	✓	✓	✓
Brazil	140					✓
Canada	20		✓	✓	✓	✓
France	220		✓	✓	✓	✓
Germany	255					✓
Greece	350		✓	✓	✓	✓
Hungary	310					✓
India	750					✓
Japan	740					✓
New Zealand	920					✓
Norway	385					✓
Spain	230		✓	✓	✓	✓

Table 1: Figures that have been completed and uploaded to GitHub

cluster near 0). That is, over time, articles published by *The Economist* that mention leaders of interest have neither a disaster-bias (which would likely carry negative ambient sentiment) nor a blessing-bias (which would likely carry positive ambient sentiment).

More importantly however, observations leader sentiment appears to exhibit a similar property — that is, over time, leader-sentiment also appears to be very neutral. This may be a credit to the apparent impartiality of the *The Economist's* editorial staff. However, it also leaves us with the problem of having little heterogeneity to exploit in future analysis. Some potential solutions are presented below.

Finally, recall that SentimentR and VADER scores are very tightly correlated. This is encouraging in that both packages are identifying similar sentiment patterns in the articles of interest. Notably, however, VADER scores exhibit much higher variance than do SentimentR scores. That is, VADER assigns higher scores to positive sentences and lower scores to negative sentences than does SentimentR. This property appears to be symmetric in that it amplifies positive scores about as much as it amplifies negative scores. Thus, the leader-neutrality problem described above cannot simply be resolved by switching to use of VADER as, over time, average leader-sentiment would remain neutral.

4.3 Countries

Below is a list of countries for which figures are available. Any unavailable figures can be produced on-demand with the `leader_sentiment_general.R` script. This figures can be accessed at the following by following this link

4.4 Sentiment Analysis: Next Steps

Using relatively naive sentiment analysis tools like SentimentR or VADER have some advantages in ease of use but are relatively rudimentary and used more often for sentiment classification (i.e. using scores as thresholds between discrete "negative," "neutral," and "positive" categories) of explicitly editorial documents (e.g. movie or restaurant reviews) or smaller documents (Tweets, for example). This problem is compounded twice over: once by the way the deeply idiomatic style in which *The Economist* writes; and twice by the type of sentiment we are trying to identify. In reality we are trying to identify the sentiment of an article's *evaluation* of a leader.

This additional layer of abstraction requires a more sophisticated method for identifying leader sentiment. I propose using topic modelling techniques detailed here:

- BERT for Text Classification with No Model Trainging

- Text Analyses & Feature Engineering with NLP

5 04 - Officer Pay

There is large variation in the official pay for world leaders. In addition to receiving a public salary, leaders may additionally be entitled to "fringe benefits" such as accommodation (in a Presidential Palace or Residence, for example) free travel when on official business, and other forms of non-pecuniary remuneration. Because it is hard to calculate the true value of these benefits, I limit definitions of compensation to salaried pay.

Below is a sample of OECD nation's leaders and their pay. (Note that this is an incomplete sample because of disagreements between whether the "head of government" or "head of state" is the true chief executive of a country. Take, for example, the conflict of such status in France where the head of state, President Emmanuel Macron, is the chief executive. His counterpart in the UK however, is prime minister Boris Johnson, the head of government.)
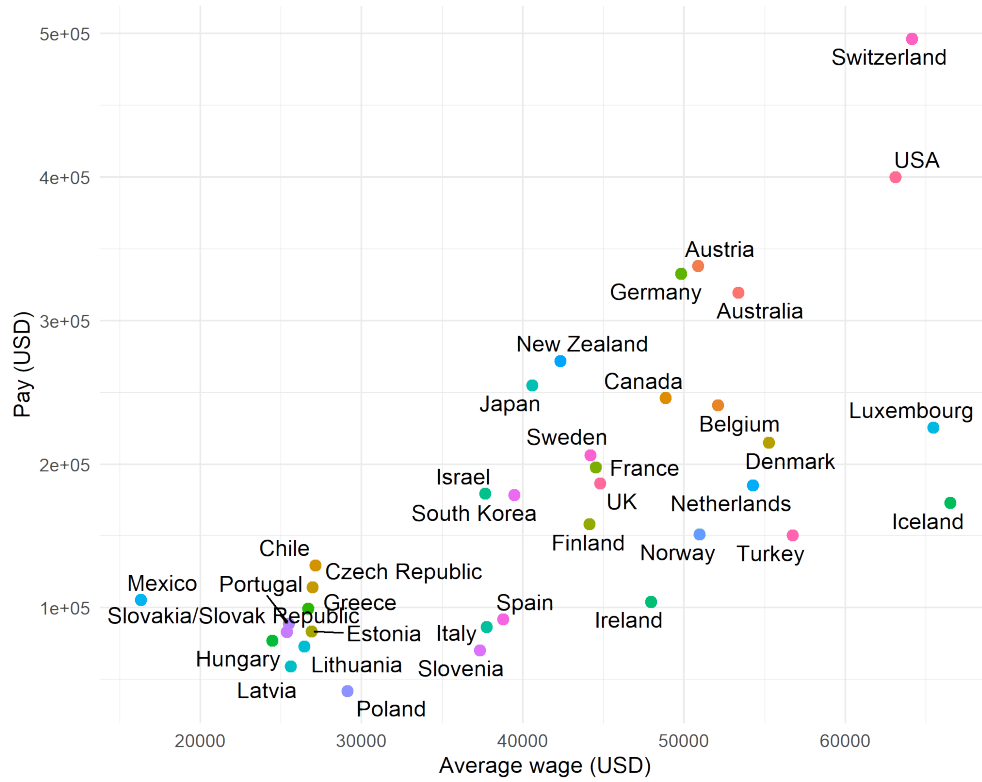
References

Bell, C. (2016). The rulers, elections, and irregular governance dataset (reign). *Broomfield, CO: OEF Research. Available at oefresearch.org*.

Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Naacl-hlt*.

Rinker, T. W. (2019). sentimentr: Calculate text polarity sentiment [Computer software manual]. Buffalo, New York. Retrieved from `http://github.com/trinker/sentimentr` (version 2.7.1)