Learning from Machines: Differentiating US Presidential Campaigns with Attribution and Annotation

Musashi Jacobs-Harukawa

Draft as of May 18, 2022

Abstract

Identifying the differing ways in which political actors and groups express themselves is a key task in the study of legislatures, campaigning and communication. A variety of computational tools exist to help find and describe these patterns, typically summarizing differences with weighted word lists representing either lexical frequencies or semantic fields. I identify two limits to the inferences that can be made based on this method: the ambiguity of the semantic value of words without wider context and an inability to detect differences outside of lexical semantics. I present a combination of text annotation and deep-learning feature attribution, a set of techniques for evaluating the relative importance of data inputs to the prediction of a neural network classifier, as an alternative means of identifying differentiating language usage in political texts. Results are evaluated with comparison to existing text-as-data tools on a dataset of US presidential campaign advertisements from Facebook between 2017 and 2020.

Word Count: 13,132 (including title, abstract, article, figures and references)

Introduction

How can we summarize differences in rhetoric between Donald Trump and Joe Biden? Observers familiar with both politicians might contrast the policy issues they choose to talk about, such as Trump favoring illegal immigration and Biden favoring infrastructure development. One can also point to differences in the sentiment they try to evoke, with Biden's rhetoric emphasizing unity, especially among Democrats, versus Trump emphasizing the existence of a movement and its enemies; or, specific phrases or slogans associated with the respective candidates "Build Back Better" versus "Make America Great Again". One can simply point to the fact that their voices sound very different. These descriptive differences in rhetoric matter for characterizing and understanding political actors, their context, and are core to a common notion of what constitutes *politics*.

Political scientists have applied theories and frameworks from linguistics to help systematize these observations. The examples mentioned in the previous paragraph can described as differences in semantics (the meaning of words and things referred to), lexis (word choice), and prosody (how things sound). In recent years, automated tools from computational linguistics have been applied with great success to political science questions seeking to characterize differences in language usage between political actors in large corpora, such as the study of legislator attention (i.e. what do congresspeople speak about? Grimmer 2013), policy framing in news media (Barnes and Hicks 2018), or identifying the words that characterize Democrats and Republican lawmakers (Monroe, Colaresi, and Quinn 2008).

I argue that while existing computational tools used by political scientists are effective for identifying some kinds of linguistic differences, they are limited in their ability to identify and describe others. I attribute this to two design decisions. First, current methods typically use words as a unit of analysis, discarding linking information relating to order, context and syntax. Second, current methods primarily output reductive numerical summaries, which are then used to weight word lists or documents. This approach links theory and empirics by inferring the semantic values and contexts associated with individual words, and then making claims about the prevalence of these phenomena based on numerical weights. This results in two widespread issues: difficulty validating the inferred semantic value of a word because of a disconnect between the source texts and the summaries, and difficulty identifying or describing patterns that occur outside of lexical semantics.

To address this gap, I offer a novel computationally-assisted approach for identifying and describing differences in language usage between political actors. In this approach, texts are annotated with salience scores representing whether a region contains language that differentiates the speaker. These salience scores are generated using Integrated Gradients (Sundararajan, Taly, and Yan 2017), a technique for measuring the importance of individual data inputs to the prediction of neural network model, extracted from a deep learning (DL) classifier based on the `BERT` architecture (Devlin et al. 2019). Instead of using these scores as a scoring method to rank the extent to which words are differentiating, I annotate the source texts with these scores in order to visualize the logic of the DL classifier in context and detect a broader range of linguistic phenomena.

I demonstrate the reliability and utility of this method with three sets campaign comparisons. First I compare the highly contrastive Trump and Biden campaigns to show that the method produces sensible and plausible results. Next, I compare the more similar Sanders and Warren campaigns to show that the method is capable of detecting subtler and unexpected differences. Finally, I compare eleven Democratic primary candidate campaigns, including Biden, Sanders and Warren, to show both the substantive importance of reference category when identifying characterizing features, and that the method can be generalized to the multiclass case.

Using the feature attribution scores as a way to identify differentiating words, I discover diverging lexical patterns such as Sanders using the words “campaign” and “donate” versus Warren using the phrases “grassroots movement” and “chip in”. While having a relatively low rank correlation compared to established word ranking methods, there is a sufficiently similar semantic interpretation of high-ranking tokens to believe that a feature attribution strategy does find substantively plausible patterns. Furthermore, using feature attribution scores to create annotative heat maps over source texts, I identify thematic patterns in individual advertisements.

This article makes two broader theoretical contributions to social science methodology. First, I demonstrate that instead of using reductive summaries, annotating documents with patterns of attribution permits valid inference and bridges qualitative and quantitative approaches to text, combining the reliability of human validation with the scalability of algorithmic analysis. Second, this paper presents a novel application of DL within a social science research framework. Despite its increasing importance in natural language processing

(NLP), machine learning (ML) and artificial intelligence (AI), DL has seen limited application in social science research because of the low explainability of DL models. Using model interpretation and validation approaches developed specifically for DL settings, I show how we can incorporate complex models into a measurement and estimation strategy.

The rest of the article is organized in four parts. In the first section, I motivate the application of computational tools to analysing political language, and highlight the limitations of existing approaches. In the second section, I make arguments for phrase-level and annotative approaches and present feature attribution on sequence classifiers as strategy that can be validated. I then present the results of this novel approach applied to political campaign advertisements, first contrasting its application to existing lexical scoring approaches, then next demonstrating a novel annotative approach. In the final section, I discuss the possibilities and limitations of my proposed approach.

Motivation

Prior to presenting the approach, I elaborate the need for a novel method by considering the aims of automated text analysis in political science and the limits of existing tools. I argue that the utility of automated tools for text extends beyond problems of scale, and that these tools should be thought of as more than a means of *amplifying* human intuition. I then consider advantages of the reductive numerical approach to text used in political science text-as-data, but also establish the kinds of claims that we are unable to make within this paradigm.

Why Text, and why Automate?

In their 2013 survey of the burgeoning political science text-as-data field, Grimmer and Stewart (2013) note the increased availability of large political corpora in electronic form as a driving factor behind the increased popularity of automated content analysis methods. This "big data" trend is the case for political campaigns as well. An indirect consequence of the major growth in internet-based political campaigning (Fowler et al. 2021) has been the creation of datasets several orders of magnitude larger than those previously available;

four such examples are detailed in the table below.[1]

Dataset	N Docs
Google Transparency (Politics, US)	619K
Facebook Ad Library (Politics, US)	13.3M
Princeton Corpus of Emails[2]	435K
ProPublica	225K

The logistical challenges presented by these large corpora makes clear that one motivation for the development and application of automated content analysis methods is scalability. Automated tools make it possible for researchers with limited funding support to analyze large quantities of unstructured data. However, justifying automated content analysis methods in terms of feasibility and efficiency presents them as a second-rate option for researchers unable to hire and train large numbers of human coders or research assistants, which misses several important parts of the picture.

For one, political scientists study texts because many political phenomena are textual. In some cases, this is in the form of incidental trace data, such as the transcripts of political activities (e.g. Diermeier et al. 2012; Gentzkow, Shapiro, and Taddy 2019). Other political phenomena are inherently of text form, such as laws (Lapinski 2008) and court decisions (Clark and Lauderdale 2012), or the communications that construct a political environment such as legislator press releases (Grimmer 2013) and news articles (Barnes and Hicks 2018). Studying these phenomena requires a range of tools specifically designed to deal with the complex way information is encoded into natural language.

While structured qualitative approaches to text can use human natural language understanding to trivialize the complexity issue at the phrase and document level, their challenge is aggregation and the identification or description of corpus-level patterns. One common approach, referred to as (Qualitative) Content Analysis (QCA), systematizes the human estimation of counts and proportions of concepts of interest within a corpus. This strategy can be applied at various levels. Examples of document level coding include Barabas and Jerit

[1] Note that such large numbers can be misleading in both directions; on the one hand they are underestimates of the total number of all political advertisements shown on these platforms (Edelson, Lauinger, and McCoy 2020) but on the other hand, most research questions are typically interested in a considerably smaller subset of advertisements, such as those run by candidate campaigns or PACs.

[2] See Mathur et al. (2020).

(2009) and Hayes and Lawless (2015), who respectively code news articles by event or tone. Classification can also be done at the quasi-sentence level (e.g. John and Jennings 2010), or flexibly between levels as an information retrieval task for finding all texts and extracts relevant to a particular issue (Stanig 2015).

What do we Estimate?

Grimmer and Stewart (2013) name quantitative analogs for tasks done with QCA, and where the estimand is unified between qualitative and quantitative approaches text-as-data methods can be thought of as tools for "*amplifying* and *augmenting* careful reading" (Grimmer and Stewart 2013, 268). But as with non-text quantitative methods used in political science, quantitative text analysis tools popular in political science applications primarily produce numerical/statistical summaries. Three categories of approach have been particularly popular in political science applications: scaling models, topic models and embeddings. Scaling models such as Word Scores (Laver, Benoit, and Garry 2003; Lowe 2008), Wordfish (Slapin and Proksch 2008) or Wordshoal (Lauderdale and Herzog 2016) score documents onto a single dimension, which is then used to compare their positions. Topic models (Blei, Ng, and Jordan 2003) decompose a corpus into k distributions over words and documents, which are generally interpreted to correspond with separate semantic fields. These topics are then used to describe documents as a mixture of these fields, or to compare different documents within the same topic. Popular variants include seeded topic models (Eshima, Imai, and Sasaki 2020), which can be used for semi-supervised lexicon discovery and parameterised topic models (Roberts et al. 2014), which are used to model topic mixture as a function of speaker covariates. Finally, embedding models represent tokens in a corpus as the dense representation of the conditional likelihood of a token given its context and other metadata in the training corpus. These vector representations are used to model words, documents and document covariates; Rodman (2019) describes the semantic shifts of key terms in newspaper corpora, whereas Rheault and Cochrane (2020) include speaker metadata to produce "party embeddings".

These approaches are all reductive in the sense that they produce a lower-dimensional summary of the source texts. They are also appropriate to their respective applications. If our estimand is ideology and we treat parliamentary speech (e.g. Peterson and Spirling 2018) or manifesto text (e.g. Slapin and Proksch 2008) as a noisy signal of ideology, then the appropriate estimator is one that decomposes or otherwise filters the input.

Likewise, if our goal is to describe the semantic shift of key terms (Garg et al. 2018; Rodman 2019) then our estimator is distances in a word embedding space, and if our goal is to discover and label the semantic fields contained in a corpus (Blei, Ng, and Jordan 2003; Grimmer 2013; Roberts et al. 2014), then our estimator will be a topic model. In the case of this article, where the objective is to select the linguistic features that differentiate and characterize political actors, past approaches have been based on lexical scoring methods, such as Monroe, Colaresi, and Quinn (2008) who compare four different approaches for identifying words that differentiate Democrats from Republicans in a congressional speech corpus.

The theoretical estimand in all of the above papers is semantic in nature, such as the ideology of a text (Laver, Benoit, and Garry 2003; Lowe 2008; Slapin and Proksch 2008; Lauderdale and Herzog 2016) or the words most closely related in *meaning* to a target word or list of words (Rodman 2019; Eshima, Imai, and Sasaki 2020). The link between theoretical and empirical estimands in these studies is made by leveraging the fact that words tend to be associated with meanings. Thus researchers combine the semantic values of words with numerical values provided by models to produce weighted or ranked word lists, and then infer the relative prevalence of semantic quantities in the data.

Many of the challenges and limitations of these methods–e.g. polysemy, ambiguity–result directly from an approach that begins with discarding word order and ends with describing the meaning of language in terms of the relative rates of usage of different words or phrases. These challenges are exacerbated by the fact that summary and reduction necessarily entail informational loss. Particularly when the representation of the results is in a different medium to the original data (i.e text to numeric), tasks such as sense disambiguation become difficult because of the disconnect between the summarized quantity and the original text.

It is not my intention to claim that inference leveraging numerical summaries for aggregation and lexical semantics for linking theory and empirics is useless or invalid; on the contrary, the careful application of these methods has produced many valid works in the study of important textual political data and phenomena. Rather, my point is that it is not always necessary to disconnect the source texts and numerical summaries, and opting not to do so has various benefits. Some questions require the contrasting of corpus-level patterns with document-level phenomena. In this article, the goal is to identify the characteristics of an advertisement that differentiate it as belonging to its campaign. This compares corpus- and document-level features. Identifying the linguistic features that typify a political advertising campaign requires knowledge of the entire corpus, and identifying which features of a document make it (a)typical of that campaign requires application of those corpus-level patterns to every aspect of the document. In these cases, I argue that there is a need for annotative methods that link high-level summaries to individual observations. I elaborate on such a tool subsequently.

Differentiating Language

In this section I present a new approach to differentiating the language use of political actors, based on the feature attribution scores of a DL sequence classification architecture. The presentation is divided into two portions; in the first, I discuss existing approaches to the task based on lexical scoring methods, and identify gaps in our methodological toolkit. In the second, I discuss the challenges of applying DL in a social science context, and provide several justifications for its application in this case.

Differentiating language usage between political campaigns is a broad task. One approach considers it as feature selection task, where the labels are the campaign, the data are the content of the advertisements, and the goal is to select the features of the data that are indicative of a label. We can add the restriction that these features should be substantively interpretable, given that we eventually want to use them to characterize the respective campaigns.

Monroe, Colaresi, and Quinn (2008) approach this task by considering lexical, i.e. word-usage, differences between the campaigns. Their method treats words as the unit of analysis in language, and the semantic values associated with words facilitates the interpretation of resulting outputs. Monroe, Colaresi, and Quinn (2008) use additional contextual information about their corpus, such as the fact that the primary topic of debate in the particular session of Congress was regarding abortion, to validate their measure and make additional claims about speaker pragmatics, such as strategy, inferable from observed word frequencies. In this example, they find that Republicans are associated with the pseudowords `kill`, `babi`, `procedur` and Democrats are associated with `woman`, `decis`, and `famili`. We are able to infer the framing strategies (Chong and Druckman 2007) employed by politicians on respective sides of the aisle based on these single words both because we infer a wider context in they occur.

However as noted, there are limits to what we can infer about the style or strategy of a campaign based on the meanings of individual words sans context. Although in the above examples, the broader intent of a statement containing the respective words can be inferred because it is less likely that the word “kill” will occur in a pro-choice statement in the context of a debate on abortion, using the word “kill” in a sentence does not inherently make it a pro-abortion stance. The semantic context of a word in an isolation can be highly ambiguous even where the word is not polysemous. For example, it is hard to guess whether the token `mother` should be associated with Republican or Democrat speech in an abortion context (as it turns out, `mother` is highly Republican).

The challenge faced by text-as-data strategies linking their theoretical and empirical estimand based on individual words and their inferred semantic context is that in most cases, individual words do not *produce* the phenomenon of interest, but only *correlate* with it. When what is being measured (the empirical estimand) is only linked to what is being studied (the theoretical estimand) by empirical regularity and not causality, it becomes important for the researchers to consider the conditions under which this regularity may not hold.[3]

An approach to directly circumvent this challenge is to recover the entire span (phrase in a broad sense) that contains the entire phenomenon of interest. This approach is challenging using current political science text-as-data methods for two reasons. The first reason relates to pre-processing: most text-as-data methods discard word order, representing documents as counts of tokens. Trying to recover span boundaries using such models is difficult, because the model does not treat language as an ordered sequence. The second reason relates to how we use the outputs of text-as-data models. Even if we are able to perfectly extract relevant spans from each text, these outputs cannot be aggregated straightforwardly like a numerical summary. In order to interpret the output of a span-extracting model, we can either use a secondary model to aggregate features of interest from these spans, but this fails to circumvent the original reason for taking a different approach.

One could simply list out the extracted spans and read them individually. Instead, I advocate an *annotative* approach, where the researcher highlights the relevant spans within each document. The time cost of this

[3] It is not my intention to say that researchers must therefore always prefer deductive strategies! On the contrary, the inferential leverage of an inductive approach provides many opportunities for progress in empirical social science research, and regardless a deductive approach to a process as complex as language will come with its own challenges.

approach is mitigated by the reduced time spent on validation, and can be further mitigated by randomly sampling a subset of texts to determine broader trends.

I apply and compare all of the above approaches in the *Results* section to show furthermore that they can be viewed as complementary and mutually validating. To the extent that each model treats the data differently but arrives at similar results, the researcher can be certain that the patterns they are observing are inherent to the data and not an artifact of the method they have chosen.

Deep Learning and Interpretation

Because of the limitations in modelling spans introduced by the standard pre-processing steps detailed above, I turn to a computational language model that models words within sequences. Among these, I use several variants of the DL `BERT` architecture (Devlin et al. 2019)[4], which has become dominant in NLP for achieving state-of-the-art performance in a broad variety of tasks with minimal task-specific engineering.

It is worth noting that despite the increasingly widespread application of ML tools in quantitative social science methodology, there has been limited transfer of complex neural network-based DL tools, themselves an important part of ML, to our field.[5] Whereas NLP has taken a strong turn towards DL in recent years, political science text-as-data applications remain primarily focused on non-DL methods. There are reasons for the lack of DL applications in political science which need to be addressed in order to justify its use in this paper.

In brief,[6] DL refers to a broad class of neural network models that pass data through *multiple* layers of nodes. Each node (sometimes called an artificial neuron) takes in a set of inputs, calculates their weighted sum, and then passes this value onto the next layer of the network. Often, before passing on the data, an "activation function" is applied to constrain the value of the output. The model is trained by passing the data through the entire network, calculating the loss compared to the true outcome, and then adjusting the weights on the

[4] *Variants of BERT*: for the sake of brevity, in this paper I refer to all transformer-based encoders for language modelling tasks as `BERT`, after the most famous architecture in Devlin et al. (2019). However, the models employed in this paper are all variants of the original model, not limited to DistilBERT, `RoBERTa`, `MiniLM`, and the Reimers and Gurevych (2019) modifications of all these. For an extensive explanation of how `BERT` works, see Rush (2018).

[5] Recent exceptions include Lall and Robinson (2021) and Torres and Cantú (2022). Additionally Rodriguez, Spirling, and Stewart (2020) include `BERT` as a benchmark to compare their method against.

[6] For recent and accessible introductions to neural networks and deep learning, I point the reader to Aggarwal et al. (2018) or Skansi (2018).

inputs at each node to minimize this error, iteratively. The loss function itself can be defined by the user. While each individual node is therefore relatively simple (analogous to a linear regression model), the ability to combine many of these units into complex configurations (called architectures) creates models that are both highly flexible and predictive.

However, in many of the typical applications for quantitative models in social science research, the ability to describe the relationship between the inputs and outputs is more important than the ability to accurately predict the output (Hofman, Sharma, and Watts 2017). Despite containing weights analogous to regression coefficients, it is difficult to make inferences about the relationship between the inputs and outputs on the basis of a DL network. In part this is because it is unclear what the weights on intermediary layers in the network describe, and in part because DL architectures typically have an enormous number of parameters[7] that make systematic evaluation impractical.

These challenges are not limited to the political science domain, and the need for explainability in various applications has motivated a "interpretable ML/DL" subfield (Linardatos, Papastefanopoulos, and Kotsiantis 2021 provide a recent review) focusing on the development of tools to interpret complex DL models. Of particular interest in this paper is a method for extracting the logic of DL text classifiers referred to as *rationale extraction* (Lei, Barzilay, and Jaakkola 2016).

This method uses the fact that when classifying a document, humans and algorithms typically base the label on a subset of the words in the text. This subset of the text that contains sufficient information to assign a label is referred to as a *rationale*. Various authors have developed axioms that rationales should fulfil (Jain et al. 2020). The first, called *faithfulness* (Lipton 2018), concerns the quality of the rationale extraction strategy. For a given classifier and rationale, a rationale is *faithful* if the extracted span "reflects the information actually used by said model". A further two criteria concern the quality of the rationale itself: it should be concise (*brevity*, Lei, Barzilay, and Jaakkola 2016) and contain all relevant information (*comprehensiveness*, Yu et al. 2019). A fourth condition concerns the relation between machine-generated rationales and human intuition; Wiegreffe and Pinter (2019) argues that rationales should make sense to human readers (*plausibility*). A further study establishes a benchmark for comparing machine and human-generated rationales (*ERASER*,

[7] Depending on the configuration, `BERT` has between 110 and 345 million parameters.

DeYoung et al. 2020).

Although rationale extraction is designed as a tool for the interpretation and evaluation of complex DL language models, I argue that it can be used to characterize the differences in language usage between political campaigns. Given a classifier for political advertisements that predicts the campaign as the label, we want to know the linguistic features in the advertisement that the model used to (correctly) label the advertisement. Provided that the extraction strategy is faithful, concise and comprehensive, rationales are the minimal span of text within a document that contains all of the information used by the model to make that prediction. When that prediction is correct, that span contains valid information for labelling the text. To the extent that the rationale is *plausible*, we can describe the relevant linguistic features that it captures.

A critical reader can point out that we do not know the kinds of difference that the model might identify, and these differences might leverage substantively meaningless regularities. For instance, when training a image classifier to label the pictures associated with the advertisements of this dataset, the model quickly learned to label images with a black bottom border as belonging to Warren's campaign because the majority of Warren images had this artifact whereas others did not. While this particular scenario can be prevented by careful pre-processing and cleaning of the data, the point remains that the discovered differences may be trivial from a human perspective.

There are several encouraging reasons to suspect *a priori* that this will not be the case. These relate to the `BERT` architecture and how it is used. Typically, researchers will download a `BERT` model that has been trained on large corpora and stored this information in the weights of the model. The pre-trained model is then "fine-tuned" on a smaller task-specific corpus. In most cases, a `BERT` model that has been pre-trained then fine-tuned will outperform a `BERT` model that has only seen the task-specific corpus. This technique of using non-task-specific information to improve performance on other tasks is referred to as *transfer learning*, and has been crucial in enabling the application of neural network architectures requiring enormous training data to data-scarce applications (Rogers, Kovaleva, and Rumshisky 2020). This strategy appears to work by encoding linguistic regularities in the weights of the model. Although researchers have found that `BERT` does not interpret language identically to humans, they have found that it has a hierarchical notion of syntax (Lin, Tan, and Frank 2019) and is able to infer semantic relations (Ettinger 2020). As such we have evidence that

BERT is able to identify the kinds of relations that we are interested in, although we are unsure whether this will hold across all applications.

Nevertheless, even if BERT does not differentiate campaigns using the same linguistic patterns as a human coder would, there are reasons to be interested in the patterns that it does use. For one, classifiers using BERT-like models have achieved superhuman accuracy on a variety of natural language benchmarks (Linardatos, Papastefanopoulos, and Kotsiantis 2021; Storks and Chai 2021). This means that researchers have found that for tasks where there is an objective truth (such as the task in this paper), BERT is able to classify correctly at a higher rate than human coders. Moreover, provided that the rationale extraction strategy is faithful, then the rationale text is in fact something that differentiates the classes. This ability to identify non-obvious differences is likely to be especially useful in cases where the differences are likely to be subtle, such as the comparison between Sanders and Warren. On this basis, I present BERT rationales as a strategy for characterizing the differences between political campaigns.

Methods and Data

In this section I explain the technicalities of the methods employed, the validation strategy, and the task and data that these methods are applied to.

Rationale Extraction and Integrated Gradients

Rationale extraction consists of two components: a feature attribution strategy that scores the input features in terms of their "contribution" to a prediction, and an aggregation strategy for generating rationales from the scored input features. In this paper, I use a feature attribution method known as Integrated Gradients (IG, Sundararajan, Taly, and Yan 2017), shown to provide faithful and consistent explanations for text classifications (Atanasova et al. 2020).

IG is a feature attribution method for neural network models, "defined as the path integral of the gradients along the straightline path from the baseline x' to x" (Sundararajan, Taly, and Yan 2017, 3), where x is the input of interest and x' is a baseline chosen by the researcher. In the case of image models, this baseline is typically a black image (of all zero inputs), whereas for BERT models we use an input with all tokens replaced with the special <pad> token used to denote an empty input. The final layer of a neural network classifier has an output node corresponding to each class; the value of the IG numerically approximates the difference in the value at the node of the target class between the baseline and the input. These raw prediction values (referred to as logits in the DL literature), are then transformed with a softmax function to sum to one to approximate probabilities.

Directly interpreting raw logits presents some challenges. At the final layer of the classifier, the monotonicity of the softmax transformation means the class corresponding to the node with the highest raw logit is the predicted class. The model is trained to minimize the cross entropy loss of the prediction, i.e. to minimise the log of the softmaxed logit corresponding to the target class. Thus because the softmax function takes all classes as arguments, in order to directly interpret raw logits for a given output we need to know the values of logits corresponding to the other classes as well.

Nevertheless, the IG scores can be interpreted in the following way. For a given IG score on an input feature, strongly positive values indicate that the input "pushes" the prediction more towards the target class, near-zero values indicate that the token provides little information, and strongly negative values indicate that the token reduce the likelihood of the target class. Within the same document, the magnitude of the raw logits[8] indicates relative importance of each token to the model's classification of the document. In the binary classification case, these values are symmetric between target classes, so negative values can be interpreted as being indicative of the negative class. In the multiclass case, positive scores have the same interpretation, but negative scores should be interpreted as being indicative of "not-" the target class.

In their typical usage, neural networks and their training can be probabilistic, but the predictions of these models and their attribution scores are deterministic. This is problematic in view of evidence of the instability of attribution methods (Atanasova et al. 2020). Given that an attribution score of zero has a clear substantive interpretation–the input does not contribute to the prediction–a measure of uncertainty is useful. This is particularly the case in a quantitative social science context where the ability to conduct statistical hypothesis

[8] Specifically, the magnitude of the attribution scores add up to the difference in predicted pre-regularization class likelihood between the given inputs and the reference inputs, which in this case is an equal-length vector containing the special <pad> token instead of token values.

tests is key.

A clever solution in the DL literature relies on the fact that many DL models include dropout layers. A dropout layer is an operation on a tensor that returns the input with some random subset of elements set to zero, with the probability of zeroing typically[9] uniform over the inputs with the probability set parametrically. Dropout layers are a proven tool for reducing overfitting (Srivastava et al. 2014), but Gal and Ghahramani (2016) shows that they can be used as a computationally efficient bootstrap for neural networks. This method is referred to as Monte Carlo Dropout (MC Dropout). I use MC Dropout to draw 30 estimates of the predicted class and attribution score. I use these 30 draws to generate bootstrapped confidence intervals for hypothesis testing, where I test whether the attribution score of a given feature is significantly different from zero.

Validation Strategies

I provide two means of validating the IG scores. In the first instance, I evaluate IG as a lexical scoring method by comparing the attribution scores to the lexical scoring methods presented in Monroe, Colaresi, and Quinn (2008). I briefly explain the two methods used for lexical scoring in this paper. The first, log-odds (LO), provides a symmetric measure of the likelihood of a token given the class of the speaker, centered at zero when the likelihood of use is equal between classes. Although I prevent tokens having an infinite/undefined log-odds ratio by adding $1e - 2$ to the odds, in presenting my results I only include the union of the vocabularies; i.e. tokens occurring in both classes (one or rest for the multiclass case). This is because including tokens only occurring in one class gives unstable results.

The second method, which I refer to as Fightin' Words (FW), is a Bayesian model for identifying distinguishing word usage presented in Monroe, Colaresi, and Quinn (2008). It builds on a simple comparison of usage rates (via LO, above) to include tokens that only occur in one set of documents by introducing a smoothing Dirichlet prior over all tokens. The z-scores indicate the number of standard deviations a word usage is away from the mean of no difference in usage between groups.

To use IG scores for lexical scoring, I normalize the IG scores within documents so that they add up to one, and calculate the average normalized score for each token lemma. Although this loses the interpretation

[9] Typically, but not necessarily, e.g. Li, Gong, and Yang (2016) propose multinomial dropout.

associated with the zero threshold, we can infer that tokens with high average normalized IG for a given candidate/campaign are tokens that are strongly indicative of that campaign in all instances where the token occurs.

These IG scores are compared with the Monroe, Colaresi, and Quinn (2008) scores both quantitatively for inter-correlation and qualitatively for *prima facie* informativeness. Although strong correlation and overlap would be indicative that IG is as valid as the established approaches in Monroe, Colaresi, and Quinn (2008), diverging results are more difficult to interpret. In one sense, we want a method that provides *additional* insights that existing methods cannot find; in another sense, diverging results require additional validation to demonstrate that the new results are also informative.

Thus we need a validation does not require comparison to existing methods. For this, I use the FRESH strategy described in Jain et al. (2020). Given the IG scores and their associated rationales, I first train a new classifier on a corpus consisting only of rationales. If the accuracy of the classifier does not drop significantly compared to the model trained on the original dataset, then we can infer that the rationales do contain an informative signal of the class (i.e. the attribution method has high precision). Furthermore, I train a third classifier on a dataset of the texts minus the rationales. If the accuracy of this classifier drops, then we can infer that the rationales themselves do not omit relevant information for inferring class (high recall).

I compare three strategies for extracting rationales from documents based on IG scores. The first two, used in Jain et al. (2020), are selecting the **top-k** tokens by attribution score and the **contiguous** length-k region with the highest summed attribution, where k is one-tenth the length of the document. The third strategy is to use the contiguous positive region with highest average attribution score. This has the advantage of being non-parametric, order-and-context-preserving and excluding any negative regions, but has the disadvantage of potentially returning no tokens whatsoever (if all attribution scores are zero or negative).

To validate the rationales, I use the FRESH strategy described in Jain et al. (2020). Given the IG scores and their associated rationales, I first train a new classifier on a corpus consisting only of rationales. If the accuracy of the classifier does not drop significantly compared to the model trained on the original dataset, then we can infer that the rationales do contain an informative signal of the class (i.e. the attribution method has high precision). Furthermore, I train a third classifier on a dataset of the texts minus the rationales. If the accuracy of this classifier drops, then we can infer that the rationales themselves do not omit relevant information for inferring class (high recall).

Finally, I use the IG scores to find and evaluate the features of an advertisement that differentiate it from other campaigns, and characteristic of its own campaign. To do this, I adopt a method used in computer vision and NLP of superimposing heat maps onto the input data to highlight high-salience regions. These annotations can intuitively indicate to a researcher the portions of a text that are characteristic or differentiating.

Data and Task

I use ProPublica's publicly available collection of political Facebook advertisements from 2016-2020. These advertisements were collected using a browser plugin installed by volunteers. Whenever a participating individual was served an advertisement on Facebook, the advertisement was scraped and sent to servers hosted by ProPublica, who then used a mixture of human and automated classification to determine the advertisement to be political or not. The raw content plus some pre-parsed information of the advertisements classified as political is published on their website as a continuously updating dataset.

I am using the subset of these advertisements for which I am able to obtain a single, sufficiently large image that accompanies the text. From these, I exclude any advertisements from advertisers that have fewer than 100 observations in the dataset. I manually merged sufficiently similar advertisers from the list of advertisers with at least 30 advertisements in this dataset (e.g. Planned Parenthood with Planned Parenthood Action).

My final dataset for analysis contains 85,664 advertisements from 228 unique advertisers. For each advertisement, I retain a unique advertisement id provided by ProPublica, the advertiser, the text of the advertisement as parsed from the raw html using the `BeautifulSoup` library, the year that the advertisement was added to the ProPublica dataset, and several manually-assigned labels based on the advertiser. These labels are the broad type of advertiser (candidate, PAC, organisation, business), a subtype (e.g. category of race if candidate, main issue if organisation), and alignment as either left, right, Democrat or Republican (again, depending on which is more appropriate).

Note that the distribution of unique advertisements in this dataset is heavily skewed towards liberal/Democrat audiences. I assume that this is because recruitment of volunteers was more successful among liberal/Democratic individuals than conservatives/Republicans. Given that I have a large number of advertisements for Democratic candidates at all levels, from gubernatorial and state legislature to presidential level, and I have close to no advertisements for republican candidates other than Donald Trump, I do not try to make inferences about campaigns other than the largest one, the Trump campaign. Although there are a number of advertisements likely tailored for conservatives in the dataset, the imbalance in coverage indicates to me that there is a large amount of missing data for smaller races.

Three comparisons are made within the corpus. The first comparison is Donald Trump and Joe Biden (527 advertisements). This comparison is made both because they were the eventual presidential nominees of their respective parties, but is presented first because we have the strongest preconceptions about what differences we should find between Trump and Biden, especially given the idiosyncratic style of Trump. Thus differences we find between the campaigns are less likely to be surprising, and more likely to be confirmatory.

The second comparison is Bernie Sanders and Elizabeth Warren (829 advertisements). This comparison is made because they represent the two most left-wing primary candidates in the Democratic primary, and therefore we have fewer expectations about the differences we expect to find in their campaign advertisements. The differences identified in this comparison therefore shows the utility of the applied methods to identify subtler differences when the expected similarity is high.

The final comparison is between Amy Klobuchar, Bernie Sanders, Beto O'Rourke, Cory Booker, Elizabeth Warren, Joe Biden, Kamala Harris, Kirsten Gillibrand, Micheal Bloomberg, Pete Buttigieg and Tom Steyer (3165 ads). This comparison shows the utility of these methods beyond the binary case, which is particularly relevant for the applicability of these methods in a multiparty setting. It also reveals information about the segments of the Democrat primary electorate that the respective primary candidates view as their target constituency. In order to adapt LO and FW to the multiclass case I employ a one-vs-rest strategy.

Pre- and Post-Processing

Pre-processing steps in automated text analysis broadly encompass the initial operationalisation of the text in a numerical (and usually tabular) format prior to its use in a model. Denny and Spirling (2018) show that steps taken at this stage typically have non-trivial impacts on the final output of the model. In order to simplify comparisons, I unify pre-processing steps between the three models where possible. I discuss two key parts of the pre-processing: text cleaning and tokenization.

Text cleaning helps to remove artefacts from the original data collection process and reduces noise, but also risks introducing systematic bias by omitting particular features. For all three models I simply removed all HTML tags, and otherwise kept the text intact (leaving artefacts such as consecutive blank spaces and so on). A custom stopword list was created to remove non-words and other parsing errors.

Tokenization refers to the conversion of sequence of characters in a text to a sequence of "tokens", the base input for all models considered. Relevant decisions include conversion to lowercase, stemming/lemmatization, and even decision of what constitutes a token (i.e. determining token boundaries). For all models, I did *not* convert to lowercase because capitalization it conveys relevant information about tone, emphasis and style to the reader. Instead of using a blunt stemming method, I used the lemmatization algorithm provided by the SpaCy library. This reduces homonymy and leaves tokens more interpretable.

For the LO and FW approaches, I also used SpaCy's language model to determine token boundaries. This has the advantage over splitting on whitespace of handling contractions as two words, and then reduce them to their respective lemmas. For IG, because the classification model is based on `RoBERTa` (Liu et al. 2019), tokenization is done with a pre-trained byte-pair encoding (BPE) algorithm. I also experiment with `MiniLM` (Wang et al. 2020), `DistilRoBERTa` (Liu et al. 2019; Sanh et al. 2019) and MPNet (Song et al. 2020), and pre-train my own model and tokenizer based on the `DistilRoBERTa` architecture (presented in the appendix).

In order to make direct comparisons between the three methods, in post-processing I match the token scores produced by the pre-trained BPE to match the boundaries of SpaCy tokens. For the most part, because BPE produces subword tokens, this is a simple case of concatenating tokens where a word is split into several word-parts, but in the rare reverse of a single BPE token corresponding to multiple SpaCy tokens (such as an apostrophe in the middle of a word, which is represented by two tokens in `RoBERTa`), I transfer the score to the first token and set the score of the latter as NaN. There were no cases of split boundaries between the two tokenization schemes.

With all comparisons, there is the decision of how to deal with tokens that occur in documents belonging to one class and not the others. This is a particular issue for LO, as the log-odds ratio is undefined/infinite. This is typically remedied by adding noise to the odds ratios (Monroe, Colaresi, and Quinn (2008) add 0.5 to counts). This still results in very large log-odds scores for tokens that only occur in one class, meaning that word lists of the top words from each class will simply capture this.

Monroe, Colaresi, and Quinn (2008) argue that this is a limitation, and account for this in the FW approach by giving a Dirichlet prior for the baseline probability of tokens occurring in a document. For transformer (and other DL) models, the baseline probability of a token is learned from the pre-training corpus, with no word containing only characters in the tokenizer vocabulary having a uninitialized corresponding gradient (although obviously, combinations of tokens may not have been encountered during the pre-training).

During fine-tuning, we should expect transformer models to converge towards predicting documents on the basis of individual tokens that perfectly predict class in the training data. In order to mitigate this, I use both a low learning rate ($1e-5$ to $5e-5$) and dropout layers both in the embedding block at the beginning of the model and in the classification block at the end of the model, to penalise localized learning.

Results

I present the results in three parts. The first part compares IG to LO and FW as a lexical scoring method. I find that whereas LO and FW provided highly correlated rankings (both by Pearson and Spearman correlations), IG is relatively uncorrelated with either. Nevertheless, by inspecting the words that each ranks highly, we see that IG does not give misleading or unhelpful results. The second part presents the results of rationale generation, finding that the IG scores are stable the rationales faithful. The final part discusses what we can infer from scores and demonstrates how they can be used to annotate documents.

Lexical Selection

I first compare the Spearman rank correlation of the three measures. Figure 1 shows the correlation matrix for the Biden vs Trump and Sanders vs Warren comparisons; the correlation matrix of all primary candidates is in the appendix. In each matrix, there are four columns/rows, corresponding to the ranking of tokens by Log-Odds Ratio, FW Z-Score, and the non-normalized feature attribution scores for each candidate (note that the rankings that these will not be symmetrical, as each is calculated only within the speaker's own documents).

Several patterns stand out. For both pairs, the rankings provided by LO and FW are highly correlated (approximately 0.86). For Biden and Sanders, there is a negative correlation due to the alignment of the LO and FW rankings treating Trump/Warren as the "positive" pole. In general IG is weakly correlated with LO and FW, but there is no fixed pattern to this weak correlation.

How do these rankings compare at the top end of the distribution? I present the top ten tokens by candidate for each comparison in figure 2 In contrast to the visualisations of Monroe, Colaresi, and Quinn (2008), I plot the information as annotated heat maps to convey the information more concisely. The figure consists of subplots, each containing the top 10 ranked tokens by candidate for each comparison, colored by score. The top row (orange) shows the absolute log-odds ratio. The middle row shows the absolute z-score from Monroe, Colaresi, and Quinn (2008). The bottom row shows the average feature attribution. The scale for all three sets of tables is shown on the right-hand-side of the figure. For LO and IG, only tokens occurring in at least two classes are included; a version of this figure that does not exclude these tokens is in the appendix.

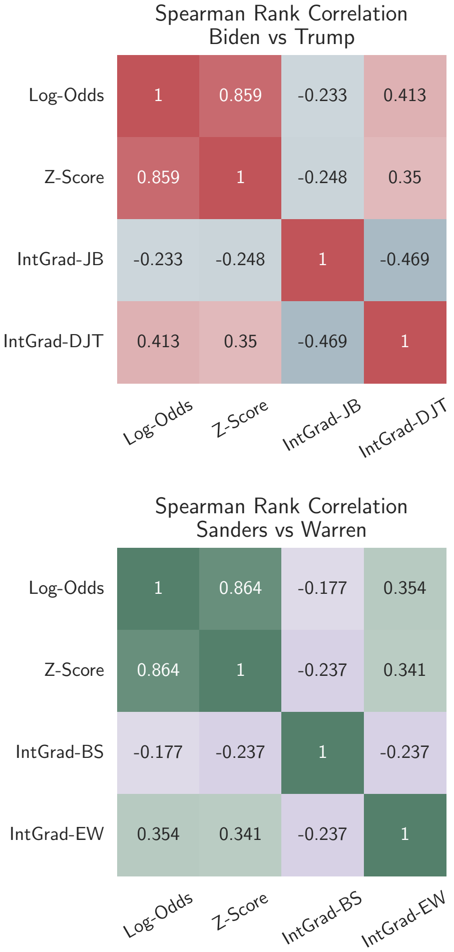
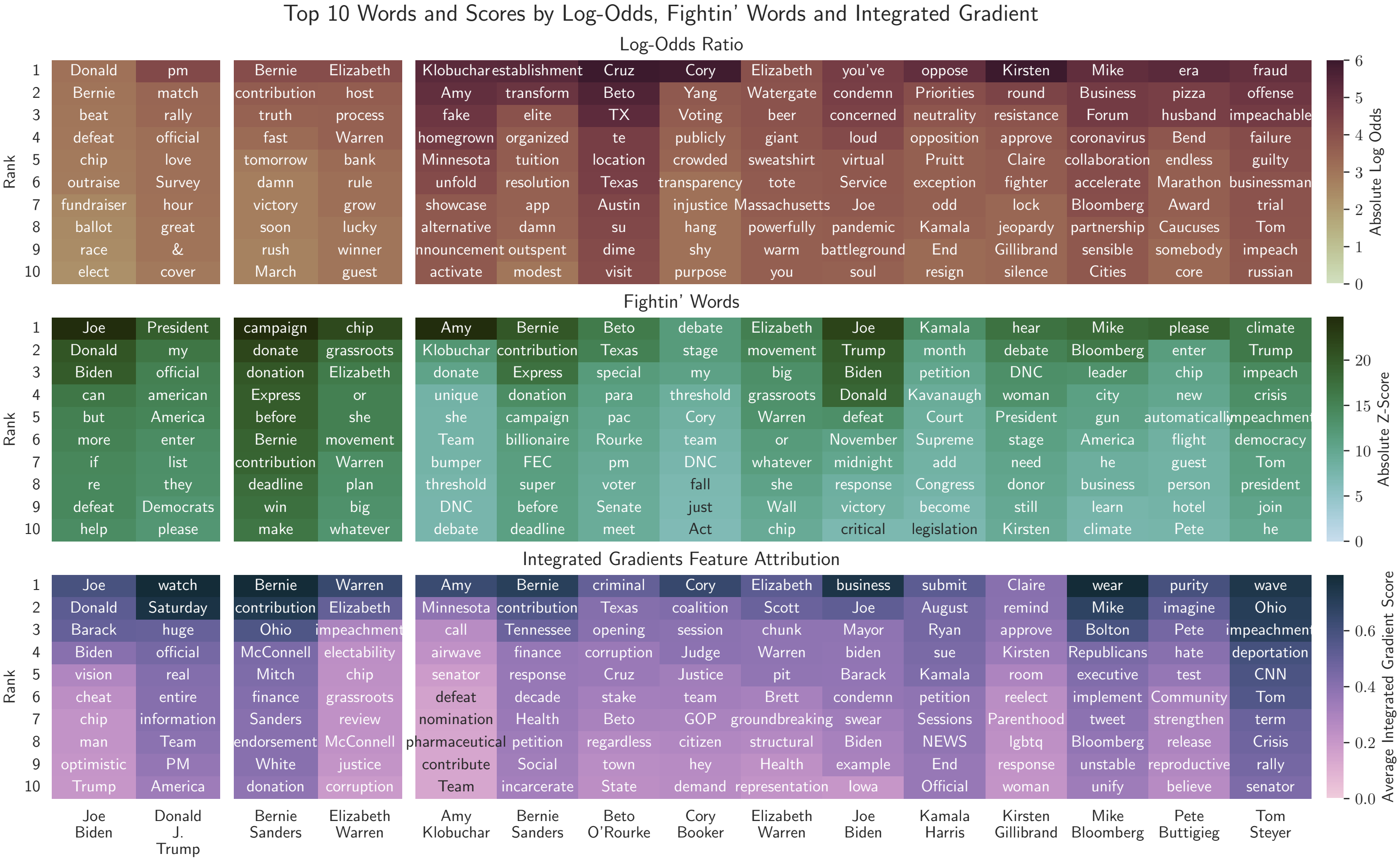


Figure 1: Spearman rank correlation matrices for LO, FW and IG scores.

Although the extent of characterization and inference that can be made from a small subset of words is limited, interesting comparisons can be made in multiple directions. One is between



Log-Odds Ratio

Rank	Joe Biden	Donald J. Trump	Bernie Sanders	Elizabeth Warren	Amy Klobuchar	Bernie Sanders	Beto O'Rourke	Cory Booker	Elizabeth Warren	Joe Biden	Kamala Harris	Kirsten Gillibrand	Mike Bloomberg	Pete Buttigieg	Tom Steyer
1	Donald	pm	Bernie	Elizabeth	Klobuchar	establishment	Cruz	Cory	Elizabeth	you've	oppose	Kirsten	Mike	era	fraud
2	Bernie	match	contribution	host	Amy	transform	Beto	Yang	Watergate	condemn	Priorities	round	Business	pizza	offense
3	beat	rally	truth	process	fake	elite	TX	Voting	beer	concerned	neutrality	resistance	Forum	husband	impeachable
4	defeat	official	fast	Warren	homegrown	organized	te	publicly	giant	loud	opposition	approve	coronavirus	Bend	failure
5	chip	love	tomorrow	bank	Minnesota	tuition	location	crowded	sweatshirt	virtual	Pruitt	Claire	collaboration	endless	guilty
6	outraise	Survey	damn	rule	unfold	resolution	Texas	transparency	tote	Service	exception	fighter	accelerate	Marathon	businessman
7	fundraiser	hour	victory	grow	showcase	app	Austin	injustice	Massachusetts	Joe	odd	lock	Bloomberg	Award	trial
8	ballot	great	soon	lucky	alternative	damn	su	hang	powerfully	pandemic	Kamala	jeopardy	partnership	Caucuses	Tom
9	race	&	rush	winner	nnouncement	outspent	dime	shy	warm	battleground	End	Gillibrand	sensible	somebody	impeach
10	elect	cover	March	guest	activate	modest	visit	purpose	you	soul	resign	silence	Cities	core	russian

Fightin' Words

Rank	Joe Biden	Donald J. Trump	Bernie Sanders	Elizabeth Warren	Amy Klobuchar	Bernie Sanders	Beto O'Rourke	Cory Booker	Elizabeth Warren	Joe Biden	Kamala Harris	Kirsten Gillibrand	Mike Bloomberg	Pete Buttigieg	Tom Steyer
1	Joe	President	campaign	chip	Amy	Bernie	Beto	debate	Elizabeth	Joe	Kamala	hear	Mike	please	climate
2	Donald	my	donate	grassroots	Klobuchar	contribution	Texas	stage	movement	Trump	month	debate	Bloomberg	enter	Trump
3	Biden	official	donation	Elizabeth	donate	Express	special	my	big	Biden	petition	DNC	leader	chip	impeach
4	can	american	Express	or	unique	donation	para	threshold	grassroots	Donald	Kavanaugh	woman	city	new	crisis
5	but	America	before	she	she	campaign	pac	Cory	Warren	defeat	Court	President	gun	automatically	impeachment
6	more	enter	Bernie	movement	Team	billionaire	Rourke	team	or	November	Supreme	stage	America	flight	democracy
7	if	list	contribution	Warren	bumper	FEC	pm	DNC	whatever	midnight	add	need	he	guest	Tom
8	re	they	deadline	plan	threshold	super	voter	fall	she	response	Congress	donor	business	person	president
9	defeat	Democrats	win	big	DNC	before	Senate	just	Wall	victory	become	still	learn	hotel	join
10	help	please	make	whatever	debate	deadline	meet	Act	chip	critical	legislation	Kirsten	climate	Pete	he

Integrated Gradients Feature Attribution

Rank	Joe Biden	Donald J. Trump	Bernie Sanders	Elizabeth Warren	Amy Klobuchar	Bernie Sanders	Beto O'Rourke	Cory Booker	Elizabeth Warren	Joe Biden	Kamala Harris	Kirsten Gillibrand	Mike Bloomberg	Pete Buttigieg	Tom Steyer
1	Joe	watch	Bernie	Warren	Amy	Bernie	criminal	Cory	Elizabeth	business	submit	Claire	wear	purity	wave
2	Donald	Saturday	contribution	Elizabeth	Minnesota	contribution	Texas	coalition	Scott	Joe	August	remind	Mike	imagine	Ohio
3	Barack	huge	Ohio	impeachment	call	Tennessee	opening	session	chunk	Mayor	Ryan	approve	Bolton	Pete	impeachment
4	Biden	official	McConnell	electability	airwave	finance	corruption	Judge	Warren	biden	sue	Kirsten	Republicans	hate	deportation
5	vision	real	Mitch	chip	senator	response	Cruz	Justice	pit	Barack	Kamala	room	executive	test	CNN
6	cheat	entire	finance	grassroots	defeat	decade	stake	team	Brett	condemn	petition	reelect	implement	Community	Tom
7	chip	information	Sanders	review	nomination	Health	Beto	GOP	groundbreaking	swear	Sessions	Parenthood	tweet	strengthen	term
8	man	Team	endorsement	McConnell	pharmaceutical	petition	regardless	citizen	structural	Biden	NEWS	lgbtq	Bloomberg	release	Crisis
9	optimistic	PM	White	justice	contribute	Social	town	hey	Health	example	End	response	unstable	reproductive	rally
10	Trump	America	donation	corruption	Team	incarcerate	State	demand	representation	Iowa	Official	woman	unify	believe	senator

Figure 2: Top 10 tokens by candidate by comparison for Log-Odds, Fightin' Words and Integrated Gradient.

models, seeing the different tokens that the respective models emphasise. The other is within the same model, comparing how changing the reference category affects the top 10 tokens of a campaign.

Each model emphasises different aspects. LO is a simple, normalized and symmetric ranking of tokens occurring very frequently in one class but not the other. We can see that `Donald` is the token with the highest LO ratio for Joe Biden when comparing Biden to Trump, followed by `Bernie`, `beat`, `defeat` and so on. In contrast, the top LO tokens for Trump when compared to Biden are `pm`, `match`, `rally`, `official` and `love`. We might infer from this that compared to Trump's advertisements, many of Biden's advertisements are about defeating his opponents, whereas Trump advertises events/deadlines (such as rallies, funding deadlines and so on).

In contrast, FW can be interpreted as the increased likelihood of using a particular term given speaker class, relative to a baseline probability of using any word. This hints at a similar interpretation for Trump, highlighting usage of the tokens `my`, `official`, `american`, but shows a semantically less interpretable set of tokens for Biden: `can`, `but`, `more`, `if`. A possible interpretation here is that Trump advertisements are more likely to use declarative statements, in which case qualifying verbs and hypothetical conjunctions are less likely to be used relative to Biden advertisements.

IG emphasizes a similar pattern overall with more resemblance to LO in interpretation. Top Biden tokens include names (`Joe`, `Donald`, `Barack`, `Biden`) whereas Trump tokens include language advertising events or other activities (`watch`, `Saturday`, `PM`, `information`). Across all three rankings, Trump advertisements are associated with adjectives pertaining to a semantic field of legitimacy and grandeur (`official`, `great`, `huge`, `real`), which appears to reflect Trump's style.

Comparing Sanders and Warren, Sanders' advertisements are more associated with campaigning and fundraising language (`contribution`, `victory`, `donation`, `donate deadline`, `finance`, `endorsement`) and Warren's advertisements are more associated with specific issues and a grassroots movement (`grassroots`, `movement`, `justice`, `corruption`, `impeachment`). This pattern holds less when the comparison includes all other democratic primary candidates. Sanders' advertisements are then more associated with his signature issues and style, with tokens such as (`establishment`, `transform`, `elite`, `tuition`, `damn`, `billionaire`, `Health`), while tokens relating to fundraising rank somewhat higher again for Warren (`beer`, `tote`).

The extent of inference that can be made from these tokens without the context in which they occur is limited. Nevertheless, it appears that all three measures highlight *prima facie* sensible characterizations of each campaign. It is also clear that from these maps that what “characterizes” a campaign very much depends on what that comparison is being made against.

Attribution and Rationales

Although feature attribution based on IG has provided plausible lexical rankings in this instance, what can we say about their performance more generally? In this section I discuss the stability, reliability and interpretation of IG in context.

Using MC Dropout, for each document I sample 30 draws from the posterior feature attribution distribution. Figure 3 compares these distributions for a Biden advertisement for the two different classifiers. The blue circles show the per-token attribution from the classifier differentiating Trump and Biden, and the orange diamonds show the same for the Democratic primary. Each point shows the estimated attribution score with bootstrapped ($n = 1000$) 95 percent confidence intervals for each token (small horizontal lines).

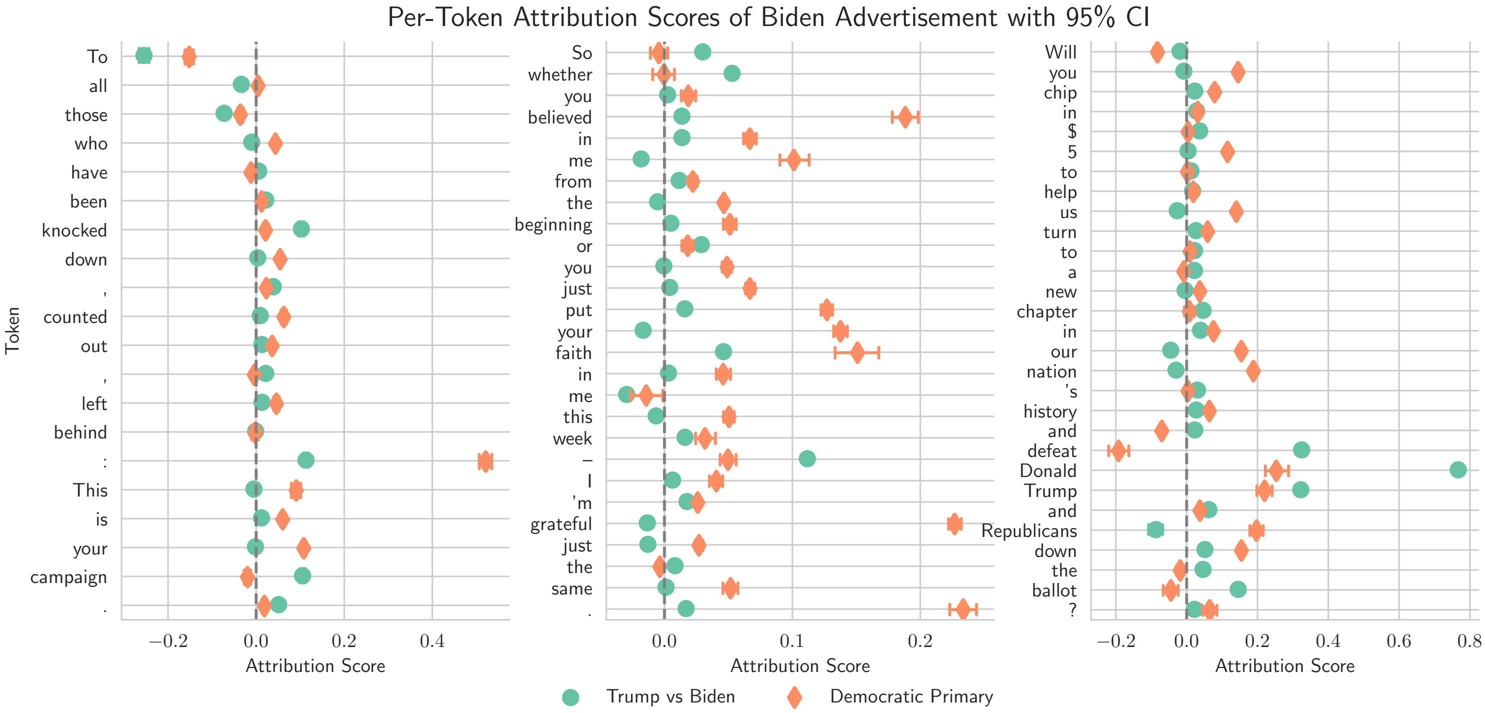


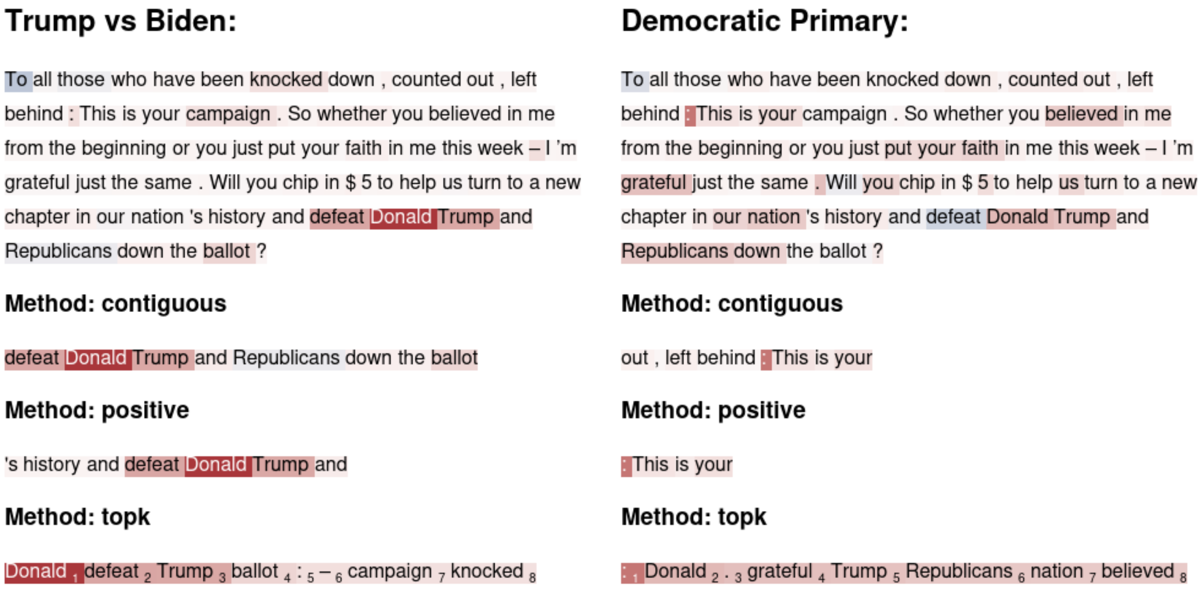
Figure 3: Per-Token Feature Attribution (bootstrapped 95% CI). *Note changing x-axis scale.*

Several patterns emerge. Firstly, there appears to be a mixture of "high", "medium" and "low" attribution regions. For the Trump versus Biden classifier, there is a single phrase with a high attribution score: "defeat Donald Trump" in the third sentence. In contrast, the Democratic primary classifier assigns high scores to "Donald Trump" but a negative score to "defeat", and its highest score is on the punctuation mark ":" in the first sentence. The Democratic primary likewise assigns significant scores to several ranges of the text that the Trump versus Biden model does not: "you believed in me", "you just put your faith in", "in our nation".

Consider the phrase "defeat Donald Trump" in the third sentence (right-hand panel, figure 3). The substantive interpretation of the attribution scores is that while the phrase "Donald Trump" is indicative of a Biden advertisement, the token "defeat" is only associated with Biden when compared to Trump advertisements, and indicative of non-Biden advertisements when comparing to other Democratic candidates. Additionally, the token "Donald" is a strong indicator of a Biden advertisement when compared to Trump, whereas this is not the case when compared to other Democratic candidates.

It is unsurprising that the phrase “defeat Donald Trump” indicates that it is the advertisement from Donald Trump’s opponent. Comparing Biden to the other democratic primary candidates, we see instead the phrases “: This is your campaign”, “put your faith”, “I’m grateful”, “our nation” and “Donald Trump and Republicans” highlighted. The former four phrases point to a rhetorical style possibly pertaining to a semantic field of Christian morality (collectivism, faith, humility) being employed by Biden’s campaign, and the latter may indicate that most Biden advertisements make reference to the opponents to be defeated.

Are these high-attribution phrases actually sufficient to differentiate Biden advertisements from other campaigns? I use the FRESH validation strategy of Jain et al. (2020) to test whether this is the case. Remember that for an explanation to be *faithful* it must actually reflect the information used by the model to come to its decision (Lipton 2018). This is equivalent to our application, where we are interested in the “logic” of the classifier because we want to know what features it uses to differentiate speakers. Therefore we need to confirm that our feature selection strategy highlights inputs that the model actually determines to be differentiating.



Trump vs Biden:

To all those who have been knocked down , counted out , left behind : This is your campaign . So whether you believed in me from the beginning or you just put your faith in me this week – I ’m grateful just the same . Will you chip in $ 5 to help us turn to a new chapter in our nation 's history and defeat Donald Trump and Republicans down the ballot ?

Method: contiguous

defeat Donald Trump and Republicans down the ballot

Method: positive

's history and defeat Donald Trump and

Method: topk

Donald $_1$ defeat $_2$ Trump $_3$ ballot $_4$: $_5$ – $_6$ campaign $_7$ knocked $_8$

Democratic Primary:

To all those who have been knocked down , counted out , left behind : This is your campaign . So whether you believed in me from the beginning or you just put your faith in me this week – I ’m grateful just the same . Will you chip in $ 5 to help us turn to a new chapter in our nation 's history and defeat Donald Trump and Republicans down the ballot ?

Method: contiguous

out , left behind : This is your

Method: positive

: This is your

Method: topk

: $_1$ Donald $_2$. $_3$ grateful $_4$ Trump $_5$ Republicans $_6$ nation $_7$ believed $_8$

Figure 4: Rationale Generation

Figure 4 shows the text of the same Biden advertisement, shaded by mean attribution score over 30 draws for each model. Darker shades indicate a stronger score, and red indicates the positive class (Biden) whereas blue indicates the negative class (not-Biden). These attribution heat maps help visualize the three rationale extraction strategies. Two are from Jain et al. (2020) (**contiguous**, **topk**), and one is an original approach (**positive**). The contiguous and topk strategies extract fixed proportions of the document as a rationale; contiguous selects the length k region with the highest sum attribution score, whereas topk chooses k tokens with the highest scores. Because my aim is to flexibly identify variable-width spans for characterization, **positive** chooses the contiguously positive segment with the highest average attribution score.[10]

The length of the rationale is an informational trade-off. Shorter rationales are more parsimonious, but are more likely to exclude informative spans. For the **topk** and **contiguous**, I set rationale length to one-tenth the length of the document, as in Jain et al. (2020). For **positive**, the average rationale length is 19.3 percent of the document. Figure 5 shows the distribution of rationale size, broken down by comparison group. Rationales for the Democratic Primary were on average the smallest proportion of the text, but the distribution also has the longest tail and smallest inter-quartile range.

Jain et al. (2020) state that to fulfil the *faithfulness* criterion, models trained on the rationales should also be able to accurately predict labels. Intuitively, this shows that extracted rationales contain enough signal that a model that only sees the extracted text would be able to infer the correct label, without leveraging any other information available to the original model. In our case, this would show that highlighted portions of the text are relevant for differentiating the speaker.

I add an additional test to measure the "recall" of a rationale extraction strategy; training a model on a dataset that contains all text except the rationales. To the extent that the accuracy of the model trained on rationales is higher than the model trained on the text-minus-rationale, we can be sure that the strategy has extracted a greater proportion of the identifying information. Note that we should not expect the text-minus-rationale model to have an accuracy of zero with fixed-length rationales; there is no way to know *a priori* the proportion of tokens in an advertisement that are indicative of its speaker.

[10] Jain et al. (2020) also train a model (`BERT`) to predict rationales, and then use this to generate new rationales. I omit this approach because of the very low accuracy achieved by the model during training.

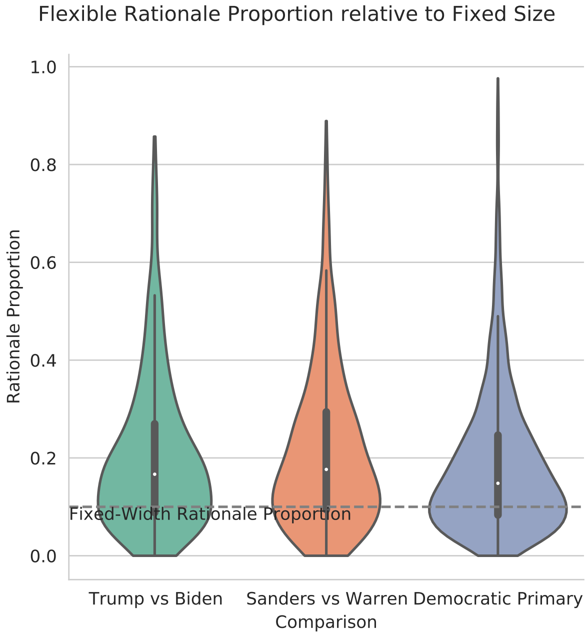


Figure 5: Distribution of Rationale Sizes per Comparison Group

Figure 6 shows accuracy of the classifiers trained on rationales or text-minus-rationale for each of the three strategies, and compares them to the accuracy of the base model with the full texts. Although all models save the model trained on the **topk** rationales have a lower accuracy than the base model, consider that the models trained on the rationales only saw a small subset of the full text. For context, consider the difficulty a trained human coder would face labelling the rationales in figure 4. Although `defeat Donald Trump` is sufficient to distinguish Trump and Biden, most coders would be hard-pressed to label the Democratic primary rationales: `out, left behind: This is your`, `:This is your` and `: Donald . grateful Trump Republicans nation believed`. Furthermore, the drop in accuracy for all strategies shows that excluding these high-attribution score tokens greatly increases the difficulty of correctly classifying the advertisement. As I discuss in this next section, this result is helpful for showing the utility of attribution-based annotation as a augmentative analysis tool.

Interpreting Annotation

Having shown that IG functions both as a lexical selection strategy and a method for explicating the patterns identified by transformer-based classifiers, I demonstrate one further utility of this tool for our question. Annotating documents with token-level attribution scores is a helpful aid for qualitative text analysis strategies such as content analysis, by incorporating information on corpus-level patterns into individual documents.

Figure 7 contains the same attribution heat maps as figure 4, but compares three advertisements each from the Sanders and Warren campaigns. The left column shows texts highlighted by attribution score from a

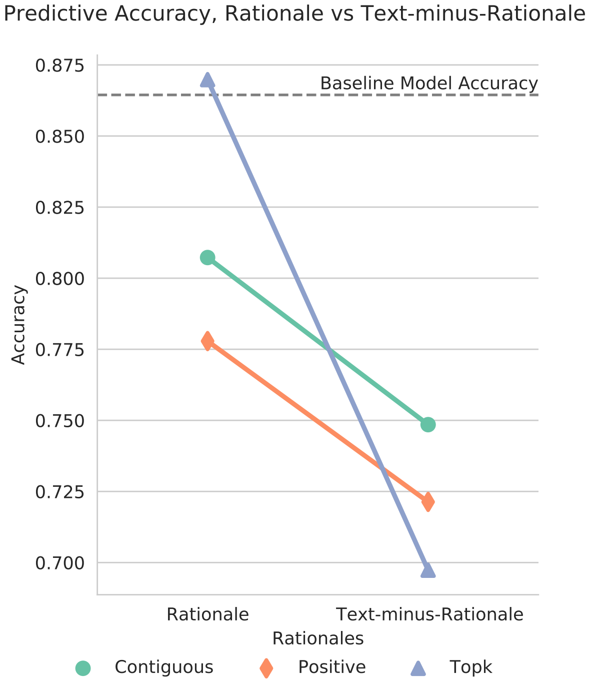


Figure 6: Accuracy of Models trained on Rationale versus Text excluding Rationale

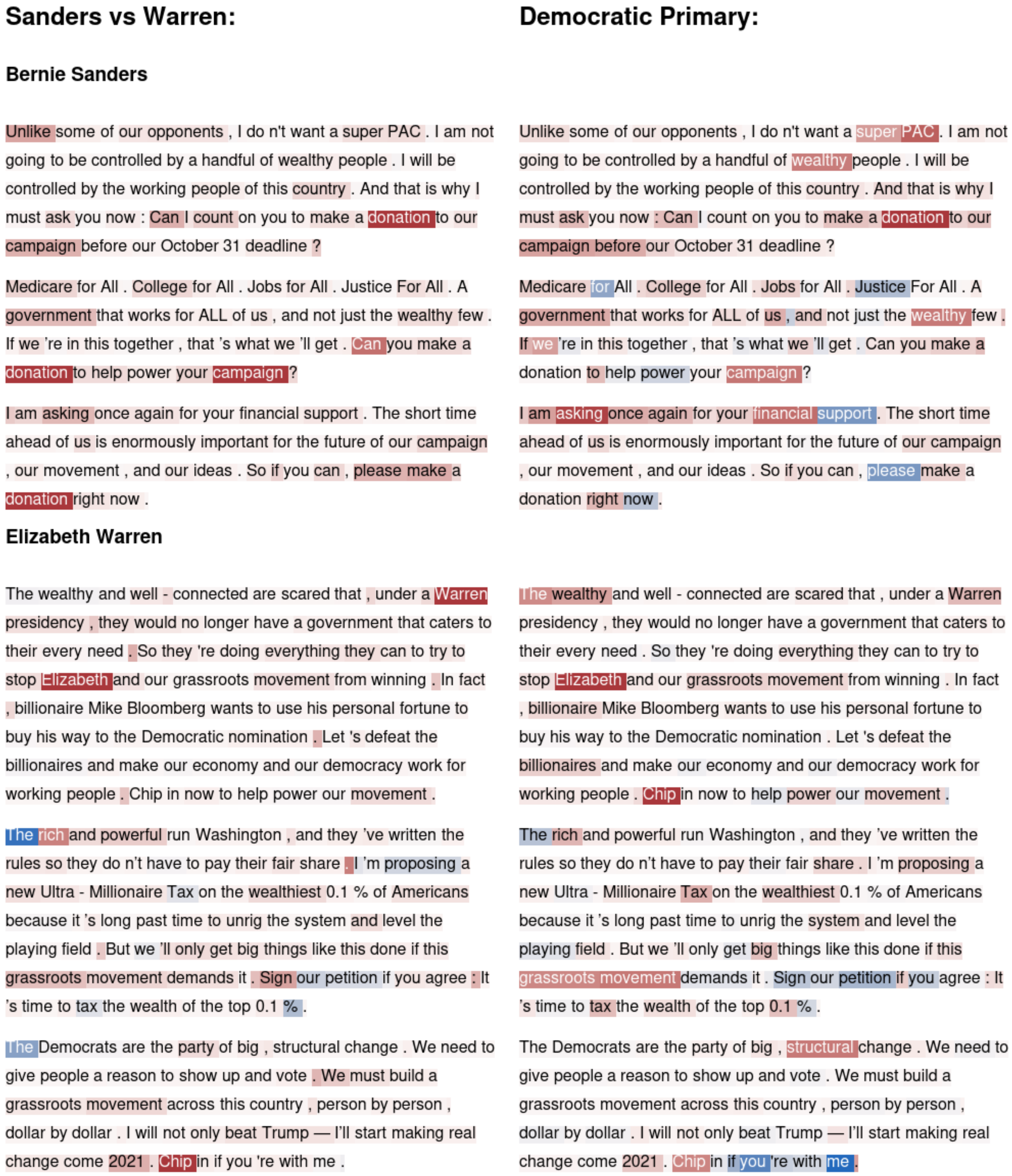
classifier trained to differentiate Sanders and Warren, while the right column shows the same for the eleven Democratic Primary candidates.

The highlighting helps the researcher by emphasizing the tokens that are distinctive, and de-emphasizing the tokens that are not. In the first Sanders advertisement, the phrases `super PAC` and `wealthy` are salient when comparing to all primary candidates, but not when comparing only to Warren. This suggests that within the eleven Democratic primary candidates, both Sanders and Warren are more likely to use these terms. In the second Sanders advertisement, several tokens reverse attribution, including `Justice`. This suggests that the theme of "Justice" is more associated with another primary candidate than either Sanders or Warren. Finally, the often referenced phrase `I am once again asking for your financial support` is scored higher when comparing to the primary candidates, whereas `donation` is no longer salient. This suggests that while `donation` is a strong predictor of Sanders when compared to Warren, this is no longer the case when comparing to primary candidates, and the model learns to search for larger identifying spans.

Similar inferences can be made looking at the three Warren advertisements. The first pair tells us that the words `wealthy` and `chip (in)` are characteristic of Warren when comparing to the larger group of eleven primary candidates, but not when only comparing to Sanders. The second pair shows a stronger focus onto the phrase `grassroots movement`, and a flip on the phrase `Sign our petition`. The third shows that a similar pattern for the phrase `structural change`, as characterizing of Warren compared to the other ten primary candidates.

There are limitations to what we can learn from this strategy. Although we have shown IG attribution is a faithful way of identifying *which* information that models use to classify texts, and that this information is provably differentiating between classes, it does not tell us *why* a given token or span is salient. For instance, we see that `wealthy` is salient for Sanders and Warren when comparing to all Democratic primary candidates, but not when only comparing to each other. It takes further contextual knowledge and interpretation from the researcher to then plausibly infer that this is because Sanders and Warren both campaign on platforms of opposing wealth inequality, whereas other primary candidates do not make it their signature issue.

On the other hand, this method is not meant to eliminate the need for domain-specific knowledge, but rather to direct the attention of a researcher to which pieces of information are salient *in context*. As with the other



Sanders vs Warren:

Bernie Sanders

Unlike some of our opponents , I do n't want a super PAC . I am not going to be controlled by a handful of wealthy people . I will be controlled by the working people of this country . And that is why I must ask you now : Can I count on you to make a donation to our campaign before our October 31 deadline ?

Medicare for All . College for All . Jobs for All . Justice For All . A government that works for ALL of us , and not just the wealthy few . If we 're in this together , that 's what we 'll get . Can you make a donation to help power your campaign ?

I am asking once again for your financial support . The short time ahead of us is enormously important for the future of our campaign , our movement , and our ideas . So if you can , please make a donation right now .

Elizabeth Warren

The wealthy and well - connected are scared that , under a Warren presidency , they would no longer have a government that caters to their every need . So they 're doing everything they can to try to stop Elizabeth and our grassroots movement from winning . In fact , billionaire Mike Bloomberg wants to use his personal fortune to buy his way to the Democratic nomination . Let 's defeat the billionaires and make our economy and our democracy work for working people . Chip in now to help power our movement .

The rich and powerful run Washington , and they 've written the rules so they do n't have to pay their fair share . I 'm proposing a new Ultra - Millionaire Tax on the wealthiest 0.1 % of Americans because it 's long past time to unrig the system and level the playing field . But we 'll only get big things like this done if this grassroots movement demands it . Sign our petition if you agree : It 's time to tax the wealth of the top 0.1 % .

The Democrats are the party of big , structural change . We need to give people a reason to show up and vote . We must build a grassroots movement across this country , person by person , dollar by dollar . I will not only beat Trump — I'll start making real change come 2021 . Chip in if you 're with me .

Democratic Primary:

Unlike some of our opponents , I do n't want a super PAC . I am not going to be controlled by a handful of wealthy people . I will be controlled by the working people of this country . And that is why I must ask you now : Can I count on you to make a donation to our campaign before our October 31 deadline ?

Medicare for All . College for All . Jobs for All . Justice For All . A government that works for ALL of us , and not just the wealthy few . If we 're in this together , that 's what we 'll get . Can you make a donation to help power your campaign ?

I am asking once again for your financial support . The short time ahead of us is enormously important for the future of our campaign , our movement , and our ideas . So if you can , please make a donation right now .

The wealthy and well - connected are scared that , under a Warren presidency , they would no longer have a government that caters to their every need . So they 're doing everything they can to try to stop Elizabeth and our grassroots movement from winning . In fact , billionaire Mike Bloomberg wants to use his personal fortune to buy his way to the Democratic nomination . Let 's defeat the billionaires and make our economy and our democracy work for working people . Chip in now to help power our movement .

The rich and powerful run Washington , and they 've written the rules so they do n't have to pay their fair share . I 'm proposing a new Ultra - Millionaire Tax on the wealthiest 0.1 % of Americans because it 's long past time to unrig the system and level the playing field . But we 'll only get big things like this done if this grassroots movement demands it . Sign our petition if you agree : It 's time to tax the wealth of the top 0.1 % .

The Democrats are the party of big , structural change . We need to give people a reason to show up and vote . We must build a grassroots movement across this country , person by person , dollar by dollar . I will not only beat Trump — I'll start making real change come 2021 . Chip in if you 're with me .

Figure 7: Annotated Sanders and Warren advertisements.

text-as-data tools, that salience must still be linked to the theoretical estimand through argumentation, but the advantage of this method is that the link between the empirical estimand (salient phrases) and the source data is more direct.

Conclusion

The application of text-as-data within computational social science is shaped by the methods and tools available. Where estimators have been based on various simplifying assumptions about language, researchers accordingly restrict their claims to what the model can justify. When representing a corpus as a matrix of word counts, the inferences that can be made from models built on this operationalization of text are based on observations about changes in the patterns of word counts.

Although part of the appeal of DL approaches driving their recent success in NLP is their high predictive accuracy, an equally important appeal of these tools is their modularity and flexibility. Contextual embedding models (of which `BERT` is one example) are a tool that allows us to model tokens within sequences, and make predictions or inferences between these levels. Although the flexibility of DL tools permits a new approach to language and a new set of estimators, estimands, and questions, this flexibility also introduces a complexity that has limited the utility of these tools in domains where explainability is key. This creates demand for research bridging this explainability gap, and showing applications of DL tools validated in the epistemology of the respective target domains.

I show that DL-based language models combined with feature attribution methods provide a way to describe the linguistic differences between political campaigns that can be justified in terms of empirical sufficiency. Although high predictive accuracy does not tell us *what* makes two campaigns differ, it does provide sufficient evidence that they are differentiable. Although the scores provided by IG cannot be provided in terms of marginal effects, by training new models that are alternately shown high-attribution or low-attribution regions, we can again infer that the scores do correspond to tokens that are informative to differentiating the campaigns. We can conduct statistical tests on these scores using uncertainty measures generated with MC Dropout, and make claims about the statistical significance of token importances.

Finally, because we are able to fit models with entire documents as a unit of analysis, we can construct instance-specific estimands for linguistic phenomena occurring only once in our corpus. This ability to superimpose corpus-level patterns into individual documents brings quantitative and qualitative approaches to text closer, allowing researchers to ask questions about specific phrases in individual documents. This approach is not limited to identifying differentiating in text either; as research moves in a multimodal direction, this work provides the theoretical foundations for applying the same saliency approaches to image classification models. Future work can also be aimed at constructing annotation tools to augment qualitative text analysis with the advances transfer learning and DL have brought to NLP.

References

Aggarwal, Charu C et al. 2018. “Neural networks and deep learning.”

Atanasova, Pepa, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. “A Diagnostic Study of Explainability Techniques for Text Classification.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–74. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.263.

Barabas, Jason, and Jennifer Jerit. 2009. “Estimating the Causal Effects of Media Coverage on Policy-Specific Knowledge.” *American Journal of Political Science* 53 (1): 73–89. http://www.jstor.org/stable/25193868.

Barnes, Lucy, and Timothy Hicks. 2018. “Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy.” *American Journal of Political Science* 62 (2): 340–54. https://doi.org/10.1111/ajps.12346.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent dirichlet allocation.” *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chong, Dennis, and James N Druckman. 2007. “Framing public opinion in competitive democracies.” *American Political Science Review* 101 (4): 637–55.

Clark, Tom S., and Benjamin E. Lauderdale. 2012. “The Genealogy of Law.” *Political Analysis* 20 (3): 329–50. http://www.jstor.org/stable/23260321.

Denny, Matthew J, and Arthur Spirling. 2018. “Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it.” *Political Analysis* 26 (2): 168–89.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. “ERASER: A Benchmark to Evaluate Rationalized NLP Models.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–58. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.408.

Diermeier, Daniel, Jean-FranÇOis Godbout, Bei Yu, and Stefan Kaufmann. 2012. “Language and Ideology in Congress.” *British Journal of Political Science* 42 (1): 31–55. http://www.jstor.org/stable/41485863.

Edelson, Laura, Tobias Lauinger, and Damon McCoy. 2020. “A security analysis of the Facebook ad library.” In *2020 IEEE Symposium on Security and Privacy (SP)*, 661–78. IEEE.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2020. “Keyword assisted topic models.” *arXiv Preprint arXiv:2004.05964*.

Ettinger, Allyson. 2020. “What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models.” *Transactions of the Association for Computational Linguistics* 8: 34–48.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. “Political advertising online and offline.” *American Political Science Review* 115 (1): 130–49.

Gal, Yarin, and Zoubin Ghahramani. 2016. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning.” In *international conference on machine learning*, 1050–59. PMLR.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. “Word embeddings quantify 100 years of gender and ethnic stereotypes.” *Proceedings of the National Academy of Sciences* 115 (16): E3635–44.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2019. “Measuring group differences in high-dimensional choices: method and application to congressional speech.” *Econometrica* 87 (4): 1307–40.

Grimmer, Justin. 2013. “Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation.” *American Journal of Political Science* 57 (3): 624–42. https://doi.org/10.1

111/ajps.12000.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Hayes, Danny, and Jennifer L. Lawless. 2015. "As Local News Goes, So Goes Citizen Engagement: Media, Knowledge, and Participation in US House Elections." *The Journal of Politics* 77 (2): 447–62. https://doi.org/10.1086/679749.

Hofman, Jake M, Amit Sharma, and Duncan J Watts. 2017. "Prediction and explanation in social systems." *Science* 355 (6324): 486–88.

Jain, Sarthak, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. "Learning to Faithfully Rationalize by Construction." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4459–73. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.409.

John, Peter, and Will Jennings. 2010. "Punctuations and Turning Points in British Politics: The Policy Agenda of the Queen's Speech, 1940–2005." *British Journal of Political Science* 40 (3): 561–86. https://doi.org/10.1017/S0007123409990068.

Lall, Ranjit, and Thomas Robinson. 2021. "The MIDAS touch: accurate and scalable missing-data imputation with deep learning." *Political Analysis*, 1–18.

Lapinski, John S. 2008. "Policy Substance and Performance in American Lawmaking, 1877–1994." *American Journal of Political Science* 52 (2): 235–51. https://doi.org/10.1111/j.1540-5907.2008.00310.x.

Lauderdale, Benjamin E, and Alexander Herzog. 2016. "Measuring political positions from legislative speech." *Political Analysis* 24 (3): 374–94.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97 (2): 311–31.

Lei, Tao, Regina Barzilay, and Tommi S. Jaakkola. 2016. "Rationalizing Neural Predictions." In *EMNLP*, 107–17. http://aclweb.org/anthology/D/D16/D16-1011.pdf.

Li, Zhe, Boqing Gong, and Tianbao Yang. 2016. "Improved dropout for shallow and deep learning." *Advances in Neural Information Processing Systems* 29.

Lin, Yongjie, Yi Chern Tan, and Robert Frank. 2019. "Open Sesame: Getting inside BERT's Linguistic Knowledge." In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 241–53. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4825.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23 (1). https://doi.org/10.3390/e23010018.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *Commun. ACM* 61 (10): 36–43. https://doi.org/10.1145/3233231.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A robustly optimized bert pretraining approach." *arXiv Preprint arXiv:1907.11692.*

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16 (4): 356–71. https://doi.org/10.1093/pan/mpn004.

Mathur, Arunesh, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. 2020. "Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle." https://electionemails2020.org.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Peterson, Andrew, and Arthur Spirling. 2018. "Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems." *Political Analysis* 26 (1): 120–28.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084.

Rheault, Ludovic, and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28 (1): 112–33.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis*, 1–25.

Rodriguez, Pedro L, Arthur Spirling, and Brandon M Stewart. 2020. "Embedding Regression: Models for Context-Specific Description and Inference in Political Science."

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A primer in bertology: What we know about how bert works." *Transactions of the Association for Computational Linguistics* 8: 842–66.

Rush, Alexander. 2018. "The Annotated Transformer." In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 52–60. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2509.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *ArXiv* abs/1910.01108.

Skansi, Sandro. 2018. *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. http://www.jstor.org/stable/25193842.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "Mpnet: Masked and permuted pre-training for language understanding." *Advances in Neural Information Processing Systems* 33: 16857–67.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15 (1): 1929–58.

Stanig, Piero. 2015. "Regulation of Speech and Media Coverage of Corruption: An Empirical Analysis of the Mexican Press." *American Journal of Political Science* 59 (1): 175–93. https://doi.org/10.1111/ajps.12110.

Storks, Shane, and Joyce Chai. 2021. "Beyond the Tip of the Iceberg: Assessing Coherence of Text Classifiers." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3169–77. Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.272.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic attribution for deep networks." In *International conference on machine learning*, 3319–28. PMLR.

Torres, Michelle, and Francisco Cantú. 2022. "Learning to see: Convolutional neural networks for the analysis of social science data." *Political Analysis* 30 (1): 113–31.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." *Advances in Neural Information Processing Systems* 33: 5776–88.

Wiegreffe, Sarah, and Yuval Pinter. 2019. "Attention is not not Explanation." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing (EMNLP-IJCNLP), 11–20. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002.

Yu, Mo, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. "Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4094–4103. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1420.