Original Manuscript

Social Science Computer Review
2024, Vol. 0(0) 1–25


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393241247420
journals.sagepub.com/home/ssc


Covering the Campaign: Computational Tools for Measuring Differences in Candidate and Party News Coverage With Application to an Emerging Democracy

Aaron Erlich[1], Danielle F. Jung[2], and James D. Long[3]

Abstract

How does media coverage of electoral campaigns distinguish parties and candidates in emerging democracies? To answer, we present a multi-step procedure that we apply in South Africa. First, we develop a theoretically informed classification of election coverage as either "narrow" or "broad" from within the entire corpus of news coverage during an electoral campaign. Second, to deploy our classification scheme, we use a supervised machine learning approach to classify news as "broad," "narrow," or "not election-related." Finally, we combine our supervised classification with a topic modeling algorithm (BERTTopic) that is based on Bidirectional Encoder Representations from Transformers (BERT), in addition to other statistical and machine learning methods. The combination of our classification scheme, BERTTopic, and associated methods allows us to identify the main election-related themes among broad and narrow election-related coverage, and how different candidates and parties are associated with these themes. We provide an in-depth discussion of our method for interested users in the social sciences. We then apply our proposed techniques on text from nearly 100,000 news articles during South Africa's 2014 campaign and test our empirical predictions about candidate and party coverage of corruption, the economy, health, public infrastructure, and security. The application of our method highlights a nuanced campaign environment in South Africa; candidates and parties frequently receive distinct and substantive coverage on key campaign themes.

Keywords

text-as-data, machine learning, South Africa, news coverage, electoral campaigns

[1]McGill University, Canada
[2]Emory University, USA
[3]University of Washington, USA

Corresponding Author:
Aaron Erlich, Department Political Science, McGill University, Leacock Building, Room 414, Montréal, QC H3A 2T7, Canada.
Email: aaron.erlich@mcgill.ca

Introduction

Scholarship on democratic politics portrays contrasting perspectives regarding whether media coverage of campaigns informs an understanding of a country's election environment. One perspective argues that information revealed in campaign coverage serves a critical role educating and persuading voters (Price & Zaller, 1993; Zaller, 1992); if true, such coverage may shape their attitudes about and choices among candidates (Gerber et al., 2009; Popkin, 1994). A contrasting viewpoint suggests that campaigns do little more than reinforce political fundamentals (Lewis-Beck & Rice, 1992), indicating coverage would at best "remind" voters of the limited choices they face, implying little to no role for persuasion (Horowitz, 1985). The tension between these perspectives is particularly salient in emerging democracies, where on the one hand campaign coverage could possibly help to address voters' informational deficiencies given the newness of democratic procedures, while on the other media often appear as simply epiphenomenal to underlying institutional realities of unconsolidated democracies.

Text-as-data techniques provide a new entry point into this debate by investigating the extent to which media coverage of campaigns meaningfully differentiates parties and candidates in substantive ways, allowing researchers to characterize more fully the information environment that voters confront. Previous studies have tackled related concerns and made advances in measuring the valence of political and economic news coverage and changes over time (Soroka, 2014), media bias (Watanabe, 2017), and personalized content of specific candidates (e.g., Hall & Lim, 2018; Vliegenthart et al., 2011) or speech (Müller, 2020). Despite qualitative analyses (e.g., McCombs & Shaw, 1972) and hand-coding of stories (Hayes & Lawless, 2018), developing automated ways to measure how the news media differentiates candidates and parties during election campaigns has remained a challenge. Automated applications have the potential to deepen our understanding of political campaigns, particularly in emerging democracies, and potentially in real time.

To address these lacunae, we first develop and present a procedure for identifying election-related news coverage from within the entire corpus of news coverage during an electoral campaign, irrespective of identifying a specific set of political actors. Second, we show how to compare candidate and party coverage from within that election-related corpus. To do so, we develop a theoretical distinction between "broad" and "narrow" election coverage and then employ new text-as-data techniques to classify the universe of election-related stories into either of these categories, which are distinguished from "not election-related" stories. Subsequently, we combine our predictions of election-related stories with BERTTopic, a topic modeling technique, which leverages word embeddings from Bidirectional Encoder Representations from Transformers (BERT), to estimate the extent to which candidates and parties are able to differentiate themselves from each other on election-related topics during an election cycle.

Applying this architecture to a news corpus from South, further informed by our knowledge of the case, our procedure helps ascertain what information is salient in election-related news coverage and what news coverage might *actually* relay to voters and what this implies about electoral competition. We identify common "election-related themes" that appear across different coverage and article aggregation schemes, and show how each party/candidate relates to them using contextual word embeddings. Our findings portray campaign coverage that not only signals political fundamentals, but in many cases provides a more nuanced set of coverage that makes important substantive distinctions between relevant political actors.

Election-Related News Coverage Classification Approach

Our approach derives from two theoretically grounded classifications of "election-related" coverage, which we define as "narrow" and "broad," and distinguish from "not election-related"

stories. At a minimum, an election is the process by which voters cast ballots to select aspirants for office; we denote "narrow" media coverage of an "election" as stories specifically mentioning electoral actors or processes (e.g., candidate/party names, voting procedures, and results declarations) which convey *only* basic information about them. Examples might include the announcement by the election commission of voter registration drives, published polls showing the horse-race, or the certification of final results.

Less obviously but fundamental to understanding the totality of election coverage, a "broad" definition constitutes *any* information about political actors, procedures, or institutions *regardless* of whether stories explicitly mention the election. This broad classification is based on the idea that general political content may gain electoral salience if covered during the campaign period because such coverage plausibly links distributional and policy outcomes with government performance or partisan platforms. Examples might include reporting on crime and policing, which does not directly portray election-related content because the police do not contest election, but may gain importance during the campaign if citizens attribute management of the police to elected officials.

Narrow and broad classifications together compose the totality of "election-related" labeling, which we differentiate from "not election-related" news (stories that do not mention politics at all, such as a wedding announcement). Narrow and broad classifications vary depending on whether media coverage of politics during the campaign includes only coverage focused on the election or also the totality of any political coverage. Therefore, what comprises election-related news quickly complicates any obvious portrayals of the election-related news coverage since coverage of politics does not always easily, clearly, or consistently delineate election-related material. Our argument is not that electoral media definitively contain either exclusively narrow or broad content; rather, simply changing the scope of inquiry around election coverage potentially leads to different substantive representations of the election-related news coverage. As a result, it is important to examine different conceptions of electoral coverage.

We apply this method to examine the coverage of political actors and events *within* narrow and broad categories to assess how variation in these labels reveals differences in coverage of electoral dynamics. Regardless of narrow or broad, if election-related news coverage provides meaningful information, campaign content should cluster around certain election-related themes and issues; that is, information "sets" should emerge following institutional and situational factors associated with political actors and government agencies. For example, parties associated with different economic platforms might enjoy more coverage on the state of the economy or economic policies; parties making issue-specific appeals (e.g., regarding health or security) should have their names associated with attendant policy promises and campaign issues; incumbent parties' content might have more coverage around government performance themes like corruption, the economy, or service delivery; and opposition parties might be associated with messaging on government failures and alternative policy directions.

From this, we propose that for any corpus of news coverage during an election period, stories can be classified into one of three types: *narrow, broad,* or *not election-related*. Hand-coders can code a random sample of stories into one of these three categories. Because many newspaper corpora will be too large to code all stories by hand, we advance supervised machine learning to classify all stories not hand-coded, a technique which has been used to find "politically-related" news from within all news coverage (Budak et al., 2016). Since the performance of machine learning classifiers will likely vary by application, we recommend evaluating a range of classifiers and the inclusion of non-textual covariates. As we discuss in the Results section, Ridge Classifier has the best performance in our application. Moreover, we suggest exploring models with only textual covariates and ones with additional features, including days to the election, article language, publication, and publication owner. We propose using accuracy and F1-scores to choose the best model. In our application, we find the model with textual covariates performs best.

Using Topic Modeling to Capture Election-Related Themes

Regardless of whether the coverage being analyzed is broad or narrow, each election has a distinct set of politically relevant campaign themes or issues,[1] which will vary by context and time period (Budge & Farlie, 1983). Historically, most scholarship has pre-coded election-related themes. For example, several articles on media coverage of U.S. politics focus on an eight-themed model of campaigns (Hayes, 2010), based on Petrocik (1996). Druckman (2004) suggests 28 campaign-related issues. Hall and Lim (2018) suggest six themes of candidate-related news coverage.

More recently, scholars have begun to use topic modeling to determine politically relevant media themes. For instance, Budak et al. (2016) code 15 themes, which were manually created with the aid of an LDA algorithm. We build on this approach. Following Rodriguez et al. (2023), we take an agnostic approach regarding what themes might emerge by not beginning with a specific set of words or election-related themes.[2] Therefore, based on our categorization of election-related stories as broad or narrow, we can then apply unsupervised topic modeling to help reveal election-related topics.

Before discussing the technical specifics of our procedure, we briefly highlight the benefits of topic modeling to studying the themes of campaign coverage. Topic modeling involves the process of finding topics that best represent the information contained in a corpus that is divided into different documents. Each topic is represented as the set of weighted words most informative of its unique topic. The weight of each word is determined by its similarity with the topic.

A traditional method for topic modeling is the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which represents each document as a combination of topics and each topic as a distribution of words. Despite its popularity and utility, the LDA has two main drawbacks. First, it requires the topic space be discretized into N topics beforehand, but the mis-specification of the number of topics N can produce uninformative results and the real number of topics is rarely known (Syed & Spruit, 2017). The second limitation relates to the representation of documents, which is generally done using bag-of-words (BOW) and ignores semantics. In other words, similar words in the semantic space (e.g., professor and teacher) are treated as different terms despite similarities.

This latter shortcoming can be addressed using word embeddings, such as *word2vec* Mikolov et al. (2013) and *GloVe* (Pennington et al., 2014). Word embeddings capture both the semantic and syntactic information underlying words. In this approach, each word is represented by a multi-dimensional vector, where each entry represents information about the word's meaning. Word embeddings later expanded to include documents in the *doc2vec* model (Le & Mikolov, 2014). This methodology is capable of learning document and word vectors that are jointly embedded in the same space, which improves the quality of the learned vectors (Lau & Baldwin, 2016). Semantically similar words and documents appear close to each other in the embedding space. As a consequence, the most similar word vectors to a document are likely good representations of the document's topic.

The *doc2vec* model can also be used in the context of topic modeling (Angelov, 2020). However, one limitation is that the embeddings are often pre-trained. In other words, they are derived without considering the relevant context. The word "virus" in "respiratory virus" and "computer virus" would be represented by the same vector encoding despite their different meanings. Training contextual representations on text corpus helps to overcome this hurdle and was first achieved by training a deep bidirectional language model, called ELMo (Embeddings from Language Models) (Peters et al., 2018) on a large corpus, where each word is assigned to an embedding that is a function of the entire input sentence and not just the specific word. ELMo is expanded by the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which pre-trains deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contextual words in all of the neural nets' layers. As a consequence,

this BERT can be easily fine-tuned for a wide range of tasks and results obtained from BERT outperformed previous state-of-the-art models in several tasks (Liu et al., 2019; Sanh et al., 2019; Wang et al., 2019).

Proposed Topic-Modeling Approach

On the subset of the data that has been machine classified as election-related (either broad or narrow) from the corpus of all news coverage, we use BERTTopic (Grootendorst, 2022), a multi-step algorithm based on the BERT embedding model, to perform topic modeling. Due to impressive recent results of contextual NLP models, we employ as our embedding model *Sentence-BERT* (Reimers & Gurevych, 2019), a modification of the pre-trained BERT network to create sentence embeddings that can be compared using cosine similarity to find other sentences with a similar meaning. This method transforms sentences or documents in a numerical representation with a semantically meaningful relationship. Each word is tokenized and mapped according to its context.[3,4]

BERTTopic is a processing pipeline algorithm, which aims to use BERT contextual embeddings together with other steps such as dimensionality reduction and similarity methods, to generate relevant topics from a large corpus. Using BERTTopic to measure how the media distinguishes parties and candidates in its coverage requires making choices related to BERTopic's intermediate steps, as well as other applied research problems. We provide guidance for researchers on the following five issues related to using the BERTTopic algorithm: (1) dimensionality reduction, (2) cluster recognition, (3) converting clusters to topics, (4) topic selection, and (5) hyperparameter selection.

Using BERTTopic alone, however, is not sufficient for our needs. We therefore suggest additional methods to: (1) convert topics to themes; (2) measure topic similarity to election-related themes in news coverage; and (3) estimate uncertainty. We address these eight issues in turn.

BERTTopic Issue 1: Dimensionality Reduction. After converting all articles into document embeddings, the BERTTopic algorithm requires that documents be grouped based on their similarity. In other words, we want to derive topics (clusters) based on the documents' content. However, due to the "curse of dimensionality," this task is often non-trivial. The sparsity of document vectors makes it challenging to find dense clusters, resulting in high computational costs (Marimont & Shapiro, 1979).

One solution to overcome this limitation is to use an algorithm to reduce the dimension of embedded documents before clustering them (Angelov, 2020). In this reduced space, clusters can be found more effectively. Out of the dimensionality reduction algorithms, two of the most applicable are the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes et al., 2018) and T-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton, 2008). According to Angelov (2020), t-SNE does not preserve global structure as well as UMAP; moreover, it does not scale well to large datasets. Thus, we use UMAP for reducing the dimensionality while preserving most of the global and local structure of the embeddings.

BERTTopic Issue 2: Cluster Recognition. With the new compressed semantic space, the task of finding similar clusters can be carried out by using a density-based algorithm such as DBSCAN (Ester et al., 1996) and their variants (Campello et al., 2013; McInnes & Healy, 2017). This class of methods' main advantage is that it does not enforce a cluster to every datapoint. In other words, documents that have no clear underlying topic are treated as noise and do not interfere in clusters of similar documents. For this step of the BERTTopic algorithm, we find the dense areas of the

reduced space using HDBSCAN (Campello et al., 2013), an extension of DBSCAN that also deals with variable density clusters and requires less parameter tuning.

BERTTopic Issue 3: Cluster to Topic Conversion. Once clusters have been identified, the researcher needs to transform them into topics. Achieving topics requires finding the most representative words within each cluster. The idea of finding the most relevant words in each document is the foundation of the term frequency–inverse document frequency (TF-IDF) method (Jones, 1972). The purpose of TF-IDF is to increase the value of a feature based on the frequency it appears in a document and based on the inverse document frequency of the same word across all documents. In other words, TF-IDF compares the importance of words between documents.

Since our goal is to determine the most important words for each topic, we employ a modified version of the TF-IDF equation to deal with clusters by grouping all documents within a topic. In this case, the description of a topic is defined by the more relevant words within that cluster. The score of a given word x in the topic c is defined as

$$W_{x,c} = tf_{x,c} * log\left(1 + \frac{A}{f_x}\right) \tag{1}$$

where $tf_{x,c}$ represents the frequency of word x in topic c, f_x refers to the frequency of word x across all topics, and A denotes the average number of words per class (topics in our case).[5]

BERTTopic Issue 4: Topic Selection. The output of the c-TF-IDF is a set of relevant words describing a collection of documents, in this case, a topic. Nevertheless, this collection of words is not guaranteed to describe a *coherent* topic. In some cases, variations of the same word or idea can end up in the topic representation. For example, imagine the top terms for a topic are: "good government," "great government," "excellent government," "innovation," "research." The first three terms are similar and define the same characteristic. Therefore, we would like to improve the diversity and coherence of words, avoiding the overlap between the words themselves.

We achieve this aim within BERTTopic by using Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998), a diversity-based ranking technique that aims to maximize the relevance and novelty of the retrieved items. In our application, the MMR can be used to improve the relevance of keywords that define a topic (Bennani-Smires et al., 2018). This metric is defined as

$$\text{MMR} = \underset{C_i \in C/K}{argmax}\left[\lambda\left(Sim(C_i, top_{emb}) - (1-\lambda)\ \underset{C_j \in K}{max} Sim\left(C_i, C_j\right)\right)\right] \tag{2}$$

where C is the set of candidate keywords, K is the ranked list of the original extracted keywords, top_{emb} is the topic embedding, and C_i and C_j are the embeddings of candidate keywords i and j, respectively. Sim is a similarity function. We employed the cosine similarity defined in equation (3). Finally, λ is the parameter that controls how diverse the keywords are. When λ is close to zero, the generated keywords are more diverse, whereas $\lambda = 1$ generates the highest relevance keywords.

In the topic modeling context, higher MMR values represent terms that are both relevant to the topic and contains minimal similarity to the other top-ranked words. We employed $\lambda = 0.1$ as greater values than this resulted in creating topics with keywords that were irrelevant to that topic.

BERTTopic Issue 5: Hyperparameter Selection. The processing steps in BERTTopic have several parameters that must be set. For the dimension reduction, the main parameters are (1) the number of neighbors, which controls the balance between preserving global structure vs. local structure in

the low dimensional embedding, (2) the target dimension of the reduced embedding space and the method for calculating the distance between embeddings. Similarly, we also need to set parameters for the clustering step. The three main parameters include the (1) minimum cluster size (minimum number of points to consider a cluster), (2) the minimum number of samples in a neighborhood for a point to be considered a core point, and (3) the metric to be used to calculate the distance between points. Finally, we also need to define some topic model hyperparameters, such as (1) the number of words that are used to define a topic, (2) the minimum number of documents in a topic, and (3) the n-gram range.

We experimented with different combinations of hyperparameters for all the steps in the topic creation. We present the values for each parameter in the Table 1, as well as a range of common values or methods that are often used in this type of problem.

Interpretation Issue 1: Converting Topics to Themes. The number of topics in our approach is defined according to how different documents cluster in the embedding space. The limitation of using a predefined number of topics before clustering is that the resulting topics could be noisy, as topics that are unrelated could be merged together if the number of topics is set too small, or there could be topics that do not have the same "theme" if the number of topics is set too large. We chose 50 topics in our application.

We recommend a two-step process to move from topics in the topic model to the primary election-related themes. First, use a dimensionality reduction algorithm to cluster similar topics in

Table 1. Hyperparameters in the Proposed Method and Suggested Values.

Processing step	Parameter name and its value	Common range of values/methods
Dimension reduction	Number of neighbors = 15	2 to 100. The choice of this parameter depends on the desired balance between local versus global structure
Dimension reduction	Target embedding dimension = 5	2 to 100. Low values tend to compress more the embeddings and could result in some information loss, and high values tend to be more expensive to calculate. The target embedding dimension will also depend on the original embedding dimension
Clustering recognition	Minimum cluster size = 20	10 to 50. This parameters is very reliant on the characteristics of the dataset. Knowledge on the structure of the documents could help defining the correct minimum cluster size
Clustering recognition	Distance metric = "euclidean"	Any distance metric could work, the more efficient/common are the Euclidean, the cosine similarity and the Manhattan distance
Clustering recognition	Minimum samples = 20	10 to 50. This parameter should have a value close to the minimum cluster size
Cluster to topic conversion	Number of words that define a topic = 10	5 to 15. This is parameters is more important for understanding the topics generated, usually using 10 words per topic yield relevant yet distinct topics
Cluster to topic conversion	Minimum topic size = 10	10 to 50. Similar to the minimum cluster size, knowledge on the corpus could improve the results
Cluster to topic conversion	n-gram range = 2	1 to 3. Bi-grams and tri-grams could provide additional information when compared to unigram, however the computational cost scales very rapidly with increasing n-grams

a two-dimensional space. In our application, we employed UMAP for a second time due to its advantages.

Once the topics are represented as points in the space, the researcher can use their contextual knowledge to define the main themes. For instance, if several topics are closely related to a specific subject, it may suggest a potential election theme. We identified such themes based on our knowledge of the context. However, an alternative approach involves using clustering metrics to define these topics, which could be explored in future research. It is worth noting that not all clusters may pertain to elections, underscoring the significance of contextual knowledge, even when clusters are determined using a metric threshold.

Interpretation Issue 2: Similarity Between Topics and Election-Related Parties and Candidates. Once the themes of election-related news coverage have been defined, the researcher can use them for a variety of purposes. Our goal is to measure the differences between parties' and candidates' associations with these themes among the broad and narrow election-related coverage. If there are differences in associations, then there is reason to believe that the media more strongly associates a particular party with a particular theme. If there are no differences, then we do not have strong reason to believe that any party is distinguishing itself on a particular election-related theme.

To estimate the association of each party with a theme (among the narrow subset, the broad subset and all-election-related coverage), we follow Rodman (2020) and employ cosine similarity between a target term and a topic, defined as

$$\underset{sim}{cos} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

where A and B represent embedding vectors.

We first detect all of the mentions of our target terms (political parties and candidates) in our corpus of text. There are different ways to estimate the similarity between two features of a corpus. For example, if we want to estimate how "Jacob Zuma" is related to the theme of "corruption," we could calculate the similarity between "corruption" and "Jacob Zuma." While this approach is intuitive, it does not leverage all contextual embedding capability. Instead, we use a different approach.

For each target term, we propose selecting a six-word symmetric context window to calculate its context embedding. In other words, we selected twelve words that surround each detected term in the corpus to estimate the context in which they are presented. Then, for each election theme, we generate its embedding by using the topic embedding most related to that theme. Finally, we calculate the similarity between the target term contextual embedding and each election-related theme.

The result of this technique is that the embeddings of each target word are generated according to the context in which they occur on the corpus. For instance, if the name of a candidate often appears in the articles close to words like dishonesty, scandal, and misconduct in narrow subset, this target embedding would be closer (more similar) to the embeddings of the corruption theme.

Interpretation Issue 3: Uncertainty Estimation. Once we have estimated the cosine similarity for different parties and candidates, it is possible to estimate the differences. However, these differences measure neither uncertainty nor how likely it is the differences between candidates/parties and topics may have occurred by chance.

To account for uncertainty, we employ a similar approach to Rodriguez et al. (2023). Specifically, we take 100 different realizations using 10% of the available data for each target term separately for our two subsets and for all election-related coverage. Then we calculate the similarity between these terms and the main-election themes. As a result, we have 100 measurements of similarity for each party/candidate and theme combination.

With these bootstrapped values, we can estimate the distribution of the target word/theme similarity using the procedure described previously and determine whether they are significantly different by calculating their confidence intervals. We present these distributions with the 5% confidence interval bounds and mean for all results we present. This method allows us to evaluate our predictions or hypotheses with typical metrics of statistical significance. Table 2 presents an overview of the steps of our multi-step procedure.

South Africa 2014

We apply our method to a novel corpus of electoral news coverage around the 2014 South African election. While our approach described above is generalizable, any analysis requires appropriate contextualization. South Africa's media and institutional environments generate national campaigns; therefore, studying election-related news reveals important dynamics about the campaign as a whole.

South Africa's print media provided significant reporting on the horse-race, political rallies, corruption scandals, strikes, and the state of public services. The government does not own or limit the publication of stories by the main print media houses,[6] which are widely perceived by South Africans as unbiased (Schreiner & Mattes, 2011). In 2014, the country enjoyed a 95% literacy rate (93% among women) and a history of accessible, widely consumed national and local newspapers.[7]

The media environment in South Africa could have simply reiterated well-worn and static statements regarding the fundamentals, including partisan politics. Since the end of apartheid and transitional elections in 1994, the African National Congress (ANC) has maintained dominance

Table 2. Research Steps.

Step	Description
Hand-coding	Handcode a sample of news coverage into "broad," "narrow," and "not election related."
Supervised learning	Classify all stories in news coverage corpus based on hand-coded subset.
Topic modeling with BERTTopic	Run BERTTopic Algorithm on the news coverage labeled as election-related (pooled) and the "broad" and "narrow" subsets separately using proposed methods and hyperparameters for dimensionality reduction, cluster recognition, cluster to topic conversion, and topic selection
Election-related theme identification	Reduce and combine the topics from the preferred topic model from BERTTopic into election-related themes using UMAP and visual inspection of 2D plots
Similarity between topics and themes calculation	Identify parties and candidates in election-related coverage and calculate the cosine similarity between these political parties and candidates and the identified election-related themes
Uncertainty estimation between topics and themes	Estimate the uncertainty of the cosine similarity measures using bootstrapping

and won a majority of legislative seats in every election. The leader of the majority party forms the government (for whom the title "president" is functionally equivalent to prime minister), and politicians therefore lean on party mobilization, and coverage of party activities, to organize support (Lodge, 2004), including through media.

The ANC's campaign dynamics revolve around turning out their base through appeals to predominantly black voters (Ferree, 2011; Southall, 2014), about 80% of the electorate. In 2014, the ANC ran its controversial incumbent president Jacob Zuma. While few observers thought the ANC would lose its parliamentary majority, *any* seat loss would be viewed as a referendum on government performance. Nelson Mandela's death a few months prior encouraged party leaders to frame many appeals as "reminding" black voters of the ANC's role in fighting apartheid. Party stalwarts argued that stressing this legacy was vital because they feared "born free" voters—adults (18+) born post-apartheid—would take the ANC for granted and not turn out in 2014, or swing to the opposition. For their part, opposition parties have had difficulty winning seats beyond pockets of regional or demographic support. The main 2014 challenger was the Democratic Alliance (DA), whose base primarily consists of white and "coloured" (mixed race) voters and was led by Helen Zille, the (white) former Cape Town mayor. Another important competitor was Julius Malema, former ANC youth leader turned founder of the Economic Freedom Fighters (EFF), a new party contesting for the first time, whose base primarily consists of young, urban, and left-wing black voters.

Electoral advantages enjoyed by the ANC do not mean campaigns lack policy substance, and the 2014 coverage provided a rich environment of themes. Opposition parties use the ANC's institutional advantages against it, hoping to increase the rate of black defection by mentioning ANC failures. Both Zille and Malema put Zuma's scandals at the front of their campaigns to leverage the fact that voters increasingly perceived him as corrupt given a long history of credible misdeeds; his use of state funds to build an opulent rural home (Nkandla) was a frequent point of attack in 2014. The DA launched a ten-point platform; the first two items focused on rooting out government corruption followed by a promise to create six million new jobs.

The parties attempted different messaging on the economy. The ANC campaigned on past performance on economic growth and further promises to address poverty alleviation. The DA employed explicit messages geared toward a rising black middle class and urban population that promised policy improvements regarding income, growth, and employment opportunities. Because the DA's messaging was less geared toward marginalized South Africans, that provided a space for the EFF to more explicitly gain lower-class and youth black support with appeals regarding redistribution, expanding workers' rights, and fighting inequality. Malema may have differentiated himself by calling for uncompensated expropriation of wealthy South Africans' land and nationalization of mines and banks (Mbete, 2015).

The quality of public services is persistently a salient campaign theme, and widespread anti-government protests—often led by ANC supporters—occur in response to the government's poor record of service delivery (Alexander, 2010). Healthcare is one such issue, given the country's massive racial and regional inequities in health access, coupled with the repercussions of AIDS and other public health crises. At the time of the 2014 election, Zille had been a prominent opposition politician with a history of criticism of health services, suggesting she may have differentiated herself on that issue. Further and throughout the election cycle, issues of the power grid were constantly in the news headlines, as were the poor quality of roads, suggesting that opposition parties could have benefited from coverage on the theme of public infrastructure. Given the violence of the apartheid regime, security has been a perennial issue for all South Africans regardless of race. Although the crime rate had initially declined after apartheid, it began to rise after 2011. Both opposition candidates and parties railed against the ANC and Zuma for failing to address crime adequately.

The 2014 results revealed the tension between institutional stability and opposition gains: the ANC won 62% of the vote (249 seats), but lost 15 seats (4% of voteshare) from 2009; the DA outperformed expectations, gaining 22% (89 seats, a pick-up of 18), and the EFF achieved 6% (25 seats).[8]

Predictions

South Africa's institutional and campaign dynamics are likely to be observed in the media coverage in several ways. Specifically, given our tracking of the campaign, we outline some predictions of how stories related to corruption, the economy, and government services might have been associated with parties and presidential aspirants. Additionally, our method will also likely pick up differentiation in the media that might not have been predicted by researchers ahead of time. We note that our retrospective approach provides a good opportunity to benchmark our method and contextualize findings: the election and its outcome are known, which allows for validation of the method drawing on expertise in the field.

Corruption

Prediction 1A Given the allegations against him, coverage of Zuma will be distinct (and of elevated quantity) from other party leaders (Zille and Malema) with respect to corruption topics.

Prediction 1B Party-level patterns will follow the same trends as Prediction 1A; the ANC will be distinct from opposition parties, but opposition parties (DA and EFF) will not be distinct from each other.

Because these predictions are plausibly the most intuitive, finding corruption-related results should provide a face validity test of our method.

Economy

Prediction 2A Given all of the parties focus on a variety of economic issues, parties will be less likely to distinguish themselves on this theme.

Our reasoning with respect to the economy is that although the policy platforms of the ANC, DA, and EFF stressed different priorities, it would be harder to differentiate themselves in coverage because these priorities all generally fall under a large umbrella of the economy-related theme.

Public Services: Health, Infrastructure, and Security

Prediction 3A Of the two opposition party leaders, Zille is more likely to distinguish herself on health.

Prediction 3B Given the parties' focus on health and infrastructure in their platforms, both opposition parties are likely to distinguish themselves from the ANC on these issues.

In light of her longer history as a governing member of the DA (as Cape Town mayor), Zille had a track record working on infrastructure and health issues and made them a focus of her campaign; in contrast, the EFF focused on economic and youth issues disproportionately. Both the DA and EFF criticized the ANC regarding services and infrastructure.

Prediction 4A Given their focus on security, opposition leaders (Malema and Zille) should be indistinguishable from each other but distinct from Zuma.

Prediction 4B At the party level, we also predict opposition parties (EFF and DA) should not be distinct from each other but distinct from the ANC.

The rising crime rate was a vulnerability for Zuma and the ANC as incumbents; the opposition candidates and parties ran on this as a failure of government performance, although they did not appear distinct from each other in this regard.

Application

Our newspaper corpus consists of 97,428 articles from 167 South African newspapers for the period from 53 days before the 2014 election (March 15, 2014) to 23 days after (May 30, 2014). We constructed the corpus by scraping the websites of South African daily or weekly print newspapers—both national and local—that publish online, along with other online-only sources (e.g., News24). We systematically reviewed every print newspaper mentioned by the South African Audience Research Foundation (SAARF) and Wikipedia in South Africa and downloaded all available stories without regard to article content. Among the 37 print publications with the highest circulation in 2014, we scraped stories from 23.[9] Because the corpus is multilingual, we first used Microsoft Azure's Translator Text API to translate all isiZulu and Afrikaans stories into English.

To classify articles' election-relatedness, with 80% agreement, two coders labeled a random subset of 1000 articles (approximately 1%) into three categories: election-related-"narrow" (11%), election-related-"broad" (14%), and "not election-related" (74%) (see Appendix B.1 on coding).[10] Political actors or institutions specifically related to or mentioning the election or election day were coded as narrow (e.g., Jacob Zuma, Democratic Alliance, voter registration). The mention of any political actor or institution was coded as broad regardless of whether the story was specifically related to the election (e.g., health ministry, police, prosecutors). Broad therefore subsumes the same coverage as narrow but also includes reference to government agencies not necessarily directly related to the election. Appendix B.1 provides examples of typical stories of broad and narrow categories.[11]

Before engaging in either classification or topic modeling, we preprocessed our data in a pipeline composed of two stages. In the first stage, numbers and dates were tagged, and we performed named entity recognition—that is, we combined all recognized named entities into single tokens (e.g.,"Jacob Zuma" was treated as a single token). We also de-duplicated important South African individuals, organizations, and places using pattern matching and converted them to a canonical form (e.g., "ANC" and "African National Congress" are standardized to "African National Congress"). In the second stage, typical bag-of-word preprocessing techniques were employed, stop-words were removed, and the text was lemmatized.

For the supervised learning component, we used both stages of preprocessing *before* running classification models. We compared the performance of logistic regression, random forest, xgboost, and Ridge classifiers, and implemented these models with only textual covariates and with additional features that include days to the election, article language, publication, and publication owner, using a 10-fold cross-validation stratified by outcome category. The Ridge classifier using the textual information and extra features achieved the best performance with an accuracy of 87% and F1-score of 0.74.[12]

Since we labeled only a subset of articles, we predict the labels for the remaining unlabeled articles using Python's scikit-learn package (Pedregosa et al., 2011). In subsequent analyses, we used the label if it was hand-coded and the predicted label otherwise.[13]

For the topic modeling component, we used only the first stage of the preprocessing pipeline. It is important to note that traditional topic models often benefit from corpus preprocessing

techniques such as lemmatization, named-entity and part-of-speech tagging, and dependency parsing (Denny & Spirling, 2018). Because BERT is based on contextual embeddings, the words surrounding each word are helpful to create the topic embedding. As a result, preprocessing techniques can hinder the quality of the topics. Therefore, we use minimal preprocessing.

Next, we applied the procedure modified from BERTTopic introduced above. Specifically, we ran the BERTTopic algorithm on six versions of the data based on document definition (2) and subset (3). For document definition, we use article as a document or five-sentence (paragraph) as a document, and we also use three subsets: the broad only subset the narrow only subset and the broad and narrow subsets combined.

Across all of the topics in the six models, by visually inspecting the two-dimensional plots and applying our contextual knowledge, we identified eight common election-related themes (discussed below) by looking at all of the resulting six models' topics. These themes appeared in at least four of the six models trained. The model trained on pooled broad and narrow coverage using five sentences as a document contained all of the election-related topics we identified.[14] Therefore, we chose this model to perform all of the subsequent analysis; however, we acknowledge these themes will vary across contexts, elections, and the number of models (e.g., sentence vs. document aggregation) researchers wish to evaluate.

We then labeled the relevant topics in our chosen model from BERTopic with one of the eight themes we identified, presenting five of those here: (1) corruption, (2) economy, (3) health, (4) public infrastructure, and (5) security; with the three remaining themes in Appendix D: (6) education, (7) housing, and (8) voting. We note that not all topics that BERTTopic detects will have an election-related theme, and these topics will not be analyzed. All election-related themes and the corresponding topics are found in Appendix Table F.6. We then calculated the cosine similarity between the three main parties (ANC, DA, EFF) and candidates (Zuma, Zille, Malema), and the eight election-related themes on the pooled data and on narrow and broad subsets separately. Finally, we estimate uncertainty with the technique described above for all estimates of cosine similarity.

Results

Before showing our core results, we first use our data to describe how election coverage varies as election day approaches, with the overall proportion of election-related articles in broad and narrow categories over time. While voting occurred on May 7, 2014, the concept of the "2014 South African election" could be interpreted as referring to a more extensive time period. Including both the broad and narrow categories accounted for on average 0.44 of articles, with a range between 0.31 and 0.67, and the maximum fraction occurring on election day itself. The bottom panel of Figure 1 displays this change in coverage, which is primarily due to the change in narrow (but not additional broad). Including narrow categories accounted for on average 0.54 of all election-related articles, with a range between 0.41 and 0.77, and the maximum narrow fraction occurring on election day. Figure 2 examines the prevalence of the mentions of the largest political parties and their leaders (Zuma/ANC, Zille/DA, Malema/EFF). Unsurprisingly, Zuma and the ANC receive the most coverage. These results show that politically relevant coverage is approximately constant, but candidate and party-focused content increased significantly two weeks before the election. Highly engaged voters would have received information in the months before the election, but closer to election day *any* potential voter consuming news was reasonably exposed to candidate/party coverage.

We now turn to the main application of our technique by examining the relationship between parties and candidates and the main-election topics. We do so under pooled coverage, and then separate broad and narrow—specifically to examine patterns or cosine similarity both between

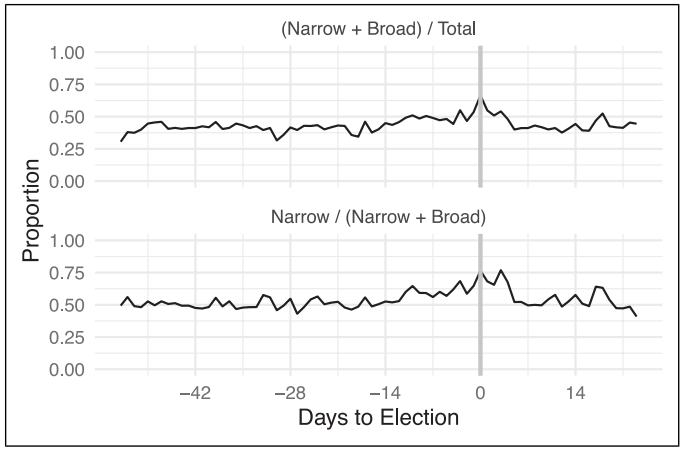

Figure 1. Proportion of articles in the Narrow, Broad, and Non-election categories by day.

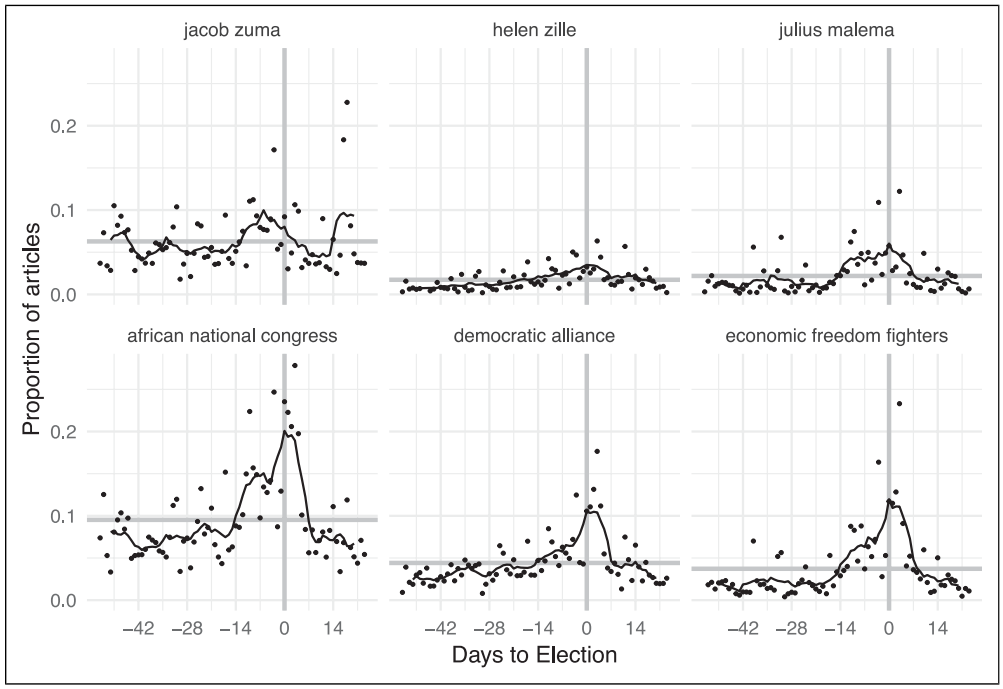

Figure 2. Proportion of articles containing at least one mention of the three largest parties and their candidates ("African National Congress" and Jacob Zuma,"Democratic Alliance" and Helen Zille, and "Economic Freedom Fighters" and Julius Malema). Dots are daily values, and lines are seven-day centered moving averages of the proportion of articles containing at least one mention of the terms. The horizontal line is the average daily proportion of the term. The *x*-axis is 42 days (6 weeks) before the election through 14 days (2 weeks) after. The vertical line is election day.

candidates and parties within topics and across topics. Because eight topics are too many to discuss in detail, here we highlight five for which we have predictions (corruption, economy, health, public infrastructure, and security) and include the figures for the other three (education, housing, and voting) in the Online Appendix (which have similar patterns to the topics presented). Overall, the cosine similarities range from 0.05 to 0.65.

The topic of corruption, per **Predictions 1A** and **1B**, provides a good prima facie test of our procedure. Figure 3 corroborates both of our predictions in relation to Jacob Zuma and the ANC: both are more associated with corruption than their competitor parties and candidates; interestingly, Zuma is more associated with corruption than the ANC. While we did not predict ex ante that the opposition parties or candidates would be differentiated from each other, the EFF is the least correlated with corruption across both broad and narrow coverage (although this result is only statistically significant with respect to narrow and pooled), and Julius Malema (EFF) appears more associated with corruption than Helen Zille (DA). Two possible explanations are that Malema used to be affiliated with the ANC such that news coverage might consciously or

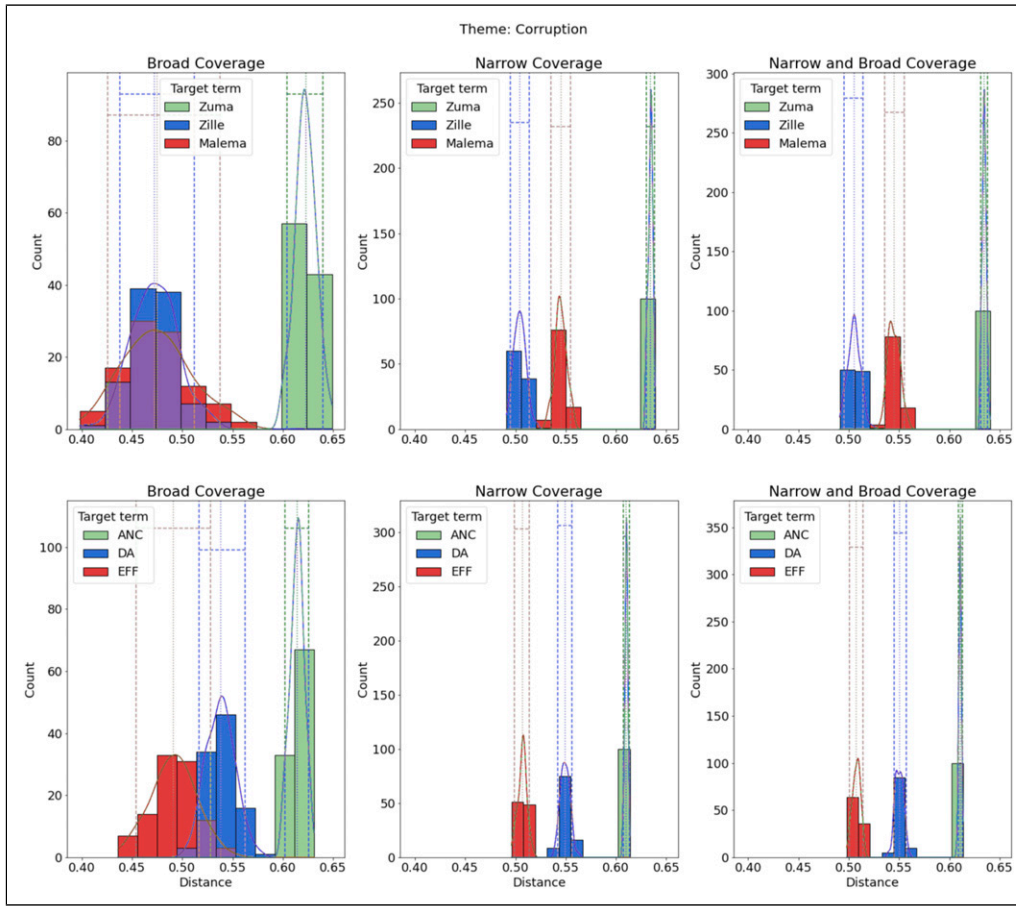

Figure 3. Corruption theme.

subconsciously use more corruption-related terms when speaking about him, or Malema might be making corruption accusations against Zuma (ANC).

Our prediction **2A** that parties would not distinguish themselves on the theme of the economy is validated in Figure 4, which shows that the media covered parties at similar rates irrespective of whether the coverage was broad or narrow. While we did not have a prediction about the candidates, Zille received more coverage related to the economy, particularly in narrow election-related coverage. These findings suggest that her messaging was reported on more than Zuma's and Malema's. This may be a result of Zuma's scandals receiving more press relative to his platform.

As seen in Figure 5 regarding health, contrary to **Prediction 3B**, there is little distinction in coverage across both narrow, broad, and pooled among the parties. Yet we find support for **Prediction 3A** regarding candidates; Helen Zille clearly stands out in the coverage on the topic of health, consistent with her historical focus on health and housing issues in this campaign. While we predicted she would stand out in narrow coverage, she also stands out in broad.

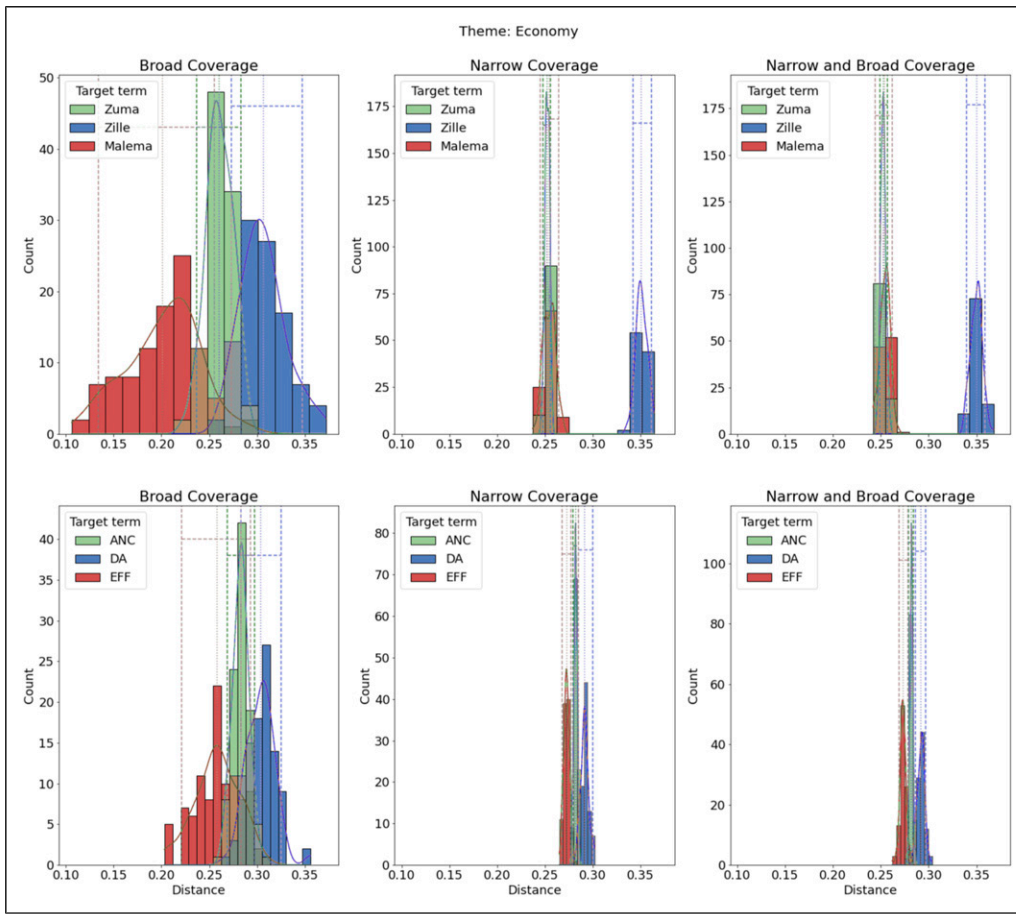

Figure 4. Economy theme.

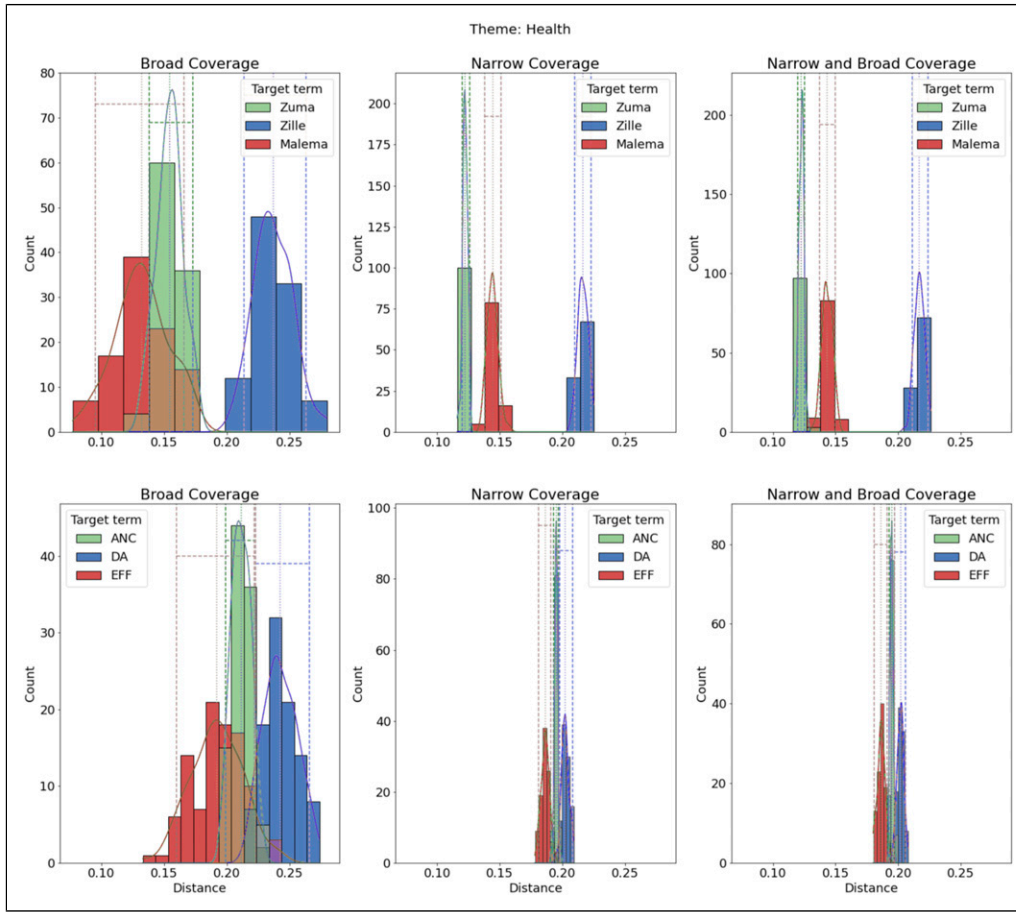

Figure 5. Health theme.

On the theme of infrastructure, we find support for **Prediction 3B**. Figure 6 shows that both the DA and the EFF were much more associated with public infrastructure than the ANC, particularly in narrow coverage.[15] We do not find that Zille distinguished herself more than Malema, but rather both candidates equally distinguished themselves from Zuma. The EFF's and DA's association with voters on these pivotal issues could have contributed to the EFF's victory margin in 2014 (having not previously contested) and the DA's improvement from previous cycles.

On security, we find some support for **Predictions 4A** and **4B**. In Figure 7, Malema and Zille are indistinguishable from one another but stand out more than Zuma, as predicted; this trend is much more evident in the narrow coverage, suggesting that the media picked up on Malema's and Zille's discussions of security-related issues. The parties were not consistently distinguished between themselves, suggesting perhaps that the security situation was so poor it did not lead to one party gaining differential coverage.

Returning to our discussion of the implications of measuring coverage as broad or narrow, the differences between candidates and parties on all election-related coverage and just narrow are almost identical, suggesting that differences between parties and candidates in their coverage are generated primarily in coverage that is directly related to the election. Within the broad category, there is much less clarity in the estimations compared to the narrow, and parties and candidates are

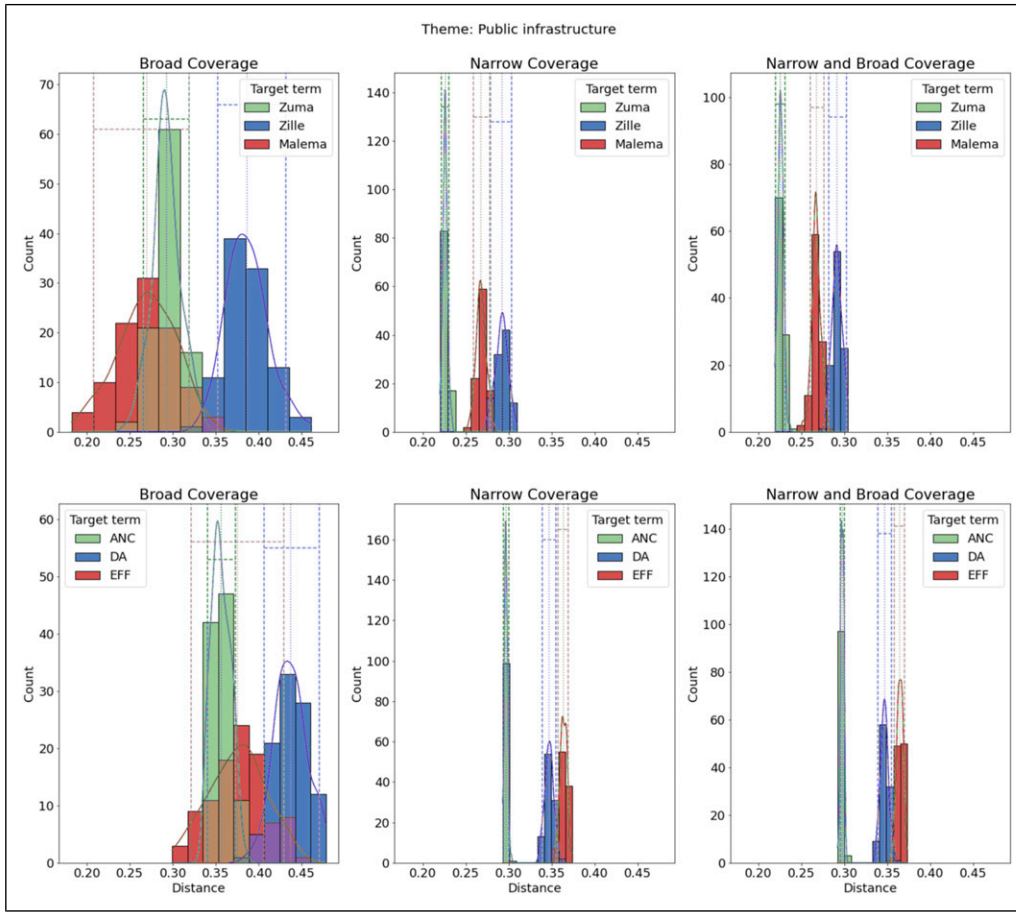

Figure 6. Public infrastructure theme.

often substantively and statistically indistinguishable. In some cases, however, while the differences are not statistically significant, there appear to be substantive differences that further research could disentangle. Nevertheless, one implication is that news readers who specifically read election-related (i.e., narrow) news articles would get more distinct impressions of the parties and candidates.

The difference between broad and narrow in how candidates and parties are associated with themes highlights the importance of coding election-related stories in multiple theoretically informed ways. Admittedly, what we found in South Africa may not appear in different campaign environments, and our findings may reflect the relatively short duration of the campaign; an avenue for further research would be to investigate how coverage varies by electoral environments. And although we expected individuals and candidates to distinguish themselves from the other parties on specific issues (e.g., Zuma regarding ongoing corruption scandals, Zille on health), we sometimes found similar estimates, particularly among opposition parties, which were often indistinguishable from one another in their coverage.

In sum, a textured picture of South Africa's election-related news coverage and its association with parties and candidates emerges from these results. Our data show that candidates and parties

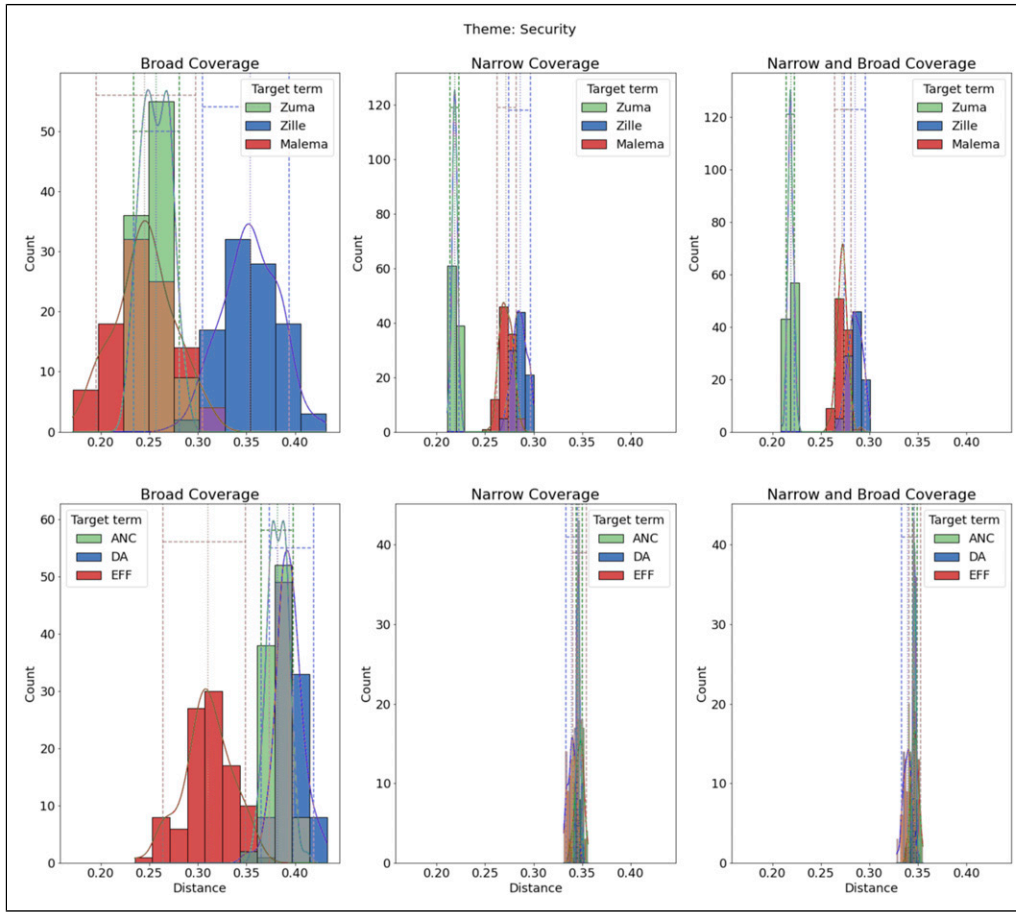

Figure 7. Security theme.

can and do distinguish themselves in meaningful ways and on substantive issues, suggesting that the South African campaign news coverage is dynamic and responsive to the actors and issues.

Conclusion

Our paper presents a multi-step process for examining how news coverage distinguishes candidates and parties in an emerging democracy. We apply our method to investigate campaign media coverage based on an extensive corpus of news stories from South Africa in 2014.

Methodologically, we contribute a new technique that combines supervised and unsupervised machine learning to predict different types of election-related news coverage and estimates which political actors are most associated with the main topics within that news coverage. While our approach does not minimize the importance of identifying robust causal relationships regarding campaigns and voting behavior, we believe machine learning methods that identify and describe empirical regularities in large datasets constitute an important complement, contributing crucial forms of evidence to understanding the information environment during an election period (Monroe et al., 2015).

Our method validates most, but not all, of our empirical predictions. Both the ANC and its presidential candidate Jacob Zuma were much more associated with corruption than opposition parties; while opposition leaders were associated with some of their pet campaign themes, such as the DA's Helen Zille on healthcare. Our method is also designed to explore whether coverage might have differed by party or candidate that a researcher may not predict ex ante; for example, we did not have a clear sense of how the presidential candidates might diverge on topics related to the economy, but this theme was most associated with the leading opposition candidate, Zille.

We also find our theoretical distinction between narrow and broad coverage yields different results. Intuitively, when examining exclusively broad coverage, candidates and parties distinguish themselves in the news coverage much less than when we examine all coverage or narrow coverage exclusively; but these findings suggest that how an analyst perceives election-related coverage matters significantly. Because narrow and broad news varies in quantity and content in many contexts, we believe our results have implications for the study of news and elections across a range of democracies, including characterizations of how media covers campaigns. Moreover, our findings in South Africa imply that voters may receive less differentiated coverage of parties when the coverage is broad, whereas more apparent and substantive distinctions may emerge when narrow.

As in all machine learning applications, our results are subject to researcher decisions. Further work could validate many of our choices regarding the processing steps and hyperparameters we set, and also apply our method to different contexts and expand the coverage of media to include transcripts of television coverage and paywalled stories. Broader electoral and media markets would present more complex electoral settings for future applications. Given constraints on time and resources, particularly in emerging democracies, automated applications such as ours offer new opportunities for research of campaigns. This method allows for us to quickly extract political coverage from all coverage and to distinguish broad coverage from narrow. We see wide-ranging implications for this application to the further study of news and elections in a variety of settings.

Acknowledgements

We acknowledge generous funding from the U.S. Agency for International Development (USAID) Development Innovation Ventures (AID-OAA-A-14-00,004); the Harvard Academy for International and Area Studies [Long]; the Center for Statistics and the Social Sciences (CSSS), University of Washington [Long]; and McGill University [Erlich]. We thank Jeff Arnold, John Beieiler, Adi Eyal and Code4SA, Wes Day, Jonathan Homola, Maura O'Neill, Phil Schrodt, Walter Mebane, Randy Stone, and seminar participants at the University of Washington's Political Economy Forum, the Centre for the Study of Democratic Citizenship's (CSDC) Speaker Series, and DevLab USAID for comments. Stefano Dantas, Ryan Sampana, Stephen Winkler, and Wesley Zudeima provided excellent research assistance. All mistakes remain with the authors and any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of USAID. We provide the code used to generate the main results at https://doi.org/10.7910/DVN/VX945J. For the underlying data used to create the results, please contact the authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Harvard Academy for International and Area

Studies, Harvard University, United States Agency for International Development (AID-OAA-A-14-00004), Center for Statistics and Social Sciences - University of Washington, and McGill University.

ORCID iD

Aaron Erlich https://orcid.org/0000-0001-6571-9081

Supplemental Material

Supplemental material for this article is available online.

Notes

1. We use the word theme interchangeably with issue. We refrain from using "topic" about election-related coverage because we use the word to describe topic models.
2. We do not rule out that in some contexts pre-specifying topics or words make sense. There is a large literature that pre-specifies topics (Green-Pedersen & Stubager, 2010). Such an approach is possible using keyATM (Eshima et al., 2024) or in BERTTopic, the method we use and explain below.
3. Transformer models have a token limit, that is, the amount of words in a document considered to derive its embedding is fixed. As a consequence, large documents might not be well represented by their embeddings. Since we work with news articles with word counts often above the token limit, we split each news article into paragraphs. Therefore, each document, in this case, refers to a paragraph.
4. Other contextual models can also be applied to sentence modeling. One such example is ELMo, mentioned in the previous section. The sentence-BERT model has some key strengths over ELMo. First, it employs a novel masked language modeling technique to represent bidirectional context, instead of using ELMo's less accurate method of right-to-left and left-to-right concatenation. Second, BERT's tokenization is done using sub-words, which is superior ELMo's character-based input (Al-Rfou et al., 2019) for several NLP tasks. Finally, BERT uses a transformer-based architecture, which enables parallelization of training, a key factor when working with a large corpus.
5. The average is used to stabilize the metric.
6. South African Broadcasting Corporation (SABC) is a state broadcaster on television and radio but not newsprint.
7. Appendix A analyzes citizens' media consumption with Afrobarometer survey data: large proportions of South Africans receive news from newspapers, either every day (27%), a few times a week (24%), or once a month (16%).
8. Although the ANC kept a majority in the 2019 election under Cyril Ramaphosa (following Zuma's 2018 resignation due to corruption charges), at the time of writing, polling suggests the ANC will lose a majority in the 2024 election.
9. Of the 14 we did not scrape, all had online presences, but six were behind paywalls, and nine had very restrictive usage agreements. Appendix C describes our data collection process and full list. Because we aimed to build a replicable open-source method, we only included those methods from which we could legally and freely download articles, and curated and sampled publications in partnership with journalists, guidance from SAARF, and the holdings of major conglomerates. Since these publications include South Africa's largest national newspapers in all three major print languages and many smaller local publications, it forms a reasonable approximation of voters' actual election-related news coverage. Nevertheless, we cannot entirely rule out some bias relating to the publications we did not scrape.
10. We report a Cohen's kappa of 0.454 which is "moderate to good" agreement (Banerjee et al., 1999, p. 6). Although there is no agreed standard for intercoder reliability, it is important to be clear when reporting it (Neuendorf, 2017, p. 167). The number of documents will depend on the total number of categories being coded and the reliability of the coders; algorithm performance can also determine if more documents need to be coded (Grimmer et al., 2022, p. 192).

11. We could not use non-experts for our coding because the task required both knowledge of elections generally and South African politics specifically. MTurk was also not feasible since it has limited its workers to India and the United States since 2013. Testing the replicability of expert coding employing crowd-sourcing in Africa is an avenue for future potential research. With Crowdflower and aggregating over 26 Crowdflower channels with 216,107 codings, Benoit et al. (2016) only obtained a total of 722 (0.3%) codings from eight South African coders, suggesting a paucity of coders.
12. This Ridge classifier method uses a ridge regression to create a classification model by converting the target variable into +1 and −1; if the Ridge regression's prediction value is greater than 0, then the predicted class is the positive class, else the predicted class is the negative class.
13. Subsequent analysis used 900 (0.9%) labeled documents after cleaning and preprocessing the corpus.
14. See Appendix Table E.5 for all 50 topics and the top words associated with this model.
15. Appendix Figure D.8 shows that the ANC is much more associated with education.

References

Alexander, P. (2010). Rebellion of the poor: South Africa's service delivery protests - a preliminary analysis. *Review of African Political Economy*, *37*(123), 25–40. https://doi.org/10.1080/03056241003637870

Al-Rfou, R., Choe, D., Constant, N., Guo, M., & Jones, L. (2019). Character-level language modeling with deeper self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 3159–3166. https://doi.org/10.1609/aaai.v33i01.33013159

Angelov, D. (2020). *Top2vec: Distributed representations of topics*. arXiv preprint arXiv:2008.09470.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23. https://doi.org/10.2307/3315487

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 221–229). Association for Computational Linguistics.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295. https://doi.org/10.1017/s0003055416000058

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250–271. https://doi.org/10.1093/poq/nfw007

Budge, I., & Farlie, D. (1983). *Explaining and predicting elections: Issue effects and party strategies in twenty-three democracies*. Allen & Unwin.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining, Osaka, Japan, 25–28 May, 2023.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, 24–28 August, 1998.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv.

Druckman, J. N. (2004). Priming the vote: Campaign effects in a U.S. Senate election. *Political Psychology*, *25*(4), 577–594. https://doi.org/10.1111/j.1467-9221.2004.00388.x

Eshima, S., Imai, K., & Sasaki, T. (2024). Keyword-assisted topic models. *American Journal of Political Science*, *68*(2), 730–750. https://doi.org/10.1111/ajps.12779

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*(34), 226–231.

Ferree, K. E. (2011). *Framing the race in South Africa: The political origins of racial census elections*. Cambridge University Press.

Gerber, A. S., Karlan, D., & Bergan, D. (2009). Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, *1*(2), 35–52. https://doi.org/10.1257/app.1.2.35

Green-Pedersen, C., & Stubager, R. (2010). The political conditionality of mass media influence: When do parties follow mass media attention? *British Journal of Political Science*, *40*(3), 663–677. https://doi.org/10.1017/s0007123410000037

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social Sciences*. Princeton University Press.

Grootendorst, M. (2022). *Bertopic: Neural topic modeling with a class-based tf-idf procedure*. arXiv.

Hall, A. B., & Lim, C. (2018). *Ideology and news content in contested U.S. House primaries*. arXiv. https://andrewbenjaminhall.com/news.pdf

Hayes, D. (2010). The dynamics of agenda convergence and the paradox of competitiveness in presidential campaigns. *Political Research Quarterly*, *63*(3), 594–611. https://doi.org/10.1177/1065912909331426

Hayes, D., & Lawless, J. L. (2018). The decline of local news and its effects: New evidence from longitudinal data. *The Journal of Politics*, *80*(1), 332–336. https://doi.org/10.1086/694105

Horowitz, D. (1985). *Ethnic groups in conflict*. University of California Press.

Lau, J. H., & Baldwin, T. (2016). *An empirical evaluation of doc2vec with practical insights into document embedding generation*. arXiv.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196). Springer.

Lewis-Beck, M. S., & Rice, T. W. (1992). *Forecasting elections*. Cq Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv.

Lodge, T. (2004). The african national congress and its allies. In A. Reynolds (Ed.), *Election '94 South Africa: The campaigns, results and future prospects* (p. 237). St. Martin's Press.

Marimont, R., & Shapiro, M. (1979). Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics*, *24*(1), 59–70. https://doi.org/10.1093/imamat/24.1.59

Mbete, S. (2015). The economic freedom Fighters South Africa's turn towards populism? *Journal of African Elections*, *14*(1), 35–59. https://doi.org/10.20940/jae/2015/v14i1a3

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176. https://doi.org/10.1086/267990

McInnes, L., & Healy, J. (2017). Accelerated hierarchical density based clustering. In 2017 IEEE international conference on data mining workshops (ICDMW), New Orleans, LA, USA, Nov. 18–21 2017.

McInnes, L., Healy, J., & Melville, J. (2018). *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. arXiv.

Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015). No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, *48*(01), 71–74. https://doi.org/10.1017/s1049096514001760

Müller, S. (2020). Media coverage of campaign promises throughout the electoral cycle. *Political Communication*, *37*(5), 696–718. https://doi.org/10.1080/10584609.2020.1744779

Neuendorf, K. A. (2017). *The content analysis guidebook*. Sage Publications, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E.

(2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*(2011), 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. arXiv.

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science*, *40*(3), 825. https://doi.org/10.2307/2111797

Popkin, S. (1994). *The reasoning voter*. Chicago University Press.

Price, V., & Zaller, J. (1993). Who gets the news? Alternative measures of news reception and their implications for research. *Public Opinion Quarterly*, *57*(2), 133–164. https://doi.org/10.1086/269363

Reimers, N., & Gurevych, I. (2019). *Sentence-bert: Sentence embeddings using siamese bert-networks*. arXiv.

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, *28*(1), 87–111. https://doi.org/10.1017/pan.2019.23

Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding regression: Models for context-specific description and inference. *American Political Science Review*, *117*(4), 1255–1274. https://doi.org/10.1017/s0003055422001228

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*. arXiv.

Schreiner, W., & Mattes, R. (2011). *The possibilities of election campaigns as sites for political advocacy*. University of Cape Town.

Soroka, S. N. (2014). *Negativity in democratic politics: Causes and consequences*. Cambridge University Press.

Southall, R. (2014). The South African election of 2014: Retrospect and prospect. *Strategic Review for Southern Africa*, *36*(2), 80–95. https://doi.org/10.35293/srsa.v36i2.170

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)* (pp. 165–174). IEEE.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Vliegenthart, R., Boomgarden, H. G., & Boumans, J. W. (2011). Changes in political news coverage: Personalization, conflict and negativity in British and Dutch newspapers. In K. Brants & K. Voltmer (Eds.), *Political communication in postmodern democracy* (pp. 93–110). Palgrave Macmillan, UK.

Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., Peng, L., & Si, L. (2019). *Structbert: Incorporating language structures into pre-training for deep language understanding*. arXiv.

Watanabe, K. (2017). Measuring news bias: Russia's official news agency ITAR-TASS' coverage of the Ukraine crisis. *European Journal of Communication*, *32*(3), 224–241. https://doi.org/10.1177/0267323117695735

Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

Author Biographies

Aaron Erlich is Associate Professor of Political Science at McGill University where he is a member of the Centre for the Study of Democratic Citizenship and the Centre on Population

Dynamics. His research focuses on the role information plays in developing democracies and quantitative methods.

Danielle F. Jung is Associate Professor of Political Science at Emory University. Her research focuses on understanding legitimacy, social organization, and governance in fragile and emerging states.

James D. Long is Professor of Political Science and Co-founder of the Political Economy Forum at the University of Washington. His research focuses on democracy, development, and corruption.