Covering the Campaign: Election Events in Emerging Democracies

Jeffrey Arnold[*], Aaron Erlich[†], Danielle F. Jung[‡], and James D. Long[§¶]

August 25, 2017

Abstract

Scholars of democratic politics differ in their conceptualization of the role of media in political campaigns and how it subsequently affects political behavior and electoral outcomes. There is therefore no consistent theoretical framework to guide defining, measuring, and analyzing election coverage, particularly in emerging democracies. We apply topic models and supervised learning methods to a news corpus of almost 100,000 news articles and approximately 137,000 tweets from South Africa's 2014 election. We use a theoretically informed classification of election coverage to demonstrate how variation in "narrow" or "broad" conceptions of elections generate variation in codings of coverage of campaign events and activities reported in the media. Our analysis results in radically distinct representations of political actors and institutions in the electoral landscape: a narrow classification includes cues to race, party, and incumbent performance in coverage; a broad definition reflects public policy concerns and service delivery outcomes. Further, examining messaging directly from political actors on social media, we find social media messages by and large reflect the news reported in traditional media and further demonstrate politicians' diverse appeals. We discuss the challenges and opportunities our research and method pose for how scholars conceive of media's role in elections based on their theoretical priors about the form, content, and impact of campaigns in emerging democracies. Our results also provide methods, evidence, and lessons learned for replication to study (electoral and non-electoral) events in other developing country settings.

[*]Department of Political Science, University of Washington, Seattle, WA. jrnold@uw.edu

[†]Department of Political Science, McGill University, Montreal, Quebec. aaron.erlich@mcgill.ca

[‡]Department of Political Science, Emory University, Atlanta, GA. danielle.jung@emory.edu

[§]Department of Political Science, University of Washington, Seattle, WA. jdlong@uw.edu

[¶]We acknowledge generous funding from the U.S. Agency for International Development (USAID) Development Innovation Ventures (AID-OAA-A-14-00004); the Harvard Academy for International and Area Studies [Long]; the Center for Statistics and the Social Sciences (CSSS), University of Washington [Arnold]; and McGill University [Erlich]. We thank John Beieiler, Adi Eyal and Code4SA, Wes Day, Maura O'Neill, Phil Schrodt, and seminar participants at the University of Washington's Forum on Political Economy and Economics, and DevLab USAID for comments. 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.

1 Introduction

In developing democracies, many factors shape voters' electoral behavior by impacting the degree to which they gain and process information about an election, such as knowledge of the candidates on offer, evaluations of parties' past performance, and the credibility of politicians' promises (Keefer and Vlaicu 2008). A country's electoral institutions, like fluid parties (Horowitz and Long 2016), party system volatility (Cox 1997, Ferree 2010, Mozaffar, Scarritt and Galaich 2003, Ordeshook and Shvetsova 1994) and competitiveness (Magaloni 2006), and electoral design (Cox 1997, 2015, Reilly 2016, Scheiner and Moser 2004) affect the choices on offer to voters and therefore the nature of their political engagement. Moreover, individual characteristics like income, gender, education, urban/rural residence, and media consumption provide opportunities and constraints towards voters' gaining political knowledge of, and participation in, democratic processes (Bratton, Mattes and Gyimah-Boadi 2005, Kasara and Suryanarayan 2015, Kuenzi and Lambright 2010). Moreover,

Anticipating the influence of these factors on voters, politicians employ numerous strategies during a campaign to shape the information environment to win citizens' support. Politicians know that voters lack complete or perfect political knowledge and are likely to use shortcuts to inform their electoral decisions (Downs 1957, Lupia and McCubbins 1998, Popkin 1994). In emerging democracies, scholars note that office seekers employ a variety of appeals to influence voters' evaluations and behavior, including signaling social solidarity to co-ethnic or co-partisan bases of support (Horowitz 1985, Snyder 2000, Stokes 2005, Wilkinson 2004), exploiting institutional advantages or employing coercion (Callen and Long 2015, Hyde 2011, Magaloni 2006), contesting the quality of government performance (Long and Gibson 2015), and promising policy outputs and service delivery (Ferree 2011, Gibson, Ferree and Long 2014). Because direct face-to-face contact of voters proves logistically hard and expensive in many developing countries (Ferree and Long 2016), candidates are likely to work to make these appeals to a wider swathe of voters through media coverage during the campaign (Conroy-Krutz and Moehler 2015, Horowitz 2012).

However, prior approaches generate conflicting hypotheses and results on how successful these efforts are likely to be given disagreement on the nature and degree to which campaigns play an active role shaping the information environment in emerging democracies. If, on the one hand, fundamen-

tals like ethnic demography, party identification, or coercive institutions rigidly determine electoral outcomes, the media environment will not serve voters with new information about the election, but rather reinforce their prior beliefs. Campaign appeals from politicians will likely cue these factors (e.g., the candidate's ethnicity, party name, party dominance) to turn out the base of ethnic or party supporters, or remind voters of the coercive apparatus of the government. Appeals are not likely to reflect attempts at providing voters' new information capable of persuasion, or cause voters to change their minds. On the other hand, if situational factors like government performance, policy outputs, or service provision are salient electoral considerations for voters, the media environment can play an important role by shaping their information on these topics, including coverage of government's success at public goods provision, candidate's debates on the quality of government action, and promises of future promises on policies and services. Campaign appeals will therefore explicitly reflect candidates' contrasting views on performance records, policy outcomes, and service provision to influence voters' information in hopes of persuading them. From either perspective, politicians' appeals in the media can play an important role reinforcing and/or shaping the information environment in elections. The question is which predominates and under what conditions.

Further complications prevent straightforward predictions about the role of campaigns in emerging democracies. While citizens in these settings increasingly consume news in a growing and rich media environment, campaign coverage may not automatically or easily influence voters' information, or processing of that information, about the election. First, news coverage may not clearly distinguish between normal everyday politics and those specific to an election. Second, even as many citizens consume media, it is not obvious at what point they start paying attention to the campaign and the election; and political coverage may or may not reflect issues salient to voters. Third, reports on electoral actors, like the government or opposition candidates, may differ in content during an election period compared to non-election periods, generating conflicting or confusing information cues. Fourth, media consumption patterns may reflect multiple "elections" within a single race, if, for example, different segments of the population receive a different informational diet. This issue is especially salient in emerging democracies, where individuals frequently face political, economic, and social

exclusion.[1] For these reasons, even if individuals regularly consume media and politicians exploit news coverage to make appeals to voters, whether the news consistently provides an important role influencing the information environment around elections remains poorly understood in theory and practice. Thus far, a lack of comprehensive data from, and analysis of, political events reported in the media during consolidating elections has prevented a systematic investigation of the form and content of campaign coverage.

To address this challenge, in this paper, we study the content of media coverage of elections in consolidating democracies. We focus on two important foundational aspects of an election's informational environment: (1) reports on events related to political actors and institutions covered in traditional media (e.g., newspapers) and (2) political actors' messaging on social media (e.g., Twitter). In line with other studies, we contend that media coverage is at least partially reflective of politicians' strategies to win support. Candidates understand that reports of campaign rallies, statements, and other events will reflect the appeals they make at those events. Therefore, measuring media coverage of politics is not epiphenomenal to the actual substance of a campaign, but rather provides an important window into the information environment available to voters. This approach establishes a wide range of news that voters could potentially draw upon to receive information about the election, form evaluations of electoral actors, and influence voting behavior. Certainly, no one voter or group of voters consumes *all* of the media available in an election, and media and candidate messaging are not the *only* sources of information available to voters.[2] Our aim is to understand the content of one critical source of the information environment (media reports) and what insights it lends into the likely electoral behavior of politicians and voters.

We argue that how one conceptualizes an election at the outset guides their assessment of the role of media and campaigns in elections. Most obviously, an election can be thought of as the day that people cast ballots to select leaders. But coverage of political actors and institutions in the news does not necessarily easily delineate election-specific material. We argue therefore that two ways to define election-related news coverage matters critically to understanding the information environment: nar-

[1] For example, in poor countries where access to education and media are inconsistent and heterogeneous, a person's race, ethnicity, region, occupation, and urban/rural residence may all affect media access and consumption.

[2] Many other factors no doubt exert direct influence on voters' perceptions and behavior, like party contact through press-the-flesh events or get-out-the-vote efforts.

row and broad. In the "narrow" case, election events are restricted to focusing on political actors and institutions directly relevant to the decisions voters must make on election day, with stories specifically mentioning the election, and cuing electoral actors and institutions. This definition derives from the assumption that voters lack information access, do not necessarily pay very much attention to the election, or discount election information in their news consumption. If voters only narrowly consume election-related news, politicians will need to make appeals through clear and obvious information shortcuts, for example, cues regarding the ethnic identity of candidates, party names contesting office, incumbent presidents' appeals for re-election, or reminders of government coercive capacity would qualify, as would mentions of specific electoral processes, such as voter registration, voting procedures, or the election commission's counting of ballots. Our narrow conception of election-related news accords with the view that voters lack information or engagement during the campaign and therefore politicians will employ cues that play to their bases, rather than require additional information or information processing on the part of voters suggestive of attempts at persuasion.

In the "broad" case, we conceive a more expansive definition of election-related events to include any coverage related to political actors and institutions in the election period, regardless of whether the election is specifically mentioned. This concept follows from the assumption that voters have access to, consume, and employ information reported in the news during the campaign that may not rely exclusively upon direct information cuing the election, like mention of a candidate or party, but involve information on political actors and institutions more generally. If voters consumer a large information diet during the campaign, politician's appeals will consider broader cues towards the issues, policies, government action, and outcomes that could impact voters' evaluations and behavior. This does not suggest that "easier" to digest shortcuts on ethnicity or party do not matter, but rather that politicians' appeals cannot take this for granted by assuming that voters do not employ information in the campaign to reason about their choices. For example, coverage of government institutions like the police do not directly present citizens with election-related content because the police are not elected and do not contest elections. But, information about the crime rate or police effectiveness may play a role influencing voters evaluations of their electoral choices if they consume that information during the campaign period if, for instance, by updating their information about government effectiveness on the part of the incumbent party or promises of crime reduction or police reform by candidates. Our

broad conception of election-related news echoes the view that voters consume and employ information during campaigns that help them reason over electoral choices. Therefore, politicians will employ cues to better inform and persuade voters by providing details about government performance, policies, and services.

Subsequently, we argue that defining election coverage either narrowly or broadly shapes both the frequency of news reports and their substantive content in ways that present voters with different information environments, and therefore, radically different representations of how politicians and voters treat campaigns. Specifically, different theoretical priors on what constitutes an "election" and election-related coverage in the news plays a powerful role in how we study, define, measure, and interpret an election's information environment. To analyze and test empirical implications arising from our approach, we examine news coverage of South Africa's 2014 national election. South Africa presents a critical case because it has many features of a consolidating democracy and allows us to hold many country-level factors constant, like the media environment and the electoral system, that may shape politicians' and voters' behavior. The country's media landscape is fairly free and open, with many different local and national papers in a variety of languages and a growing ICT infrastructure to support news consumption, especially on social media. This provides an unusually rich and comprehensive data environment of election-related events.[3] But South Africa also provides a hard test for exploring the role of campaigns given the political dynamics arising from social identity and the party system, which question whether the media environment actually matters for understanding political behavior in South Africa. Since the country's transition from apartheid to democracy, electoral outcomes have been considered a foregone conclusion because of the dominance of the African National Congress (ANC) party, backed by the support of a majority of the country's black population. Given the legacies of apartheid, race correlates strongly with party support in South Africa (Ferree 2011, 2006, Garcia-Rivero 2006, Lodge 1995). Therefore, South Africa is demonstrative of other emerging democracies where media consumption is increasing and the news reports extensively on electoral races (Horowitz and Long 2016), but social identity and political features of consolidating institutions, like dominant parties (Magaloni 2006) or restrictions to competition (Levitsky and Way

[3]South Africans are by and large heavy consumers of media (see Table 1), outpacing other developing countries with lower incomes and less media access, such as much of sub-Saharan Africa. But South Africa also represents where these societies are headed given the rapid expansion of ICT and media coverage.

2010), also constrain how voters and politicians employ information in their electoral behavior.

To evaluate how variation in conceptions of election coverage generates variation in the form and content of the campaign, we first match our theoretically-informed priors on what constitutes an election, whether narrow or broad, with the actual news events in South Africa. To do so, we leverage the fact that most South African newspapers are free and digital, and collected a large corpus of South African news articles, almost 100,000 from over 100 publications, and over 137,000 social media posts from political actors.[4] We used experts to label a sample of articles and tweets as election-relevant or not, and from that, employed supervised prediction methods to classify the remaining articles in the corpus into narrow and broad classifications of election-relatedness. We apply a topic model to summarize the overall news coverage from traditional and social media in South Africa, generating qualitative and quantitative measures applied to political actors, institutions, and events. By regressing the labelled articles on these topics, we describe the news content most predictive of election relatedness as well as estimate the topical composition of the articles in the election-related categories for narrow and broad classifications.[5] With a similar procedure, we examine how political actors used social media (Twitter) to complement or substitute the election coverage in traditional media. Specifically, we examine the types and content of their output to relate to tradition media as a check on whether news coverage reflects politicians' direct appeals and messaging. We also examine how messaging content changes over time and to different groups as a reflection of strategies that directly reach potential supporters.

To preview a first set of results, variation in defining election-related material narrowly or broadly shapes the amount and content of events. First, a narrow definition generates election coverage focused on the main political parties, candidates, and the incumbent ANC president's Jacob Zuma candidacy for re-election, including material related to his associated corruption scandals.[6] From this narrow classification, this election coverage includes cues to race and party in coverage of the main political

[4]In contrast, many previous studies pre-select stories that they believe to contain election coverage in order to examine campaign events (Ferree 2011, Horowitz 2012) or use reported media attentiveness from survey data (Banducci and Karp 2003, Horowitz and Long 2016).

[5]Our method somewhat parallels other studies that use media coverage to examine reported events like political violence (Bagozzi and Schrodt 2012, Schrodt 2012), however, we focus our techniques in one country and hone in on political coverage of elections (broadly-defined).

[6]A scandal relating to President Jacob Zuma's usage of state funds to build himself an opulent rural home (Nkandla) dominated much of the presidential coverage during the 2014.

parties and candidates, and also includes information about the incumbent's performance record. Second, a broad definition generates additional election coverage reflective of public policy concerns on poverty and unemployment, flag-bearer campaign issues in this particular race, as well as other service delivery outputs regarding education, electricity provision, health care, and workers' rights that were less directly linked to specific parties or candidates, but were nonetheless salient in campaign coverage and could have influenced voters' evaluations of political actors. This suggests that appeals related to government action and policy promises provided a more multi-dimensional information environment under a broad conception. Therefore, a stricter understanding of election coverage is rich in details about the identity and party characteristics of the main political actors contesting office, but misses qualitatively important factors that are captured in the media and election environment compared to a broader conception. Additionally, either a narrow or broad definition provides news coverage that contain racial and partisan cues, but both also reveal a wider and richer information environment that voters could have consumed, including with respect to performance and policy.

In a second set of results, we find that direct appeals from politicians over social media largely reflect campaign coverage in traditional media First, we find that either from a narrow or broad definition, the content of appeals on social media cue the identity of candidates and parties, as well as government performance and relevant policy issues. We interpret this as evidence that media reports in traditional media represent politicians' strategies in their coverage and are not random noise or systematically biased away from the appeals that politicians make directly. Further, examining the content of social media specifically, we show variation in the types of appeals made by party. The incumbent ANC was much more likely to appeal to voters' party identification and "duty" of upholding the ANC legacy as a reason to turn out to support the president and party; conversely, the opposition Democratic Alliance (DA) and Economic Freedom Fighters (EFF) messaged more directly on the failures of the ANC and the state of the economy, while also providing substantive policy promises on the issues. Therefore, while an array of appeals matters generally, parties may differentially employ these appeals in the campaign to directly target their supporters. Therefore, the question of whether and how campaigns are covered by media coverage is not just a question of whether campaigns "matter," but also for whom they matter and in what ways. Our results demonstrate that while social media is a complement, not a substitute, to strategies and appeals picked up and conveyed in traditional media,

social media also allows politicians to harness and more directly influence the direction of information conveyed through news to voters.

All told, our results show differences in levels and proportions of the type of coverage of a campaign depend on the conceptualization of what constitutes election coverage in the media, both as reported in traditional news and over social media. We believe our findings provide important implications about the content of the information environment for voters and the explicit strategies and appeals used by politicians to persuade voters. Our results do not undermine the well-established correlation between race, party identification, and voting behavior in South Africa, and it is beyond the scope of this project to definitively or directly link the information environment to voters' electoral decisions. But our study contributes to prior work by underscoring the importance of a more nuanced view of campaigns and campaign coverage in the media to understand fundamental dynamics of political competition in emerging democracies. How one defines election-related content from the outset importantly shapes the nature and extent of the information environment, providing different appeals and information shortcuts available to voters. Therefore, the degree to which scholars, politicians, and voters treat the importance of campaigns rests fundamentally on assumptions about how to conceive election coverage and voters' engagement with it. Following from these assumptions, our results show that whether voters consume a narrow versus broad election media diet, they are likely exposed to a multiplicity of information shortcuts. Racial and party cues no doubt play a role, but even a narrow reporting on election news provided voters with information beyond ethnic and party appeals that could have informed their choices. Our findings importantly demonstrate that previous approaches to the role of campaigns in emerging democracies may want to more systematically analyze media to better understand how politicians and voters potentially engage the information environment, and whether and how this differs in direct appeals on social media versus events reported in traditional media. Our approach lends weight to a burgeoning literature that factors beyond social identity and party institutions, such as government performance and policy outcomes, matter towards the form and content of electoral competition and outcomes in emerging democracies.

Our paper proceeds as follows. In the next section, we review common approaches to studying electoral campaigns in emerging democracies. We delineate two broad views from the literature that provide contrasting perspectives on the role of campaigns towards shaping the information environ-

ment. We then explicate our theoretical approach. In Section 3, we provide background to elections, campaigns, and media consumption in South Africa. Next, we discuss our measurement strategy and how our theory informed our coding in Section 4. In Section 5, we present results, first with respect to news media coverage and second regarding political actors' messaging on social media. We conclude in Section 6 by discussing the opportunities and challenges afforded by our approach to studying media coverage of campaigns in consolidating democracies, and relate the comparative insights of our theory, methods, and results to diverse social science and data science literatures.

2 Theoretical Foundations

2.1 Campaigns in Emerging Democracies

Voters require information about political actors and institutions, and they use this information when making electoral choices (Downs 1957, Lupia and McCubbins 1998). Therefore, the campaign should prove an important time period in allowing politicians to employ strategies to appeal to voters (Popkin 1994). What information do voters have by the time they reach the ballot box to vote on election day in emerging democracies? Many factors directly and indirectly shape individuals' information, information processing, and behavior. In developing countries, voters face important resource and institutional constraints towards casting ballots, like socio-demographic factors including education, income, gender, residence, and media consumption (Bratton, Mattes and Gyimah-Boadi 2005, Kasara and Suryanarayan 2015, Kuenzi and Lambright 2010). Institutions including parties (Horowitz and Long 2016), party systems (Ferree 2010, Mozaffar, Scarritt and Galaich 2003, Ordeshook and Shvetsova 1994), and electoral systems (Cox 1997, 2015) may exert influences over voters as well. These personal and contextual variables generate variation in the amount and quality of information voters receive to adjudicate between alternative candidates. Politicians therefore have strong incentives to use media coverage to their strategic advantage by influencing the information environment. But do campaigns shape opinion?

In a first body of literature, scholars downplay expectations about the role of campaigns towards providing voters with new information to help inform their decisions. First, many scholars advocate

that citizens in these countries face informational deficiencies, from problems in educational attainment to media consumption, and therefore rely on information shortcuts that provide the easiest and most readily available cues to help inform their choices. Scholars argue that in societies where social identity proves salient to politics, ethnic cues become the most predictive factor of vote choice (Chandra 2004, Posner 2005).[7] Ethnic cues are valuable because they are easily determined by voters and may provide a shortcut to politicians' past or future behavior (Chandra 2004). Voters may rely on ethnic cues because they lack knowledge or high quality information regarding government performance or policy outcomes. In ethnic party systems, voters may use short-cuts based on candidate's ethnicity or the ethnic reputations of party labels (Ferree 2011). Additionally, the hybrid nature of democracy in transitioning societies means that semi-authoritarian constraints on competition and influence over institutions can create strong expectations about likely electoral winners, like in dominant party systems or electoral autocracies (Blaydes 2011, Magaloni 2006), or those that restrict competition and employ coercive or corrupt tactics (Callen et al. 2013, Hyde 2011, Magaloni 2010). Voters may therefore not update beliefs or evaluations without meaningful choices when competition is restricted and rely on cues related to these dominant political forces.

This perspective provides implications for campaigns and the information environment. If fundamentals like ethnic demography, party identification, or party coercion rigidly determine electoral outcomes, the media environment will not provide voters with new information or cues that are likely to update their opinions about candidates or persuade. Rather, the media will likely reinforce their prior beliefs, such as the ethnic affiliation of party labels or a dominant party's coercive capacity. Politicians need not devote campaign resources and messaging towards persuasion since electoral outcomes are considered a foregone conclusion insofar as the predicted vote choice for voters is determined by their social identity or party dynamics. We therefore expect the coverage and content related to political actors and institutions not to change significantly in response to an election, for example, by altering messaging or appeals made in the campaign. Politicians' appeals should reflect strategies derived from ethnic or institutional opportunities to turn out their bases.

[7]Strong attachments of in-group feelings and expressions of ethnic solidarity cause voters to strongly prefer co-ethnic candidates or ethnic coalitions (de Kadt forthcoming, de Kadt and Sands n.d., Friedman 2004, Horowitz 1985, Rabushka and Shepsle 1972). Voters may also select co-ethnics because they deem promises more credibility when made by co-ethnic politicians (Dawson 1994, Ferree 2006, Keefer and Vlaicu 2008), including over the direction of policy and patronage (Bates 1974, Chandra 2004, Posner 2005).

In a second body of literature, scholars take a contrasting approach and argue that campaigns may play a vital role shaping the information environment as voters seek information beyond ethnic and party cues. First, they may use information related to government performance to form evaluations of incumbent politicians. "Bread and butter" issues (Barkan 1976), like government service delivery of needed public goods (Long and Gibson 2015), could matter to voters perceptions of whether they want to continue support of an incumbent (Bratton and Kimenyi 2008, Bratton, Mattes and Gyimah-Boadi 2005, Hoffman and Long 2013, Lindberg and Morrison 2008). Voters may also obtain information about candidate's policy positions adopted and promoted in the campaign (Horowitz and Long 2016, Long and Gibson 2015) to select those who match them on the issues and they deem likely to pursue their desired policy outcomes regarding public services or government reforms (Gibson, Ferree and Long 2014). Performance and policy evaluations are not fixed, but can change over time, if, for instance, an incumbent party who performed well in the past performs poorly in the present, or if an opposition candidate without a record and only promises gains office and then has a record. If performance and policy are salient election considerations, voters will seek and employ information that helps them evaluate incumbent governments, opposition parties, and the credibility of promises made by both.

According to this second perspective, voters may seek a multiplicity of information cues. The media environment should provide important information as candidates contest and debate these issues to persuade voters, and signal with respect to government performance and policy adoption. Campaigns, and how they are covered in the media, therefore play an important role in highlighting issues and conveying information about candidates to help voters form evaluations. We would expect a diversity of kinds of things covered by the media during the campaign, with explicit appeals from politicians based on more malleable and situational factors, like coverage of performance and polices, to convince voters as opposed to just fixed ethnic or institutional variables that voters already have information about and politicians cannot change. This could first mean that coverage during an election would significantly change in form or content from non-electoral periods. It could also mean that politicians' appeals and messaging, as picked up by media, reflect strategies that are not simply about turning out the base based on fixed levels of support, but include factors indicative of attempts at persuasion.

From either perspective, politicians' appeals in the media play an important role reinforcing and

shaping the information environment in elections. Adjudicating between the two broad theoretical orientations regarding the role of campaigns in new democracies generates opposing hypotheses and divergent predictions about the role of media and information in campaigns. In the first view, campaigns are not likely to differ from normal non-election periods of coverage, and campaign coverage should reflect appeals made to ethnic and party shortcuts. Conversely, if campaigns do matter, politicians' appeals should grow more diverse and substantive, possibly providing voters with information on performance and policy concerns. Both sets of factors can have an impact, but the question is which predominates and under what conditions. Whereas these previous studies have advanced a series of mechanisms to advance claims about the relevance of campaigns based on information short-cuts available to voters, they have not yet investigated the degree to which media coverage lends insights into politicians' appeals and use of a variety of cues to shape the information environment.

2.2 Defining 'Election' Coverage in the Campaign

In our study, we examine media coverage to gain theoretical and empirical traction on the content of campaigns and the electoral information environment in emerging democracies. Our aim is not to measure whether the information conveyed in media reports ultimately affected whether people participated, how they voted, or whether they were persuaded or changed their minds. Moreover, we do not claim that information in the news perfectly reflects the totality of voters' information environment, that media reports are the only empirical evidence of politicians' appeals, or that our approach allows us to test the relative salience of specific information cues for specific voters or the impact of these cues on behavior. Rather, our goal is to explore the content of the media landscape to examine how variation from theoretical priors about what constitutes election-coverage shapes the information environment and whether the form and content of news reports accord with likely strategies suggestive that politicians take campaigns seriously, and if so, in what ways based on the appeals they make. We accord with prior studies that many voters lack high quality information about electoral features and many individuals will only partially be engaged with election news to help inform their decision. This includes information about ethnicity and party. We also agree that politicians have incentives to utilize coverage of the campaign in media to employ strategies beyond ethnic and party appeals to help reach voters. This potentially creates new and more information

cues and those that are relevant to an election, including with respect to government performance and policy.

We proceed by first asking "what is an election?" to understand the theoretical parameters by which we understand whether and how coverage of political events during an election likely shaped the information environment. Most obviously, an election can be thought of as the day that people cast ballots to select leaders, but we also think that the informational environment around the election matters in how people are engaged. We argue that how one conceptualizes an election at the outset guides how one assesses the role of media and campaigns in elections. All coverage of political actors and institutions generally does not necessarily easily delineate election-specific material. We therefore highlight two ways to define news coverage related to elections that will matter towards understanding the information environment.

First, election events can be conceived of *narrowly* as specifically focusing on political actors and institutions directly relevant to the decisions voters must make on election day. Since voters may not necessarily pay very much attention to the events, actors, and issues with the campaign, a narrow definition may focus on cues like party name or candidate ethnicity. Such voters would require strong signals about election candidates and electoral institutions. This definition follows from the assumption that voters do not necessarily pay much attention, and therefore must be cued very clearly and obviously to election-related content. For example, reports about about the parties contesting office or the incumbent president would qualify as narrow, as would mentions of specific electoral practices or institutions, such as voter registration, voting procedures, or the election commission. Party names, candidate names (which cue ethnicity), and mentions of election day would all form cues that voters receive in a narrow definition.

Second and alternatively, election events can be conceived more *broadly* to include any coverage related to political actors and institutions in the election period. This concept follows from the assumption that voters are more open to more information during the campaign, and may not only rely upon direct information cuing the election specifically, but may consider broader issues that could impact evaluations and behavior. Coverage of government institutions may not directly present citizens with election-related content, but consumed during an election period, may gain salience in a voter's information environment and therefore impact their evaluations and choices. For example, the

health ministry or the police (both government institutions) may not feature directly in an election where voters are not selecting public officials who staff government agencies. But coverage of these otherwise non-election-specific actors during an election could contribute considerations for voters. If anything at all related to politics near an election is considered an election event, the possibilities of what appeals politicians make and what information voters have is large. This is not to say that everything matters – obviously, for individual politicians or voters they will lend differential attention to certain strategies or information sources. But what it does portray is a very large and potentially diverse information setting for voters who would potentially employ any event related to politics near an election in their calculus.

We subsequently argue that these different theoretical priors on what constitutes an "election" and whether election-related coverage is narrow or broad in the news plays a powerful role in how we study, define, and measure the information environment. Defining election coverage either narrowly or broadly shapes the frequency of news reports and their substance in ways that present voters with different types of information cues. As coverage of the election is more narrow, we expect voters to be more focused on easily accessible shortcuts requiring the least additional information, like racial or party cues. This reflects appeals aimed at turning out the base, but not typically persuasion. As election coverage grows more broadly, we expect voters to be exposed to more and more complex cues that require more information, like on government performance and policy. Therefore, politicians' use of media in campaigning will reflect more diverse appeals indicative of attempts at persuasion and beyond only turning out the base. Whether narrow or broad, we argue that fundamentally, different theoretical priors on what constitutes an "election" and election-related coverage in the news matter critically to measuring the form and content of the information environment.

Additionally, whether or not candidates specifically convey certain messaging through appeals carried by traditional media, the news will also report on events over which politicians have no control, such as a gaff or misstatement. Therefore politicians may try to use media to directly influence information and the media environment. To do so, they can harness social media to speak to the voters they want to make specific appeals to. Precisely because there are different groups who may receive different election information diets, appeals may be made to different segments of the population allowing us to more directly gain empirical traction on politicians' strategies, and whether they mirror

the appeals carried by traditional news or suggest other strategies at play.

Our approach leads to a number of potential testable implications. First, we expect that variation in a narrow or broad approach generates differences between types of campaign coverage and campaign coverage as a proportion of total news coverage that indicates whether and how campaigns matter towards shaping the information environment. Second, the content and form of appeals could vary by characteristics like party or language. For example, incumbent or dominant parties may make different substantive appeals based on their powerful standing, compared to opposition parties. If dominant parties are worried about turning out the base, they may appeal to good partisans and loyalty. However, if opposition parties want to take votes away, they may say incumbent has performed poorly or is corrupt. Third, on social media, we may observe different appeals when looking at traditional news compared to direct messaging. If the reported news reflects well candidates' appeals and strategies, we would expect to see differences in the distributions due primarily to the medium itself (e.g., tweets are shorter), rather than major divergence on the content of the messages (e.g., party cues versus performance cues).[8]

3 Setting: South Africa

South Africa provides an important case to investigate the role of campaigns and the information environment in emerging democracies. The benefit of focusing our empirical scope to one country is that it allows us to hold country-level factors constant to better test the observable implications of our theoretical approach to measure variation in election-relatedness to variation in coverage. However, in this section, we review South Africa's institutional environment to situate our study in comparative analysis by contextualizing how important political dynamics contribute qualitatively to South African electoral campaigns and voters' information environment.

In 1994, South Africa held free elections with universal franchise for the first time transitioning from apartheid (Johnson and Schlemmer 1996, Reynolds 1994). Although a minority in South Africa, descendents of white settlers and European immigrants solidified a white-dominated apartheid regime after the country gained independence from the United Kingdom that enacted policies to curtail the

[8]Within each of these categories we may also observe changes over the course of the campaign.

rights of the majority black population and suppress their political and economic activity. But in 1994, blacks overwhelmingly supported Nelson Mandela of the African National Congress (ANC) and elected him to the presidency, and the ANC gained a large majority of seats in parliament.

South Africa's electoral system provides an important window into many of the likely campaign dynamics at elections. The country forms a single nation-wide legislative constituency that yields 400 members to its National Assembly (parliament). The seats are apportioned proportionally to parties based on vote shares. Parties determine party lists and candidate rankings, and the leader of the majority party in the National Assembly forms the executive. While voters typically have knowledge of party leaders and top-ranked party officials who are likely to gain seats, they have considerably less knowledge of individuals down the list (Reynolds and Reilly 1997). This electoral system creates powerful incentives for politicians to build strong parties and move up the party ranking, perhaps at the expense of building strong personal linkages with constituents (Mattes and Southall 2004, Reynolds 1999). Appeals to party and party mobilization is therefore an important campaign strategy to win seats (Lodge 2004, Mattes, Gouws and Kotze 1995).

While South Africa has many robust democratic institutions, this institutional environment overlaid with the remnants of apartheid creates challenges regarding the nature and level of political competition that may play an important role shaping politicians' and voters' electoral engagement and behavior. Since 1994, the ANC has won national contests with wide margins in subsequent elections, therefore maintaining executive and legislative dominance since transition. Numerous studies document the well-established correlation between race and party support: blacks overwhelmingly support the ANC and since 1994, elections have mirrored an "ethnic census" (Ferree 2011, Garcia-Rivero 2006, Lodge 1995). The relationship between race and party support is not surprising given that the apartheid regime promoted white dominance and the expense of black exclusion and curtailed black activism under the ANC. Since 1994, under a dominant ANC, smaller parties have faced challenges winning at the national level, including the largest opposition party in 2014, the Democratic Alliance (DA), which draws large number of white and "Coloured" (mixed-race) supporters. The identity of party leaders not only present voters with stark racial choices, with a black ANC leader (Jacob Zuma) and white DA leader (Helen Zille), but given South Africa's political development, party labels also contain heuristics that guide voters' evaluations of a party's racial credentials (Ferree 2011, 2006).

For these reasons, many observers of South African elections typically do not think campaigns matter very much in terms of conveying new information to voters. The institutional and racial environment perennially point to ANC dominance. The campaign dynamics are very much about turning out the base. Related to our concepts, this means that the ANC will continue to make racial appeals to consolidate its black support, and also remind ANC members to do their duty as good ANC supporters to turn out and vote for them (Southall 2014*a*). Given the death of Nelson Mandela in December 2013, a few months before the May 2014 election, these appeals could have been especially successful at reminding black South African voters of the ANC and Mandela's legacy. 2014 was also the first election in which "born free" voters could vote—that is, adults of voting age (18+), born after the end of apartheid. It was especially feared that these voters would take the ANC for granted and not turn out, so reminding them of the ANC's liberation struggle during apartheid was especially important (Mattes 2012). Such a strategy would accord with the ANC's typical outlook to turn out their base using the party's strong mobilizing capacity. One could conceive of the Democratic Alliance as trying to increase their margins by turning out their base of white and coloured voters in Gauteng and the Western Cape (Booysen 2005), two provinces where they are strong.

However, the racial and institutional advantages enjoyed by the ANC does not mean that campaigns lack substance in South Africa, even if electoral results are consistent. Indeed, South African campaigns reflect many bread and butter issues (Habib and Sanusha 2006, Mattes and Piombo 2001). The DA tries to use the ANC's institutional advantage against it and the party has also tried to make cross-ethnic appeals to gain black support. Part of this strategy involves point out the ANC's lackluster performance, particularly regarding its controversial party leader, Jacob Zuma. Although now in the political majority, many blacks do not feel that the ANC's performance lives up to the promises made as apartheid ended, especially young voters. The 2015 unemployment rate of 26% was the highest in a decade, and over half of black youths are jobless. Regardless of race, many voters increasingly perceive the ANC, and incumbent president Jacob Zuma, as corrupt (Southall 2014*b*). And while the ANC dominates black support, South Africa has a long tradition of activism, strikes, and protest going back to the anti-apartheid struggle (Lodge 1983, Lodge and Nasson 1991), and there have been perennial and widespread protests against the government for its poor record of performance on delivering public services (Alexander 2010, Southall 2014*b*), allowing many supporters

who vote for the ANC nontheless to express their frustrations at the government.

Political and electoral opposition to ANC dominance has not just arisen from the DA or the ANC's traditional supporters. Factions within the ANC have tried to break away and gain support as the new voice of a black South Africa under various party labels. While these appeals still include race, they also typically harness an explicit economic message. In 2014, this included the Congress of the People (COPE) and the leftist Economic Freedom Fighters (EFF), led by Julius Malema. These divisions reflect the complications faced by the ANC to hold together a large coalition of black South Africans from different regions, economic classes, and age groups, including the difficult position of either holding the center at the expense of the left or tacking to the left at the expense of the center. Therefore, we might expect people who even at one point broadly supported the ANC and black cause to shift their appeals towards a party that is a sub-group of the ANC based on other economic or policy priorities.

There are additional reasons to suspect that the media landscape provides access to significant campaign exposure. First, South Africa has a long history of national and local newspapers in a variety of languages, and a high literacy rate. Second and more recently, in terms of technological development, South Africa has enjoyed rapid growth of ICT in recent years. It boasts the highest cellular phone connections per capita in Africa,[9] and the fifth highest internet access rate. Cell phone saturation was almost 90% in the 2011 census and has since risen to almost 100%. Web-enabled feature phones and smartphones currently have a saturation rate of 70%. More economically developed areas of South Africa have higher usage rates, as do younger and more male populations. This would provide citizens with the tools to more easily access published news and social media. Overall, as Table 3 shows, large proportions of South Africans receive news from radio and newspapers. This generally holds across socio-demographic categories, with some differences based on race and gender.

These realities at the outset paint a very mixed picture of South African campaigns. On the one hand, the remnants of apartheid and importance of race along with ANC dominance all suggest that beyond racial and party cues, little could matter beyond that, and campaigns insofar as they matter, work towards getting out the base. However, the political dynamics of the race and various other economic conditions could have made government performance and policy issues also important, and

[9] As of 2014, 149 connections per 100 citizens; Nigeria has 77.84 per 100 (World Bank 2014).

	n	(%)	0	1	2	3	4
Sex							
Female	1206	50	13	5	9	10	11
Male	1184	50	10	4	7	13	16
Race							
Black/African	1665	70	20	8	13	17	17
White/European	251	11	1	1	1	2	5
Colored/mixed	327	14	2	1	2	3	3
South Asian	142	6	0	0	0	1	2
Other	5	0	0				
Age							
[18,32.4]	978	41	8	5	9	10	10
(32.4,46.9]	749	31	7	2	5	9	9
(46.9,61.3]	456	19	5	2	2	4	6
(61.3,75.7]	169	7	3	1	1	1	1
(75.7,90.2]	28	1	1	0	0	0	0
(90.2,105]	2	0	0				
Education							
No formal schooling	69	3	2	0	0	0	0
Some primary / primary completed	289	12	6	2	2	2	2
Some secondary / secondary completed	1399	59	13	6	11	15	14
Post-secondary training (not uni.)	351	15	2	1	2	4	6
Some university / university completed	281	12	0	0	1	2	6
Water source							
Inside home / compound	1897	79	15	7	12	19	24
Outside compound	478	20	7	3	4	4	3
Party affiliation							
ANC	965	40	13	5	7	10	9
COPE	21	1	0		0	0	0
DA	344	14	2	1	2	3	5
EFF	148	6	1	1	1	2	2
No party	746	31	6	2	4	7	8
Other	87	4	1	0	1	1	1

Newspaper Consumption as Percentage of Demographic Group (0 = Never, 4 = Every day)

the media environment is conducive to having a lot of information, not just a little.

4 Research Design

In this section, we summarize our data collection and then outline our research design. Our empirical aim is to examine news reports and whether campaign coverage indicates various appeals reflective of politicians' strategies and how this shaped voters' information environment. Specifically, we will look at the content of the differences between narrowly and broadly drawn definitions of elections. News reports necessarily miss some political events that may have occurred, and therefore, the appeals of which voters not in attendance would have missed. But, insofar as politicians want the press to cover their activities (e.g., speeches, rallies, policy proposals, debates over issues) to get out their appeals to influence voters' information environment, news reports should be an important tool reflecting the actual content of the campaign. Moreover, social media allows political actors to make appeals to voters directly through messaging. Of course, if certain events, or certain types of events are not systematically reported in the news, media reports may bias our understanding of the true underlying features of the campaign, particularly when compared to social media posts. We address these potential sources of bias below and employ data from social media as a check on our prior about news coverage. While imperfect, we think that overall, our data and analysis provide an important first step in mapping tools to study election-relatedness of the campaign in the news, with the most complete dataset available to us.

4.1 Data and Methodology

The total event corpus consists of nearly 100,000 articles from a wide range of South African newspapers and approximately 137,000 tweets from the accounts of political actors and institutions (e.g., President Zuma, members of parliament, the electoral commission).

Newspaper Corpus

We generated the corpus of news stories from 167 South African publications obtained via news sites that had a digital web presence. The corpus consists of news articles captured by the scrapers between

March 15, 2014 and June 1, 2014, the period around the 2014 South African general election, and includes many publications with the largest distribution (shown in Table 2), along with many smaller and local publications. We also scraped several online only news sources such as *News24*. In summary, the corpus included everything possible that we could digitize and to which we could legally obtain access. We believe this is a reasonable approximation of the actual information environment to which voters might be exposed.

The articles were scraped with customized software developed in partnership with Code for South Africa (Code4SA), a South African non-profit civic technology lab (Eyal 2016). We scraped news sites both in cases where there existed RSS feed(s) and cases where a newspaper had a web presence but no RSS feed.[10] For all publications, we collected all available articles, and therefore did not specify *ex ante* any rule to determine whether a story was politically or election related. Since the scraper relied on publicly available RSS feeds and news sites, the sample of publications did not include newspapers which had a pay wall, did not have an Internet version, or which explicitly prohibited scraping in their terms of service.[11]

Following Davenport and Moore (2015), our goal is to capture as many election-related stories as possible, but we could not obtain all of them for two reasons. First, to receive news coverage at all an event would need to be "newsworthy" insofar as it received any reporting and publication. Given the number of national and local news sources in our corpus, we believe that any event missing from coverage completely likely did not generate unique appeals or affect the information environment to such a large degree as to suggest bias in the nature of appeals we do pick up from reported events. Second, as noted, our corpus includes all news stories available to us, a sample neither completely universal nor random. Our sample is therefore designed to obtain as complete as possible coverage of election-related reporting in the South African print media but may miss some coverage only reported in papers not in our sample. Nevertheless, we believe that our sample goes far beyond other efforts and attenuates source bias that may exist by only employing only a small number of well known media

[10] However, if the news site had RSS feed(s) we used the RSS feed rather than scraping the stories off the html rendered page.

[11] Publications such as *The Diamond Fields Advertiser* only market a PDF version of their print copy. *The Sowetan* has a very restrictive terms and conditions that specifically prohibits downloading stories and our partners had concerns about several titles with the Times Media Group. The paywall restriction excluded three major Afrikaans publications, *Die Burger*, *Beeld* and *Volksblad*.

outlets (Reeves, Shellman and Stewart 2006).

Our sample includes newspapers published in English, Afrikaans, and Zulu (isiZulu) titles.[12] Table 2 shows the top publications by number of articles in our data. The majority of all titles in the South African market publish in English. English is the first language of less than 10 percent of the population, but is spoken widely as a second or third language. We explicitly sought diverse language coverage because as Holmes (2015, 278) notes, the majority of South African media analysis and research ignores the non-English press. Therefore, Table 2 shows a large Afrikaans market, which is the first language of approximately 14 percent of the population.[13] In recent years, South Africa has witnessed growth in isiZulu publications. IsiZulu is the African language which has witnessed the largest media growth, as a result of their relatively large population size and economic standing among black African groups (Ndlovu 2011). Importantly, our data contains *Isolezwe*, the largest isiZulu publication in circulation. In the corpus, 80,240 (81%) articles were in English, 16,581 (17%) articles were in Afrikaans, and 532 (1%) articles were in isiZulu.

In addition to language diversity, our sample includes titles from all of the main South African media conglomerates.[14] In May 2014, the time of this study, newspaper ownership was highly concentrated. Four media houses owned the majority of print media in South Africa: Caxton,[15] Independent News and Media South Africa (INMSA), Naspers/Media24, and the Times Media Group (TMG)[16] (Angelopulo and Potgieter 2016).[17]

In summary at the time of scraping, most newspapers in South Africa had free online availability, so our sample includes many of the major newspapers in South Africa. However, as shown in Table 2, our method unavoidably missed several large publications. While our corpus aims for coverage of the South African media landscape, we do not, at this juncture aim for representativeness.[18]

[12]Some smaller publications are multilingual, reporting in English and Afrikaans or Zulu.

[13]English language dominance and the secondary market for Afrikaans language news reflect the white dominance under British colonial rule and subsequent apartheid era of development of media in South Africa (Tomaselli 1997).

[14]Media ownership in South Africa in recent years has both integrated into a few large conglomerates and globalized, with increasing foreign ownership.

[15]Caxton publishers are known for their small free and community newspapers. Our samples included both their largest newspapers, as well as many smaller local newspapers.

[16]Formerly Avusa.

[17]All conglomerates have English titles, however there is some differentiation in whether they own Afrikaans or Zulu publications. Naspers/Media24 owns the major Afrikaans newspapers. INMSA has the only large Zulu publication in our sample, *Isolezwe*.

[18]One possibility for future analysis would be to include weights based on readership numbers.

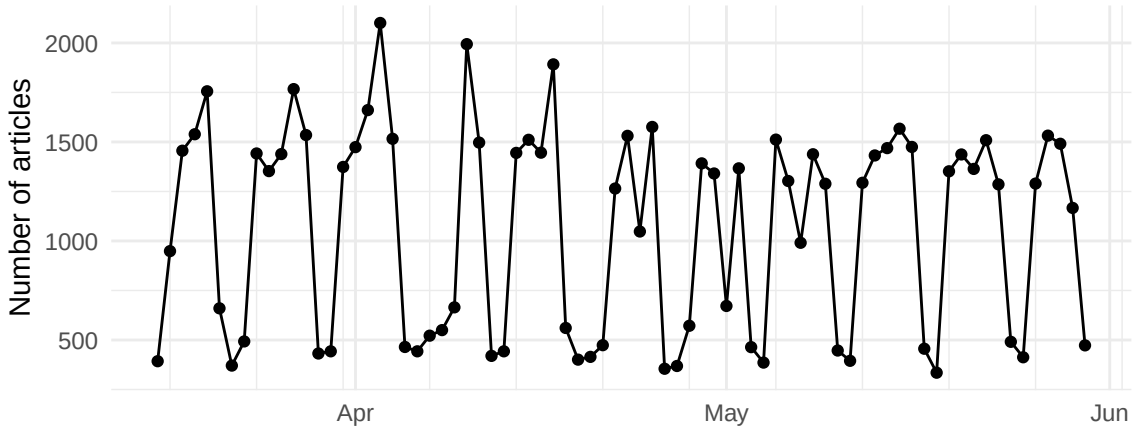

Figure 1: Number of articles per day in the corpus. The cyclical pattern is due to the day of the week, with the troughs occurring on weekends. Major grid lines are on the first of the month, and minor grid lines are weekly.

Social Media Corpus

In addition to newspapers, we collected 137,969 tweets from 30,558 usernames on Twitter, a popular social media site for political actors in South Africa to message directly. Users were candidates, official party accounts, or otherwise clearly political users. These tweets originate from the Twitter accounts of leading candidates in the 2014 South African election with a Twitter account, e.g. Jacob Zuma (e.g., @SAPresident), and a curated list of other important political South African Twitter accounts, including sitting members of parliament, institutional accounts at national and provincial levels (@DA_KZN, @DAGauteng)[19], other government officials, and government accounts (e.g., @Parliament_of_RSA).[20] These tweets were collected from April 2, 2014 to May 15, 2014.

Election Relatedness Category Labelling

To our knowledge, there is no *a priori* method to determine whether a news article is or is not politically relevant and then related to an election. To address this problem, we randomly sampled 1,000 (approximately one percent) of the articles to hand label from the complete set of newspaper articles in our corpus after the scraping.

First, each coder was shown the full title and full text for an article and asked to select one of three possible labels: "Election-related," "Not election-related," and "Indeterminate." Before manually labelling the stories, two of the co-authors first coded a sample of 100 of the 1,000 stories and met to discuss inter-coder differences. After closely examining the differences in the understanding of "election-related," three of the authors reached the agreement that the difference between two of the authors was a difference between a narrow conception of electoral coverage and a broad conception of electoral coverage. In the narrow conception, political actors or institutions specifically related to or mentioning the election or election day were coded. For example, this would include mention of the incumbent president Jacob Zuma (running for re-election), the ANC and DA parties (both contesting for office), or registration and ballot counting performed by the electoral commission. In the broader conception, mention of any political actor or institution was coded whether specifically related to the election or not. A broad classification would pick up the same coverage as narrow, but

[19]DA refers to Democratic Alliance and KZN is an abbreviation for Kwa-Zulu Natal province. Gauteng is the name of another province.

[20]Additionally, the corpus includes retweets of those users and tweets in which those usernames are mentioned.

also include mention of government agencies, such as the health ministry or police, or other political actors, such as the courts, which were not specifically related to the election race or election day. The two coders continued coding using these two different conceptions of election-relatedness, generate a three category classification system for all the sample: "narrowly election-related", "broadly election-related", and "non-election related".

We took two approaches to understand the coverage of the election in terms of these categories. First, in order to understand the content of the articles in each of these categories, we estimated a topic model to generate topics for each article, and then regressed these topics on these categories. Second, we describe how these categories varied across time, and between publications.

Table 5 highlights the type of stories that are typical of both the "broad" and "narrow" categories. In the narrow category, a story that reports on potential misconduct of the election commission (IEC) chairwoman Pansy Tlakula is coded as narrowly election related. Other stories such as one about police being deployed for election related-security or another story about Mmusi Maimane, the Black African candidate for the premier of Gauteng are also coded as "narrow," since they explicitly discuss election related matters. A story reporting about a court case involving mineworks is not coded as narrow. But in the broad category, stories related to a mineworkers' strike that had been ongoing since 2013 and was about their wages was coded. While this strike did not reference anything particular about the election and just happened to occur during the election cycle, it is coded broadly even though it the stories do not specifically mention anything related to the campaign or to politicians running for office. Another example of the broad category involves stories about trucks damaging roads, specifically highlight a public policy problem that could theoretically become an electoral issue, but again the story does not tie the issue directly to the election campaign.

Of the 984 documents labelled, 103 (10%) were labelled as "narrowly" election related, 154 (10%) were labelled as "broadly" election related, the remainder were labelled as not election related.

4.2 Summarizing News With Topic Models

Given the size of the corpus, we first summarize the content of articles using a 32-topic topic model. Second, we find those topics most associated with each election-relatedness (narrow or broad) category by regressing these topics on the labelled documents.

Table 6 shows the top seven words for each topic using several labelling methods (FREX, Lift, Score, and most frequent words) and the two documents with the highest estimated posterior proportion of the topic. Topic models were estimated using a correlated topic model (CTM) (Blei and Lafferty 2009).[21]

We tokenize and annotate our documents with part-of-speech and named entity tags with SpaCy.[22] After tagging these documents, we kept only content words (adjectives, adverbs, nouns, and verbs) and named-entities (people, locations, and organization). Named entities with multiple words were treated as a single token, meaning that an occurence of "jacob zuma" , the incumbenent president, in the document is included in the analysis as `jacob_zuma`, rather than as `jacob` and `zuma`. We pruned the vocabulary by removing words appearing in less than 2 documents and keeping words with the highest term-frequency inverse-document frequency (TF-IDF). Using TF-IDF to filter words effectively drops corpus specific stop-words since they will have a high document frequency (low inverse-document frequency) (Blei and Lafferty 2009).

The choice of the number of topics is domain and question dependent. We estimated models with 8, 16, 32, 64, and 128 topics. Using the topic fit measures of exclusivity and semantic cohesion (Lucas et al. 2015), the models with 32- and 64-topics have the highest semantic cohesion. We chose 32 because it performed best for all measures of predictive accuracy considered (see Section 5.2).

5 Results

5.1 Interpreting Categories with Topic Models

To explain and understand the narrow and broad election categories from section 4.1 topics, we regressed these data on the topics estimated in section 4.2. We ran a multinomial logit regression with a lasso penalty on the labelled subset of data with the topics as features.[23] Table 4 plots the coefficients for each topic from the regression of all topics on the election relatedness categories.

[21] We used the implementation of the CTM in the **stm** Rpackage (Lucas et al. 2015, Roberts, Stewart and Tingley 2016, Roberts et al. 2013).

[22] SpaCy is a Python package providing an end-to-end NLP pipeline: `https://github.com/explosion/spaCy`.

[23] The lasso regularization imposes sparsity in the coefficient estimates, meaning most coefficients will be zero, allowing us to focus on the coefficients that most reliable predict with a topic is predictive of narrow election coverage, broad election coverage, or neither.

We will focus on three core results of the model. First, the individual coefficients and predicted probabilities associated with each topic tell us which topics are associated with each category. Second, the overall number of non-zero coefficients for each topic tell us how coherent each category is, meaning how many different topics are related to a category. Third, the measures of predictive accuracy help us to understand the ease with which we can filter political or non-political news coverage.

As expected, the narrow election coverage is explained by fewer topics. Only 4 topics have non-zero coefficients. For example, in Topic 27, a topic which almost perfectly predicts the narrow topic, the most associated tokens for this topic include three political parties that have narrow racial and ethnic electoral bases — The Inkatha Freedom Party (IFP), the Economic Freedom Fighters (EFF), and the National Freedom Party (NFP), and the location where the worst election violence occurred in 2014, in addition to words that are almost always used with voting.[24] The other two topics most associated with narrow election coverage also contain the names of party leaders and parties, but additionally include words that were the center of the main parties' election campaigns. Particularly, Topic 24 focuses on the tokens "empowerment," "unemployment," "poverty," and "inequality." Both the ANC and DA campaigns put the issues of unemployment and poverty front and center in their campaigns.[25] This finding suggests that a voter consuming news along a narrow dimension would receive information specifically about both the largest political and more minor political parties, including with respect to their reputations regarding race, and their specific economic messaging and their strategies to highlight their proposed strength on the economy, as well as issues specifically related to voting processes, such as violence.

The broad election coverage is explained by more topics as expected, with 8, which is twice as many as the narrow category.[26] Of the 8 topics associated with broad elections, the tokens in 7 load fairly neatly onto distinct areas of public policy and service provision. For example, topic 16 focuses on issues of financial regulation and taxation, which would be attributed to the incumbent government. Topic 18 clusters neatly onto issues of education (including tertiary education), Topic

[24]The IFP is a party that represents Zulu interests, the EFF is a party that splintered from the ANC, which espouses more left-wing redistributive politics for blacks, and the NFP is a splinter party from the IFP.

[25]The 2014 ANC campaign made reducing inequality a central campaign promise.

[26]Although the lasso penalty imposes sparsity in the estimation, there was no constraint enforced in the regression that coefficients could load on only one category. Despite that, most coefficients did, with only topics 32 and 23 having non-zero coefficients in two categories, and in both cases having a negative coefficient for the non-election category, and a positive coefficient for the broad category.

7 focuses on healthcare and medical services, and Topic 14 relates to water and power utilities (the two often linked through hydroelectricity). Topic 32 is perhaps the most unique to South Africa and relates to the country's large and politically active mining sector. The Association of Mineworkers and Construction Union (AMCU) was on strike for the entire electoral cycle in the most expensive strike in South Africa's history. Topic 5 also covers a politically salient issue in South Africa related to crime and criminality. Siboniso Miya and Radovan Krejcir were two defendants in a high profile criminal case including political corruption.[27] The broad election category would make election coverage much more expansive than the narrow topics discussed above not just in amount of news but the content, by also including public policy and service provision issues around topics like education, utilities, healthcare, and workers' rights.

Last, the remaining topics that are highly predictive of non-election related coverage are associated with music (topic 11), sports (31, 3, 19), local events (15), and foreign affairs (17).[28]

Figure 4 compares the predictive performance of the topics for each of the categories using multiple metrics: Brier score, log score, accuracy, precision, recall (sensitivity), and specificity.[29] Just using topics, we achieve 85–92% accuracy across the categories. The narrow election topic and non-election topic perform best in different classification statistics. In precision (proportion of true positive values to predicted positive values) and recall (proportion of true positive values to actual positive values), the predictions for the non-election category perform better. But in the aggregate metrics: accuracy, Brier score, log score, and in specificity (proportion of true negatives to actual negatives), the narrow election topic does the best. The takeaway from these results are that narrowly election-related articles are easy to identify, while the broader category of potentially election-relevant articles is much harder to classify.

Although only the labelled documents were used in the regression, the topics were estimated using

[27] A criminal case with some political overtones again highlights the difficulties in teasing out the relationships between media coverage and elections.

[28] In South Africa, tokens associated with foreign affairs do not appear co-located with election-related tokens. This results lends itself to the supposition that foreign affairs do not play a large role in the South African context, at least as far as elections and campaigns are concerned.

[29] Each statistic is the mean value from a 10-fold cross-validation. During training, the lasso model weighted observations so that the classes were assigned equal weight even though they are unevenly balanced in the sample; in doing so, we are trading off performance in classifying non-election stories for better performance in classifying election stories. Across all the metrics (except specificity - proportion of true negatives to real negatives).

the whole corpus, allowing us to draw strength from all the available text.[30]

5.2 Overall Election Coverage

In order to get a complete view of the election coverage in our sample, we used the labelled set of documents to predict the election relatedness categories for the full sample. As in the previous section, we run a lasso regression with the categories as the output variable and the topics from the previous section, publication, functions of the date, and language as features. In the previous section, we did not include non-topic features in the estimation because our goal was to understand the unconditional relationship between topics and election-relatedness, not the election relatedness of a topic conditional on it being in an article in certain publication, in a certain language, and at a certain time. However, for estimation, our objective is to obtain the best predictions possible without regard for interpretability, so we include the available information and use regularization and cross-validation to avoid over-fitting.

Figure 6 plots the proportion of documents in each category. The share of stories in each category is stable after July 2013 - with non-election articles at 50–65% of the sample, broad election coverage at 25–35% of the sample, and narrow election coverage at 10–15% of the coverage. The only major change in overall coverage is the week of the election in which almost 50% of articles are in the narrow election coverage category. Figures 7 and 8 plot the proportion of topics over time within articles classified as narrow and broadly politically relevant.

There is also variation between political coverage by language: the English and Zulu publications in the sample have more both narrow and broad election coverage. Table 3 lists the proportion of articles in each election coverage category aggregated over all the English publications average higher coverage of broad and narrow election coverage (26% and 13%) relative to Afrikaans publications (11% and 2%). The Zulu publications are similar in coverage to that of the English publications, with 15% narrow political coverage, 26% broad political coverage. Figure 9 plots the proportion of each topic within each language (Afrikaans, English, and Zulu) for each day in the sample. The Zulu proportion of coverage is more variable than the others, due to the smaller number of articles. The

[30]The shrinkage parameter (λ) was chosen by cross validation. In estimation with the training set the observations were weighted in order to balance the classes. The function `cv.glmnet` from the R package **glmnet** was used for estimation.

Figure 2: Coefficients for each topic from multinomial logit regression with a Lasso penalty of all topics on the election relatedness categories. For the plot coefficients are truncated at ± 4 since these are extremely large numbers on a logit scale, and there is little information beyond that size.

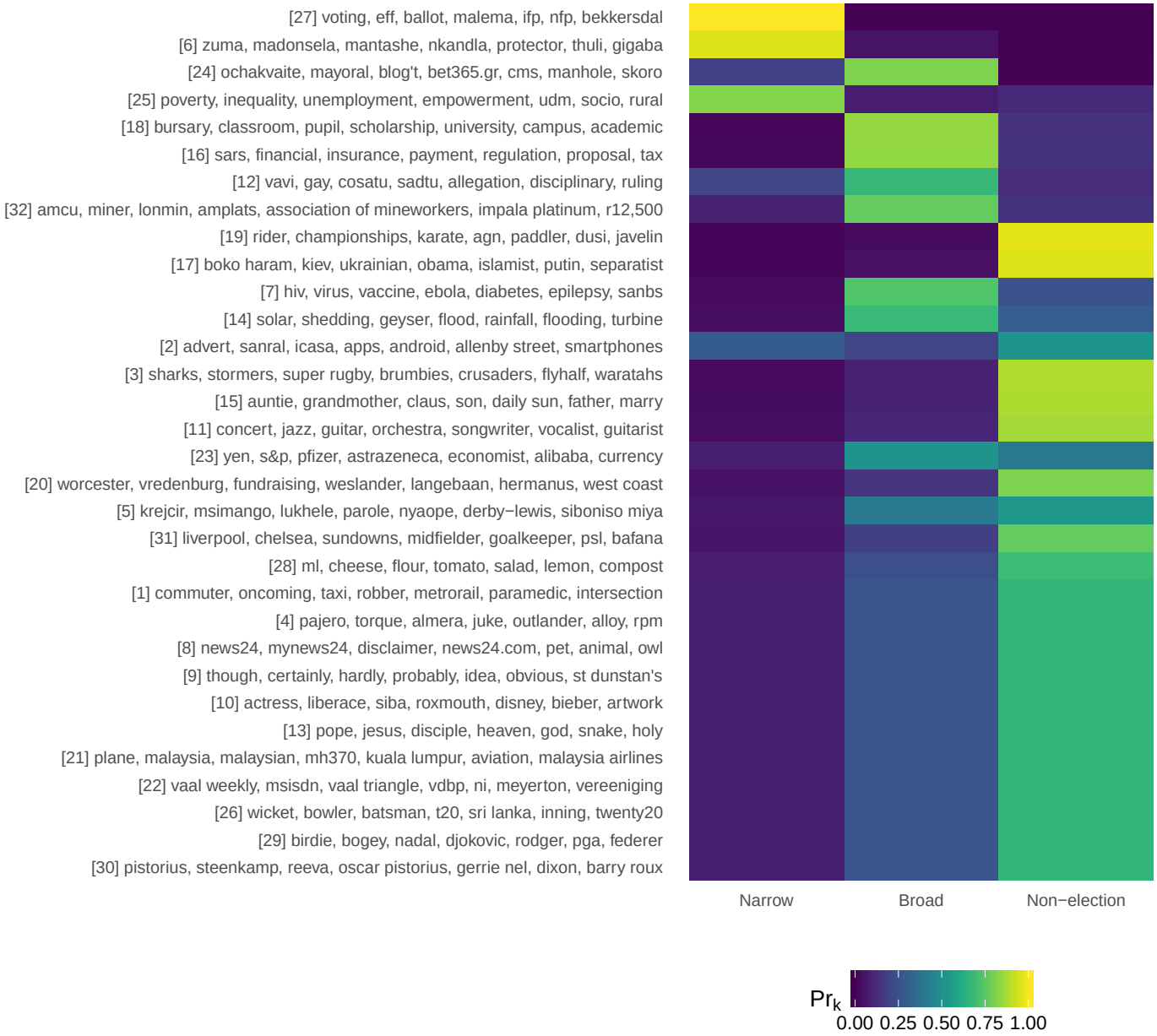

Figure 3: Coefficients for each topic from multinomial logit regression with a Lasso penalty of all topics on the election relatedness categories.

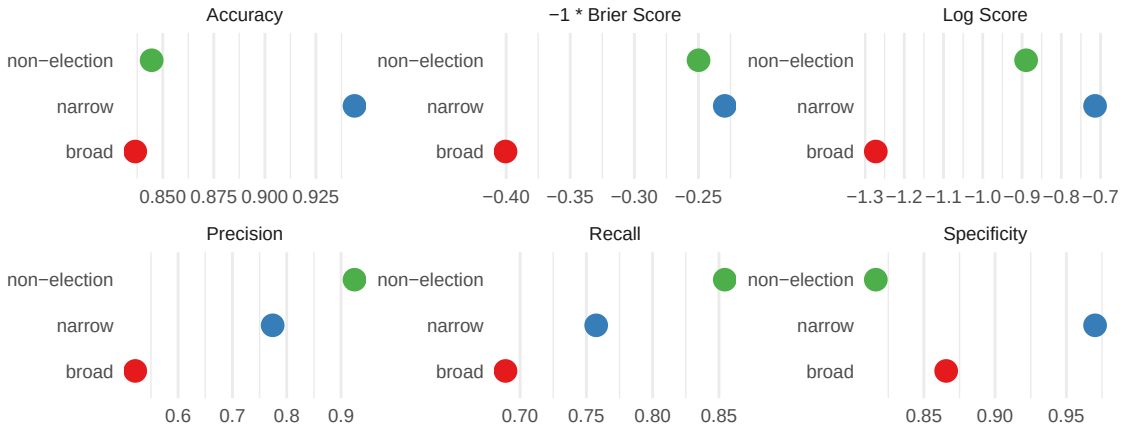

Figure 4: The broad category is the hardest to predict. The plot shows the results of several classification metrics from a 10-fold cross-validation of the lasso regression of election category on 32 topics. For all measures, higher is better

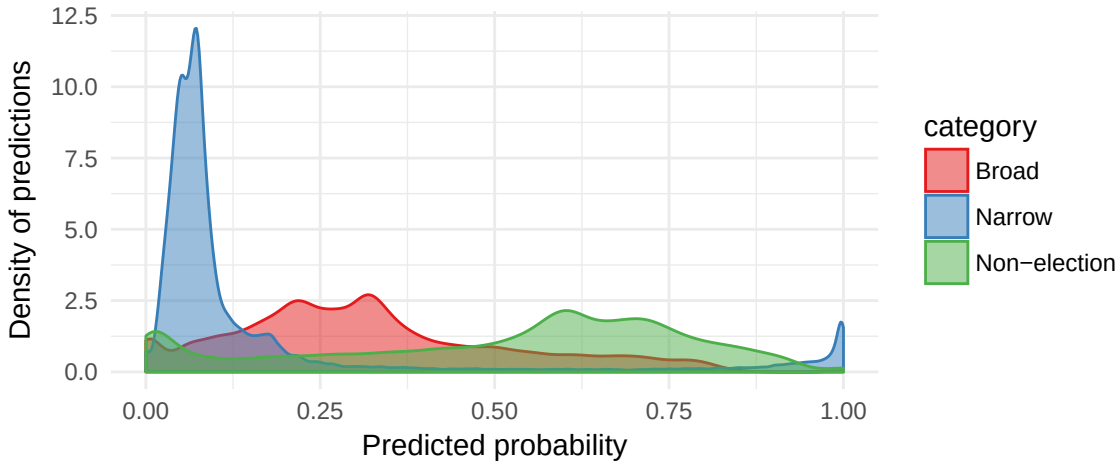

Figure 5: This figure plots distributions of the predicted probabilities for each category for the lasso regression model of the election categories on 32 topics. The narrow election topics are easy to predict, with the predicted probabilities either close to zero or one. However, the broad election related topic is difficult to predict from the content alone; in few cases does it even have probabilities above .75.

English articles show an increase in narrow-political coverage in the period leading up to the election, but the Afrikaans articles do not.

5.3 Comparing Social Media and Newspaper Coverage

How do the topics of political discussion during the election covered in the news differ from those coming from social media? To investigate this, we estimated the topics present in a corpus of both tweets from politicians, parties, and political accounts, as well as the previously discussed newspaper articles above. Although there are some differences due to differences used in the types of language used in the two media, we find that the topics of political discussion are similar.

This analysis uses the corpus of news articles discussed in Section 4.1 and the Twitter corpus discussed in Section 4.1. However, combining these two corpora poses several difficulties due to differences in the lengths and languages of documents in each. It is difficult to estimate topic models using individual tweets as documents since the small number of characters in tweets provide little information about the co-occurrence of words. A common method for dealing with this problem is to aggregate tweets into larger documents for analysis (Mehrotra et al. 2013). Thus, we aggregate tweets on two levels. First, we identify the top 100 usernames by the number of tweets in our corpus. All usernames not in that list are combined into single "others" user. Second, we aggregate all tweets from a single username, including the "others" user, into a single document for each date. Combined, these pre-processing steps aggregated the tweets into 2,042 documents.

In addition to the size of tweets, language differences in Tweets and newspaper articles can pose a problem for analysis. We employed several pre-processing steps to remove some of these differences.[31] Hashtags and mentions form an important part of discourse on Twitter, but do not appear in the formal writing in the news paper articles. However, many Twitter usernames and hashtags consist of one or more words that do appear in the newspaper articles. We pre-possessed the texts of tweets by replacing Twitter usernames with the names of the person or organization associated with the username and replacing splitting hashtags into their constituent words. These replacements were generated from a dictionary of politician accounts and the most common hashtags and heuristically

[31] The structural topic model (STM) provides a statistical method for modeling differences in language models conditional on topics. We estimated several models with a twitter indicator variable in the STM content equation, but the resulting topics were not interpretable. We may explore this issue in the future.

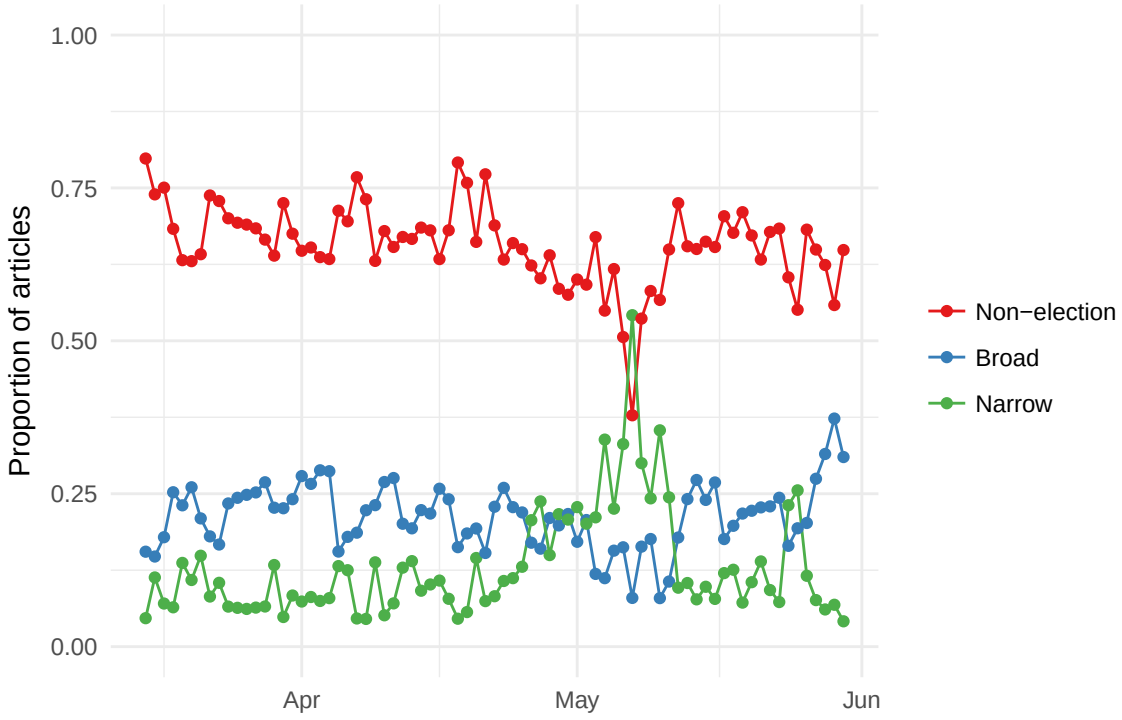

Figure 6: Proportion of articles in corpus in the broad, narrow, or non-election related coverage, by date, from March 15 to June 1, 2014. This uses the assigned labels of the labelled subset, and the predicted probabilities for the unlabelled subset.

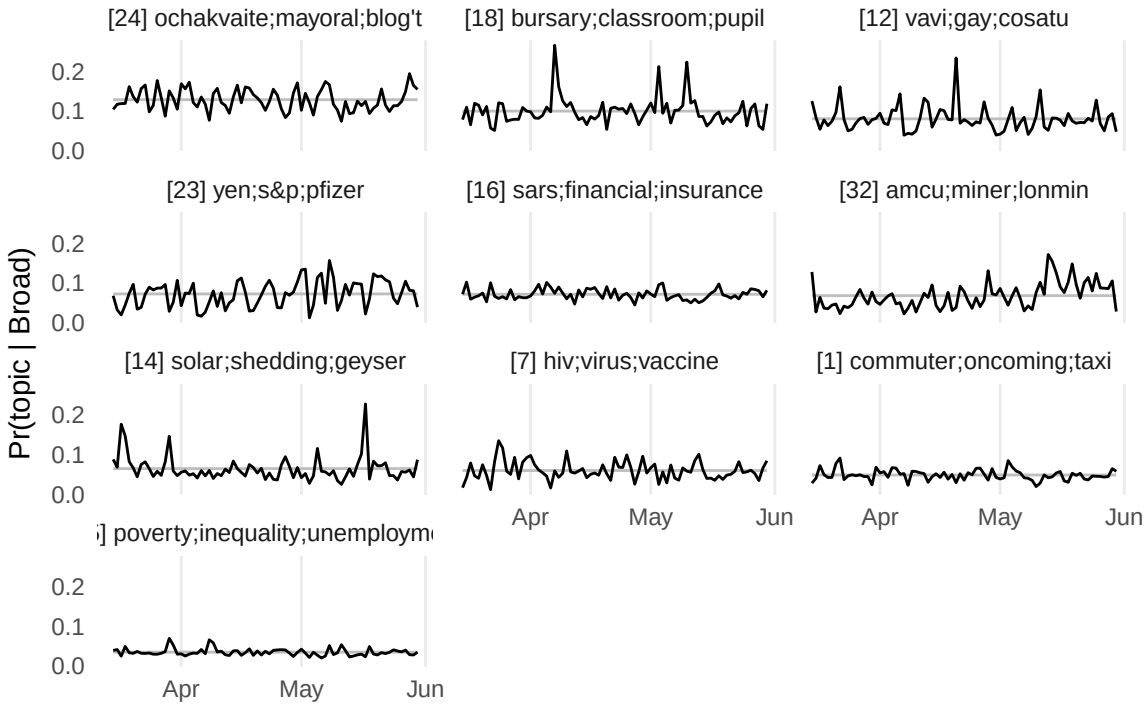

Figure 7: Proportion of topics by day for articles classified to in the "Broad" category (March 15-June 1, 2014). The gray line within each topic is the mean weekly proportion for that topic.

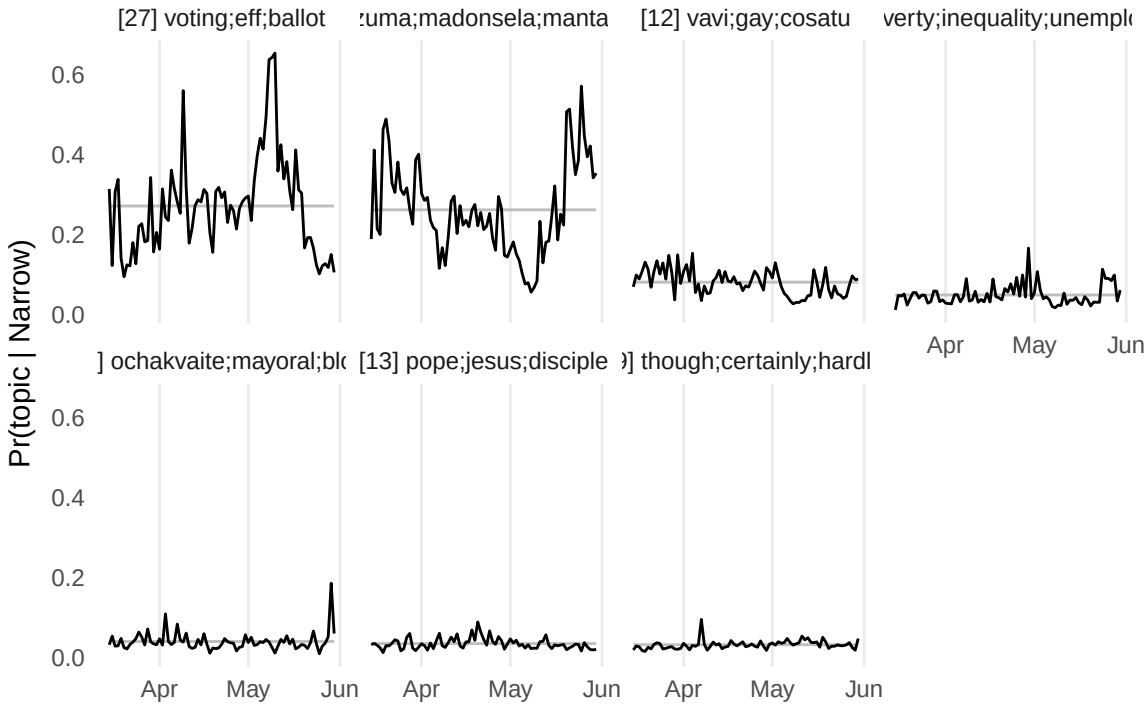

Figure 8: Proportion of topics by day for articles classified to in the "Narrow" category (March 15-June 1, 2014). The gray line within each topic is the mean daily proportion for that topic.

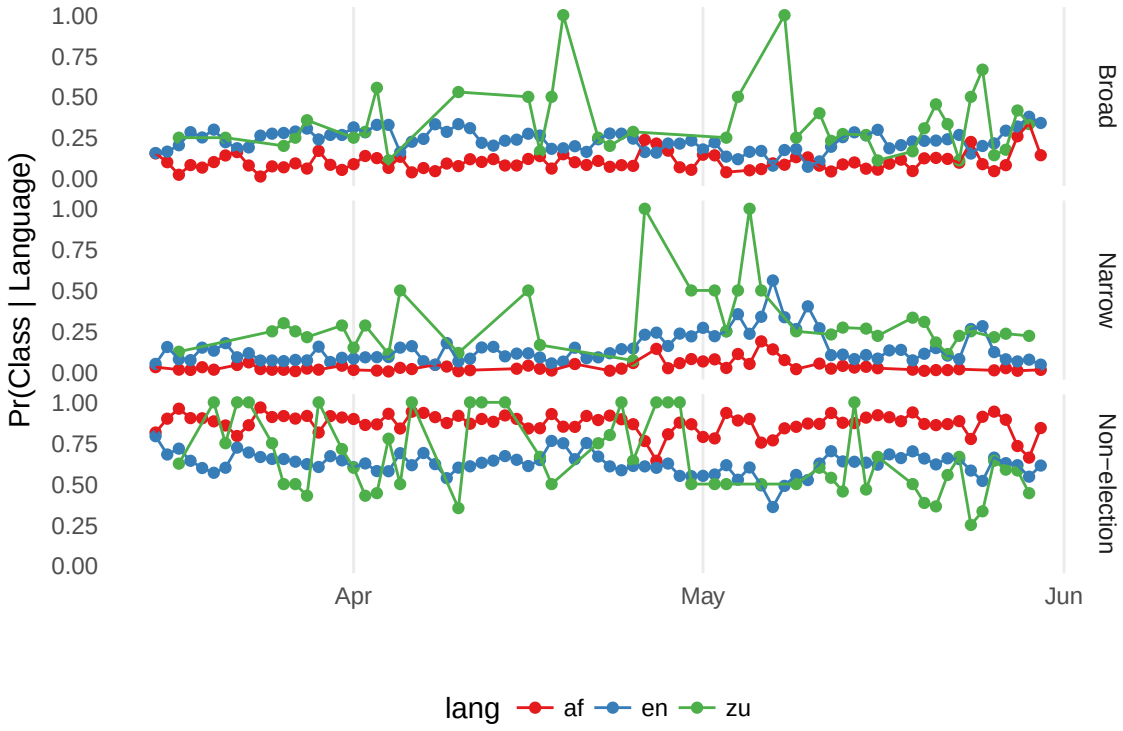

Figure 9: Proportion of topics by day

splitting the hashtag or username on capital letters or underscores, since many hashtags and usernames use either "snake_case" or "CamelCase" to combine words.[32] For example, the capitalization heuristic splits the username "@DANews" is into two tokens, "DA" and "News".

The other pre-processing steps consisted of translating all tweets into English, combining named entities into single tokens, keeping only content words, and removing words that appeared either too infrequently or too frequently. Since tweets were written in languages other than English, the language of articles and tweets was identified with the Python package `cld2-cffi`, and translated to English using the Google Cloud Translation API. Like the newspaper articles we tokenized and annotated Tweets with parts of speech and named entity recognition using SpaCy's English language model. Only content words, defined as a word with a part-of-speech of adjective, adverb, noun, or verb, were retained. Words were replaced by their lower-case lemmatized form, with the exception of numeric and time entities ("date", "time", "percen", "quantity", "ordinal", and "cardinal"), which were replaced with a token indicating the entity type. Words appearing in more than 10 documents or fewer than 80% documents were removed. After pre-processing, the corpus consisted of a vocabulary of 47,180 words, 99,404 documents, 22,944,652 tokens, and an average of 231 tokens per document.

We estimate a single 64-topic correlated topic model on this corpus of both news articles and tweets using the R package `stm` (Roberts, Stewart and Tingley 2016). Estimating a topic model on a corpus that includes both tweets and news articles places both newspaper articles and tweets in the same space so we can compare them using the same set of topics. However, it is not interesting to directly compare the distribution of topics between the news and Twitter subsets of the corpus, since we expect these distributions to be different since we include all news articles, whereas the tweets were selected such that they consistent almost entirely of politically related material. Thus, we will compare the distribution of topics in tweets to the distribution of topics in the subset of articles that are predicted to be in the narrow election-related category, and those predicted to be in either the narrow or broad election-related categories.[33] This ensures that we are comparing the political Twitter-sphere to the political coverage in the news, excluding all the other non-political stories in the news.

[32]Splitting hashtags into words on capital letters is a preprocessing step is used by to generating word vectors from Twitter (Pennington, Socher and Manning 2014*a*,*b*).

[33]More accurately, we compare the distribution of topics in tweets, to the distribution of topics in news-articles weighted by the probability of being election-related. As in Table 4, these predicted probabilities come from a Lasso regression of topic on the probability of being in each category.

Figure 10 visualizes the proportion of each topics appearing in three categories of documents: tweets, news articles in the narrow election-related category, and news articles in either the narrow or broad election-related categories. While we caution that these results in particular are preliminary, we notice some interesting patterns. First, there are many areas of substantive coherence—President Zuma's misuse of funds scandal was widely discussed in both traditional (narrow) newspaper coverage and political social media. We believe these early results are generally indicative of two political conversations that are relatively reflective of one another—meaning, what politicians and political entities are discussing directly does not diverge too significantly from the political coverage in the newspapers. What does seem to emerge is that differences seem particularly tied to issues that may be particularly complex for Twitter's character limitations (e.g., legislative alliance formations) but favors discussions involving well known politicians, but those who might fall below party leader of a major party. We plan to interrogate these questions more directly in future analysis.

6 Discussion and Conclusion

In this paper, we investigate media coverage of the campaign in an emerging democracy's election. Using nearly the universe of news reports with machine learning on a large corpus of events reported in traditional media and direct appeals from political actors through social media, our results indicate that how one conceives of an election at the outset plays a critical role in measurement and results. We show that even a narrow conception of election coverage provides richer details to the information environment beyond the racial and institutional cues generated from politicians' appeals, including regarding the performance of the incumbent government. We also show that a broad definition further provides news consumers with information about public policies and service provision outcomes. Direct appeals by politicians over social media largely confirm these results, demonstrating that media reports do not simply reporting random noise or systematic bias away from politicians' appeals, but rather reflect the information environment. This provides important evidence as to what information voters have when they cast ballots resulting from politicians' appeals.

South Africa provides important comparative insights towards understanding campaigns and elections in other emerging democracies. Conducting the study only in South Africa allows us to hold

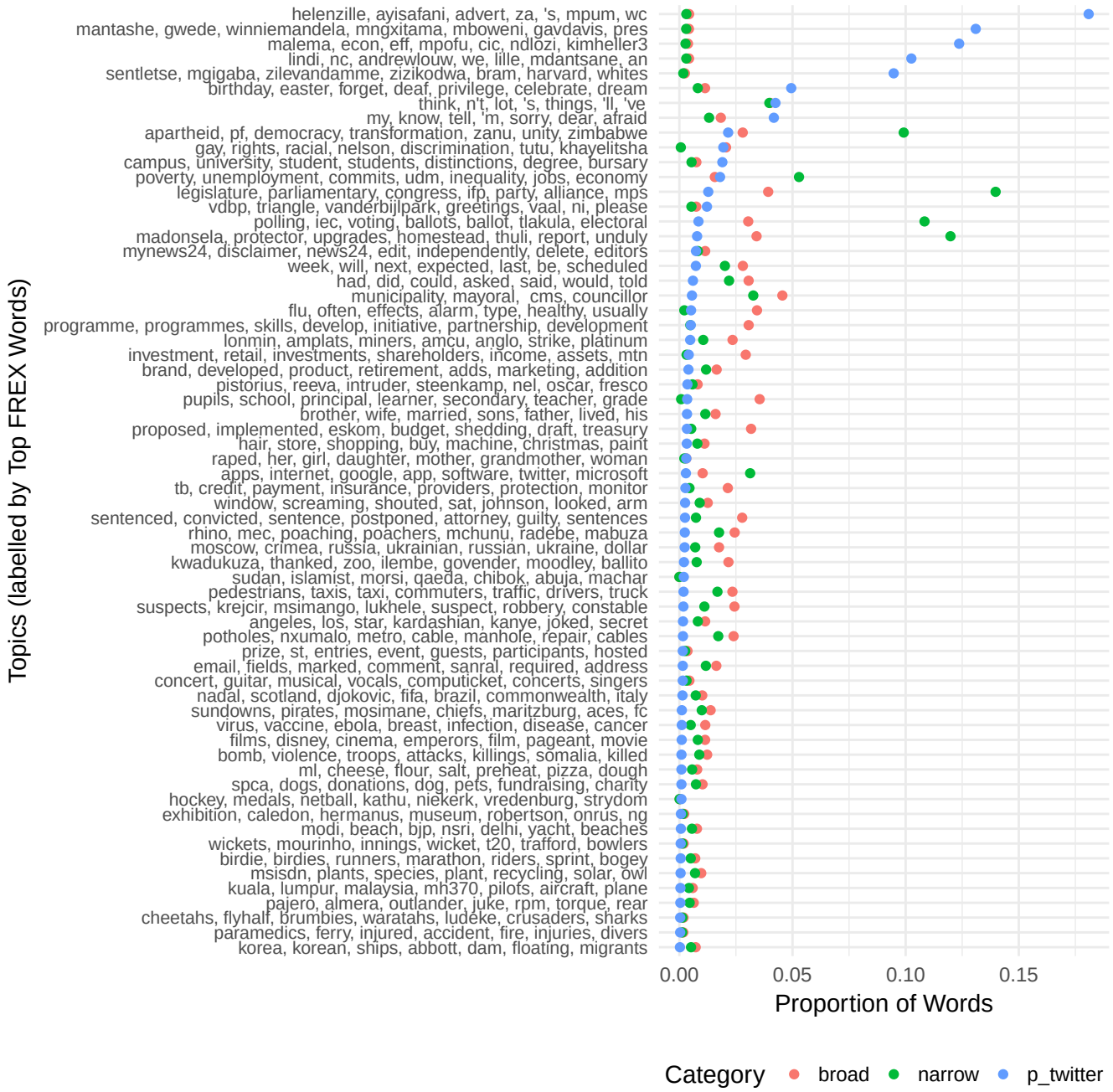

Figure 10: Estimated proportions of topics in Tweets, news articles in the narrow election-related category, and news articles in the narrow or broad election-related categories. The topics are from a 64-topic model estimated on a corpus including both news articles and tweets. News articles are classified into election-relatedness categories with an L1-regularized multinomial regression.

country-level and institutional factors constant. Similar to other countries where social identity is salient, electoral participation in South Africa is largely driven by race and characteristics of a dominant party system where competition may be constrained. This presents a harder case of the potential influence of campaigns. And similar to other emerging democracies, media consumption is increasing among the population and elections do not simply appear to reflect very basic levels of information, but rather suggest that media, politicians, and voters all look like more than just ethnicity and party matter. It is beyond our ability to say whether this is suggestive of greater consolidation as the campaign is becoming more "substantive," but we do note that since the timing of this election which resulted in the re-election of Jacob Zuma and an ANC majority, issues of corruption and poor government performance have plagued his administration and the ANC.[34] Whether and to what degree these issues may play a role in the upcoming 2019 election is yet to be seen, but our results given some indication that politicians can and do consider more than race and party appeals when seeking to gain, and retain, office.

Our study makes several contributions to diverse social science and data science literatures. First, we lend insights to studies of the role of campaigns towards understanding electoral outcomes in emerging democracies. Two contrasting views shape how scholars have treated these elections. In the first, the fundamentals of ethnic demography and electoral institutions suggest that campaigns for office rarely, if ever, matter in the sense of persuading voters. Media reports during the campaign are unlikely to provide voters with any new information that helps inform decision-making beyond turning out the base. Our results suggest that this view does not give a complete picture of the nuance and richness of the campaign as reflected in the media environment. We more closely accord with a second view, that politicians deploy diverse appeals that could also include attempts at persuasion so that their strategies are not just about playing towards ethnic demography or institutional factors, but also

[34]Zuma most recently, but barely, survived a no confidence vote in parliament in August 2017.

to persuade voters on substantive performance and policy issues.[35] While it is beyond the scope of our study to test whether reports in media directly influence voting behavior, our analysis does provide important evidence to suggest that campaigns, even narrowly defined, may be more substantive and nuanced than previously thought, at least insofar as the media reports on it both in traditional media and from direct appeals made by politicians over social media. We find that media reports suggest politicians do both – there is not necessarily an inherent trade-off at the media event level, regardless of whether one adopts a narrow or broad definition. Voters may be exposed to new information that could possibly influence their opinions by causing them to update their prior beliefs, in contrast to a literature that says African voters lack or do not care about this information.

Second, we support our conclusions with multiple data sources and techniques including media news reports, social media, and machine learning, supporting prior work that employs survey, observational, and experimental data. Divergent findings on the role of information (via media or campaigns) may arise at least partially as result of examining and testing hypotheses at different levels of measurement, including with public opinion surveys (Bratton and Kimenyi 2008, Bratton, Mattes and Gyimah-Boadi 2005, Ferree 2011, Gibson, Ferree and Long 2014, Hoffman and Long 2013, Long and Gibson 2015), observational electoral data (Arriola 2009, Chandra 2004, Posner 2005, Posner and Simon 2002), experimental manipulations (Bush et al. 2016, Conroy-Krutz and Moehler 2016), or politicians' speeches (Horowitz 2012). Whereas public opinion data show mixed results on new information mattering towards evaluations and therefore campaigns not mattering much, examinations of direct campaign strategy suggest that politicians may attempt persuasion even as they also turn out the base. And of course, not all that is reported in the media is purely a function of campaign strategies or what politicians intended—coverage includes many potentially negative stories that reflect poorly on people, or could be interpreted in that way.

[35] In so doing, we also join a larger literature on the political economy of campaigns and elections in industrialized democracies (e.g., Dalton and Wattenberg 2000, Powell and Whitten 1993) Here too, scholars divide on the question of whether, how, and to what degree media mediates campaign information and strategies, and the degree to which this matters towards voter information and evaluations. In the US, scholars have downplayed the importance of campaigns given the stability of party identification and "fundamentals" with respect to economic performance and incumbent party approval rating towards understanding electoral outcomes (Barro 1973, Fearon 1999, Ferejohn 1986, Fiorina 1981), this does not preclude the possibility that at least some voters some of the time are influenced by some campaign information, but rather complements it (Popkin 1994, Popkin et al. 1976). Moreover, scholars in American Politics also argue that American voters likely lack the requisite information to form evaluations of parties beyond their own partisanship and social identity (Achen and Bartels 2016). Understanding the role that both fundamentals and malleable things matters supports a holistic approach to studying elections and media in consolidating and consolidated democracies.

Third, we offer a discussion and procedure to create and analyze datasets of political events, emphasizing casting a wide net, particularly in developing contexts. We advance this by examining elections, a much harder class of events given potential for multiple interpretations that are difficult to observe and significantly more news sources. Importantly, our investigation has revealed that how one defines the coding itself is a product of one's theoretical orientation and classification scheme, suggesting multiple plausible pathways to studying elections with no agreed-upon definition. We provide insights on the initial qualitative coding of stories, and what this reveals about measurement more generally. One important theoretical and empirical implication from this is that context and case knowledge matter, and projects such as ours often require iteration between induction and deduction. That is, constructs of things (an "election event") and actual things (a rally held by President Zuma) that are pieces of data do not always perfectly match up. While this relates to issues of measurement fundamentally, it also may help, at least in part, explain mixed results from prior views in the literature that either campaigns do or do not matter to voters. That is, the literatue may suffer from some confirmation bias; that is, if scholars assume campaigns only matter to drive up co-ethnic or co-partisan support, it is not surprising that that is what they find covered in the news; or, if researchers think that other things matter, it is also not surprising that they find evidence consistent with their priors. Studies therefore should be very explicit at the outset in terms of what they are doing and how defining so as to be firmer on scientific and research design grounds, but also for replicability and external validity. Overall, when it comes to measurement and testing, big data and machine learning require an earlier attention to a part of the research design that social scientists typically ignore since we they tend to take measurement for granted following agreed upon ontologies of events and phenomena.

Last, we approach the topic of campaigns and elections using both supervised and unsupervised machine learning to help us gain insight into what constitutes election and campaign coverage. We believe these insights have broad application to machine learning and big data research addressing social science problems of classification and prediction (Alvarez 2015, Ruths and Pfeffer 2014), and more narrowly to such big data applications to the study of elections, including election outcomes (Mebane et al. 2017), voting behavior (Nickerson and Rogers 2014), media stories (Bagozzi and Schrodt 2012, Beieler 2016), politicians' statements and legislative text (Proksch and Slapin 2010, Spirling 2015,

Wilkerson, Smith and Stramp 2015), and campaign finance.

References

Achen, Christopher H and Larry M Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton: Princeton University Press. OCLC: 953028210.

Alexander, Peter. 2010. "Rebellion of the poor: South Africa's service delivery protests – a preliminary analysis." *Review of African Political Economy* 37(123):25–40.

Alvarez, R. Michael, ed. 2015. *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press.

Angelopulo, George and Petrus Potgieter. 2016. Media Ownership and Concentration in South Africa - Oxford Scholarship. In *Who Owns the World's Media?: Media Concentration and Ownership around the World*, ed. Eli M. Noam. Oxford U. P. pp. 987–1014.

Arriola, L. R. 2009. "Patronage and Political Stability in Africa." *Comparative Political Studies* 42(10):1339–1362.

Bagozzi, Benjamin E. and Philip A. Schrodt. 2012. "The Dimensionality of Political News Reports.". http://www.benjaminbagozzi.com/uploads/1/2/5/7/12579534/bagozzi.schrodt.epsa12.pdf

Banducci, Susan A. and Jeffrey A. Karp. 2003. "How Elections Change the Way Citizens View the Political System: Campaigns, Media Effects and Electoral Outcomes in Comparative Perspective." *British Journal of Political Science* 33(3):443–467.

Barkan, Joel D. 1976. *An African Dilemma: University students, development, and politics in Ghana, Tanzania, and Uganda*. New York: Oxford University Press.

Barro, Robert J. 1973. "The Control of Politicians: An Economic Model." *Public Choice* 14(1):19–42.

Bates, Robert. 1974. "Ethnic Competition and Modernization in Contemporary Africa." *Comparative Political Studies* 6(4):457–485.

Beieler, John. 2016. The Generation and Use of Political Event Data PhD thesis Pennsylvania State University.
https://etda.libraries.psu.edu/catalog/13347jub270

Blaydes, Lisa. 2011. *Elections and Distributive Politics in Mubarak's Egypt*. Cambridge: Cambridge University Press.

Blei, David M. and John D. Lafferty. 2009. Topic Models. In *Text Mining: Classification, Clustering, and Applications*, ed. A. Srivastava and M. Sahami. CRC Data Mining and Knowledge Discoverty Series Chapman & Hall.

Booysen, Susan. 2005. The Democratic Alliance: Progess and Pitfalls. In *Electoral Politics in South Africa: Assessing the First Democratic Decade.*, ed. Jessica Piombo and Lia Nijzink. Place of publication not identified: Palgrave Macmillan. OCLC: 951524863.

Bratton, Michael and Mwangi S. Kimenyi. 2008. "Voting in Kenya: Putting Ethnicity in Perspective." *Journal of Eastern African Studies* 2:272–289.

Bratton, Michael, Robert Mattes and Emmanuel Gyimah-Boadi. 2005. *Public opinion, democracy, and market reform in Africa*. Cambridge University Press.

Bush, Sarah Sunn, Aaron Erlich, Lauren Prather and Yael Zeira. 2016. "The Effects of Authoritarian Iconography." *Comparative Political Studies* 49(13):1704–1738.

Callen, Michael and James D. Long. 2015. "Institutional Corruption and Election Fraud: Evidence from a Field Experiment in Afghanistan." *American Economic Review* 105(1):354–381.

Callen, Michael Joseph, Saad Gulzar, Syed Ali Hasanain and Muhammad Yasir Khan. 2013. "The political economy of public employee absence: Experimental evidence from Pakistan." *Available at SSRN 2316245* .

Chandra, Kanchan. 2004. *Why ethnic parties succeed*. New York: Cambridge University Press.

Conroy-Krutz, Jeffrey and Devra C. Moehler. 2015. "Moderation from Bias: A Field Experiment on Partisan Media in a New Democracy." *Journal of Politics* 77(2):575–587.

Conroy-Krutz, Jeffrey and Devra C. Moehler. 2016. "Eyes on the ballot: Priming effects and ethnic voting in the developing world." *Electoral Studies* 42:99–113.

Cox, G. 1997. *Making votes count: strategic coordination in the world's electoral systems*. Cambridge: Cambridge University Press.

Cox, Gary W. 2015. "Electoral Rules, Mobilization, and Turnout." *Annual Review of Political Science* 18(1):49–68.

Dalton, Russell J. and Martin P. Wattenberg, eds. 2000. *Parties Without Partisans*. New York: Oxford University Press.

Davenport, Christian and Will H. Moore. 2015. "Conflict Consortium Standards & Best Practices for Observational Data." Working Paper.
http://conflictconsortium.weebly.com/uploads/1/8/3/5/18359923/cc-datastandardspractices7apr2015.pdf

Dawson, Michael C. 1994. *Behind the Mule: Race and Class in African-American Politics*. Princeton, NJ: Princeton University Press.

de Kadt, Daniel. forthcoming. "Voting then, voting now: The long term consequences of participation in South Africa's first democratic election." *Journal of Politics* .

de Kadt, Daniel and Melissa Sands. n.d. "Segregation drives racial voting: New evidence from South Africa." Working Paper.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper & Row.

Eyal, Adi. 2016. "jrnold/various-scrapers v1.0.0.".
https://doi.org/10.5281/zenodo.208322

Fearon, James D. 1999. Electoral Accountability and the Control of Politicians: Selecting Good Types Versus Sanctioning Poor Performance. In *Democracy, Accountability, and Representation*, ed. Adam Przeworski, Susan C. Stokes and Bernard Manin. Cambridge University Press pp. 55–61.

Ferejohn, John. 1986. "Incumbent Performance and Electoral Control." *Public Choice* 50(1/3):5–25.

Ferree, Karen. 2011. *Framing the race in South AFrica: the political origins of racial census elections*. Cambridge University Press.

Ferree, Karen E. 2006. "Explaining South Africa's Racial Census." *Journal of Politics* 68(4):803–815.

Ferree, Karen E. 2010. "The social origins of electoral volatility in Africa." *British Journal of Political Science* 40(4).

Ferree, Karen and James D. Long. 2016. "Gifts, Threats, and the Secret Ballot in Africa." *African Affairs* forthcoming.

Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections*. New Haven, CT: Yale University Press.

Friedman, Steven. 2004. "Why we vote: the issue of identity." *Election Synopsis* 1(2):2–4.

Garcia-Rivero, Carlos. 2006. "Race, Class and Underlying Trends in Party Support in South Africa." *Party Politics* 12(1):57–75.

Gibson, Clark C., Karen E. Ferree and James D. Long. 2014. "Voting behavior and electoral irregularities in Kenya's 2013 Election." *Journal of Eastern African Studies* 8(1):153–172.

Habib, Adam and Naidu Sanusha. 2006. "Race, Class and Voting Patterns in South Africas Electoral System: Ten Years of Democracy." *Africa Development* 31(3):81–92.

Hoffman, Barak D and James D Long. 2013. "Parties, Ethnicity, and Voting in African Elections." *Comparative Politics* 45:127–146.

Holmes, Carolyn E. 2015. "Marikana in Translation: Print Nationalism in South Africa's Multilingual Press." *African Affairs* 114(455):271–294.

Horowitz, Donald. 1985. *Ethnic groups in conflict*. Berkeley: University of California Press.

Horowitz, Jeremy. 2012. Campaigns and Ethnic Polarization in Kenya PhD thesis University of California, San Diego.

Horowitz, Jeremy and James Long. 2016. "Strategic voting, information, and ethnicity in emerging democracies: Evidence from Kenya." *Electoral Studies* 44:351–361.

Hyde, Susan D. 2011. *The Pseudo-Democrat's Dilemma: Why Election Monitoring Became an International Norm.* Ithaca, NY: Cornell University Press.

Johnson, R.W. and Lawrence Schlemmer, eds. 1996. *Launching Democracy in South Africa: the first open election, April 1994.* Yale University Press.

Kasara, K. and P. Suryanarayan. 2015. "When do the Rich Vote Less than the Poor and Why?: Explaining Turnout Inequality across the World." *American Journal of Political Science* 59(3):613–627.

Keefer, Philip and Razvan Vlaicu. 2008. "Democracy, Credibility, Clientelism." *The Journal of Law, Economics, and Organization* 24(2):371–406.

Kuenzi, Michelle and Gina MS Lambright. 2010. "Who votes in Africa? An examination of electoral participation in 10 African countries." *Party Politics* p. 1354068810376779.

Levitsky, Steven and Lucan A. Way. 2010. *Competitive Authoritarianism: Hybrid Regimes after the Cold War.* Cambridge: Cambridge University Press.

Lindberg, Staffan and Minion K.C. Morrison. 2008. "Are African Voters Really Ethnic or Clientelistic? Survey Evidence from Ghana." *Political Science Quarterly* 123(1):95–122.

Lodge, T. 1983. *Black politics in South Africa since 1945.* Longman Publishing Group.

Lodge, Tom. 1995. "The South African General Election, April 1994: Results, Analysis and Implications." *African Affairs* 94(377):471–500.

Lodge, Tom. 2004. The African National Congress and Its Allies. In *Election 94 South Africa: The Campaigns, Results and Future Prospects*, ed. Andrew Reynolds. New York: St. Martin's Press.

Lodge, Tom and Bill Nasson. 1991. *All, here, and now: Black politics in South Africa in the 1980s.* New Africa Books.

Long, James D. and Clark C. Gibson. 2015. "Evaluating the Roles of Ethnicity and Performance in African Elections." *Political Research Quarterly* 68(4):830–842.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2):254–277.

Lupia, Arthur and Matthew D. McCubbins. 1998. *The Democratic Dilemma: Can citizens learn what they need to know?* Cambridge: Cambridge University Press.

Magaloni, Beatriz. 2006. *Voting for Autocracy: Hegemonic Party Survival and its Demise in Mexico.* Cambridge: Cambridge University Press.

Magaloni, Beatriz. 2010. "The Game of Electoral Fraud and the Ousting of Authoritarian Rule." *American Journal of Political Science* 54(3):751–765.
http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2010.00458.x/full

Mattes, Robert. 2012. "The Born Frees: The Prospects for Generational Change in Post Apartheid South Africa." *Australian Journal of Political Science* 47(1).

Mattes, Robert, Amanda Gouws and Hennie Kotze. 1995. "The Emerging Party System in the New South Africa." *Party Politics* 1(3):381–395.

Mattes, Robert and Jessica Piombo. 2001. "Opposition Parties and the Voters in South Africas General Election of 1999." *Democratization* 8(3):101–128.

Mattes, Robert and Roger Southall. 2004. "Popular Attitudes Toward the South African Electoral System." *Democratization* 11(1):51–76.

Mebane, Walter R., Jr., Alejandro Pineda, Logan Woods, Joseph Klaver, Patrick Wu and Blake Miller. 2017. "Using Twitter to Observe Election Incidents in the United States." Prepared for presentation at the 2017 Annual meeting of the Midwest Political Science Association.

Mehrotra, Rishabh, Scott Sanner, Wray Buntine and Lexing Xie. 2013. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *Proceedings of the 36th International*

ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '13 New York, NY, USA: ACM pp. 889–892.

Mozaffar, Shaheen, James R. Scarritt and Glen Galaich. 2003. "Electoral Institutions, Ethnopolitical Cleavages, and Party Systems in Africa's Emerging Democracies." *American Political Science Review* 97(3):379–390.

Ndlovu, Musa. 2011. "The meaning of post-apartheid Zulu media." *Communicatio* 37(2):268–290.

Nickerson, David W. and Todd Rogers. 2014. "Political Campaigns and Big Data." *Journal of Economic Perspectives* 28(2):51–74.

Ordeshook, Peter C. and Olga V. Shvetsova. 1994. "Ethnic Heterogeneity, District Magnitude, and the Number of Parties." *American Journal of Political Science* 38(1):100–123.

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014*a*. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543.
http://www.aclweb.org/anthology/D14-1162

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014*b*. "preprocess-twitter.rb.".
https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

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

Popkin, S., J. Gorman, C. Phillips and J. Smith. 1976. "Comment: What Have You Done for Me Lately? Toward an Investment Theory of Voting." *American Political Science Review* 70(30):779–805.

Posner, D. 2005. *Institutions and ethnic politics in Africa*. Cambridge: Cambridge University Press.

Posner, Daniel N. and David J. Simon. 2002. "Economic conditions and incumbent support in Africa's new democracies." *Comparative Political Studies* 35(3):313–336.

Powell, G. Bingham and Guy D. Whitten. 1993. "A Cross-National Analysis of Economic Voting: Taking Account of the Political Context." *American Journal of Political Science* 37(2):391–414.

Proksch, Sven-Oliver and Jonathan B. Slapin. 2010. "Position Taking in European Parliament Speeches." *British Journal of Political Science* 40(03):587–611.

Rabushka, A. and K. A. Shepsle. 1972. *Politics in plural societies*. Stanford University Press.

Reeves, Andrew, Stephen Shellman and Brandon Stewart. 2006. "Fair & balanced or fit to print? The effects of media sources on statistical inferences." Working Paper.
http://scholar.princeton.edu/sites/default/files/bstewart/files/occassional0.pdf

Reilly, Benjamin. 2016. Timing and Sequencing in Post-Conflict Elections. In *Building Sustainable Peace: Timing and Sequencing of Post-Conflict Reconstruction and Peacebuilding*, ed. Arnim Langer and Graham K. Brown. New York: Oxford University Press pp. 72–86.

Reynolds, Andrew. 1999. *Electoral Systems and Democratization in Southern Africa*. London and New York: Oxford University Press.

Reynolds, Andrew and Ben Reilly. 1997. *The International IDEA Handbook of Electoral System Design*. Sweden: International Institute for Democracy and Electoral Assistance.

Reynolds, Andrew, ed. 1994. *Election '94 South Africa: the campaigns, results and future prospects*. St. Martin's Press.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2016. *stm: R Package for Structural Topic Models*. R package version 1.1.3.
http://www.structuraltopicmodel.com

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi et al. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Ruths, Derek and Jrgen Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346(6213).

Scheiner, Eric and Robert G. Moser. 2004. "Mixed electoral systems and electoral system effects: controlled comparison and crossnational analysis." *Electoral Studies* 23(4):575–599.

Schrodt, Philip A. 2012. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38(4):546–569.

Snyder, Jack. 2000. *From voting to violence: Democratization and nationalist conflict.* New York: Norton.

Southall, Roger. 2014*a*. "The South African Election of 2014: Retrospect and Prospect." *Strategic Review for Southern Africa* 36(2).

Southall, Roger. 2014*b*. "Zuma: party leadership as electoral liability." *Nelson Mandela and the political economy of unfinished liberation* http://www.tandf.co.uk/journals/pdf/spissue/crea-si-southall.pdf.

Spirling, A. 2015. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *Journal of Politics* 78(1):120–136.

Stokes, Susan C. 2005. "Perverse Accountability: A Formal Model of Machine Politics with Evidence from Argentina." *The American Political Science Review* 99(3):315–325.

Tomaselli, Keyan. 1997. "Ownership and control in the South African print media: black empowerment after apartheid, 1990–1997." *Ecquid Novi* 18(1):67–68.

Wilkerson, John, David Smith and Nicholas Stramp. 2015. "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach." *American Journal of Political Science* 59(4):943–956.

Wilkinson, Steven I. 2004. *Votes and Violence.* Cambridge: Cambridge University Press.

World Bank. 2014. "Mobile cellular subscriptions.". data retrieved from World Development Indicators, http://data.worldbank.org/indicator/IT.CEL.SETS.P2.

	Dec 2014		Other Info		
	Readers (000s)	Pct	In Sample	Language	Reason
Beeld	416	1.10	No	Afrikaans	Paywall
Die Burger	502	1.30	No	Afrikaans	Paywall
Business Day	79	0.20	No	English	TMG
Cape Argus	421	1.10	Yes	English	
Cape Times	235	0.60	Yes	English	
The Citizen	391	1.00	Yes	English	
Daily Dispatch	215	0.60	Yes	English	
Daily News	289	0.80	Yes	English	
Daily Sun	5256	14.00	Yes	English	
Daily Voice	516	1.40	No	English	PDF only
Diamond Fields Advertiser	81	0.20	No	English	PDF only
Herald	194	0.50	No	English	TMG
Isolezwe	1180	3.10	Yes	isiZulu	
Mercury	185	0.50	Yes	English	
New Age (Tna)	130	0.30	No	English	
Pretoria News	141	0.40	Yes	English	
Die Son	1035	2.70	Yes	Afrikaans	
The Sowetan	1655	4.40	No	English	T&C
The Star	598	1.60	Yes	English	
The Times	340	0.90	No	English	TMG
Volksblad	123	0.30	No	Afrikaans	Paywall
The Witness	84	0.20	Yes	English	
Any "Amps" Daily Newspaper	10773	28.60			

Table 1: Newspaper readership of major South African Newspapers. Data from the South African Advertising Research Foundation

Publication Name	No. of Articles	Language	Days in Corpus
News24	13125	English	87
Independent Online	6000	English	109
The Citizen	4406	English	73
Tame Times	4081	English	103
City Press	2992	English	86
Business Report	2555	English	73
Pretoria News	2347	English	84
VaalWeekblad	1655	Afrikaans	95
Worcester Standard	1551	Afrikaans	63
Daily News	1466	English	45
Mail & Guardian	1399	English	89
Eikestad Nuus	1317	Afrikaans	84
Daily Sun	1136	English	88
The Witness	1136	English	49
Weslander	1130	Afrikaans	194
Kalahari Bulletin	1100	Afrikaans	92
Boland Gazette	1073	Afrikaans	125
Paarl Post	1003	Afrikaans	53
Potchefstroom Herald	991	Afrikaans	79
Tygerburger	982	English	107

Table 2: Top 20 publications in the corpus, by number of articles. There are 99,166 articles total in the corpus.

Language	Broad	Narrow	Non-election
Afrikaans	0.11	0.02	0.87
English	0.26	0.13	0.62
isiZulu	0.25	0.16	0.60

Table 3: Proportion of articles in each political coverage category (Broad, Narrow, Non-election) aggregated by the language of the article (Afrikaans, English, Zulu).

7 Appendix

Category	Topic	Labels	$\beta_{k,j}$	$\Pr_{k,j}$
Narrow	27	voting, eff, ballot, malema, ifp, nfp, bekkersdal	11.71	1.00
	6	zuma, madonsela, mantashe, nkandla, protector, thuli, gigaba	4.17	0.95
	25	poverty, inequality, unemployment, empowerment, udm, socio, rural	3.60	0.82
	2	advert, sanral, icasa, apps, android, allenby street, smartphones	1.51	0.28
Broad	18	bursary, classroom, pupil, scholarship, university, campus, academic	2.70	0.84
	16	sars, financial, insurance, payment, regulation, proposal, tax	2.66	0.84
	7	hiv, virus, vaccine, ebola, diabetes, epilepsy, sanbs	2.00	0.72
	14	solar, shedding, geyser, flood, rainfall, flooding, turbine	1.75	0.67
	32	amcu, miner, lonmin, amplats, association of mineworkers, impala platinum, r12,500	0.94	0.76
	23	yen, s&p, pfizer, astrazeneca, economist, alibaba, currency	0.68	0.52
	5	krejcir, msimango, lukhele, parole, nyaope, derby-lewis, siboniso miya	0.66	0.41
	24	ochakvaite, mayoral, blog't, bet365.gr, cms, manhole, skoro	0.26	0.81
Non-election	19	rider, championships, karate, agn, paddler, dusi, javelin	2.55	0.96
	17	boko haram, kiev, ukrainian, obama, islamist, putin, separatist	2.24	0.95
	3	sharks, stormers, super rugby, brumbies, crusaders, flyhalf, waratahs	1.39	0.88
	15	auntie, grandmother, claus, son, daily sun, father, marry	1.33	0.88
	11	concert, jazz, guitar, orchestra, songwriter, vocalist, guitarist	1.20	0.86

20	worcester, vredenburg, fundraising, weslander, langebaan, hermanus, west coast	0.77	0.80
31	liverpool, chelsea, sundowns, midfielder, goalkeeper, psl, bafana	0.51	0.76
28	ml, cheese, flour, tomato, salad, lemon, compost	0.13	0.69
23	yen, s&p, pfizer, astrazeneca, economist, alibaba, currency	-0.50	0.40
25	poverty, inequality, unemployment, empowerment, udm, socio, rural	-0.51	0.11
32	amcu, miner, lonmin, amplats, association of mineworkers, impala platinum, r12,500	-1.63	0.15
12	vavi, gay, cosatu, sadtu, allegation, disciplinary, ruling	-2.63	0.12
6	zuma, madonsela, mantashe, nkandla, protector, thuli, gigaba	-5.07	0.00
24	ochakvaite, mayoral, blog't, bet365.gr, cms, manhole, skoro	-6.27	0.00

Table 4: Non-zero coefficients for a multinomial regression with lasso regularization of the election-relatedness category on topics from the 32-topic correlated topic model with the labelled subset of documents in the corpus. $\beta_{k,j}$ is the lasso regression coefficient for category j and topic k. $\Pr_{k,j}$ is the probability that a document is in category j given it consists only of tokens from topic k. It is calculated as $\Pr_{k,j} = \Pr(\text{category} = j|\theta_k = 1) = \exp(\alpha_j + \beta_{j,k})/\sum_{j' \in J} \exp(\alpha_{j'} + \beta_{j',k})$, where α_j is the intercept for category j, and J is the set of categories: narrow, broad, non-election. The topics are sorted by category and descending size of their coeffient. Topics with coefficients of zero are not included.

Category	Article
Broad	CAUTIONARY notices issued in the same week by packaging group Bowler Metcalf and dairy brands producer Clover have caused a speculative fizz in the market. Bowcalf specialises in plastics packagin... ("THE INVESTMENT WEEK: Bowler Metcalf" *Financial Mail*, 2014-03-13)
	"She's been offered a place in perhaps the most prestigious programme in politics anywhere, which is a year masters' in public administration at Harvard," she told the broadcaster. The programme at... ("Mazibuko taking sabbatical: Zille" *News24*, 2014-05-11)
	Annah Monamodi of ADHASA Soweto encourages the community to attend upcoming meeting. Residents of Soweto are invited by the Attention Deficit and Hyperactivity Support Group of South Africa (ADHASA... ("" *West Side Urban*, 2014-05-09)
	Ballito is to be the second South African town to have its own nuclear power plant. Government gazetted this extreme solution to Ballito's on-going electricity problems on Tuesday, April 1. Followi... ("Nuclear solution to Ballitos electricity problems" *North Coast Courier*, 2014-04-01)
	BONTEHEUWEL residents complained about the shortage and visibility of police and said gang shootings and other crimes could be prevented if policing was effective. During a public meeting in Bonteh... ("Police losing war on crime - residents" *Cape Times*, 2014-05-05)

NELSPRUIT The Mpumalanga Department of Public Works, Roads and Transport (DPWRT) is urging taxi, bus and any other public transport operators to apply for temporary operating licenses, if they wi... (“” *Mpumalanga News*, 2014-04-11)

Poverty is the greatest challenge to effective HIV-AIDS treatment in Saldanha Bay. Sunday, December 1 was World AIDS Day. The media fuss over the country's shortage of antiretroviral drugs, but lo... (“Armoede knel Vigsstryd aan die Weskus” *Weslander*, 2013-12-06)

THE executive manager at the Zululand Chamber of Commerce and Industry (ZCCI), Charmayne Pountney, has been suspended with immediate effect. A letter on behalf of the ZCCI President, Sizwe Khumalo,... (“Chamber manager suspended” *Ulundi Fever*, 2014-01-30)

The SA Reserve Bank (Sarb) is keeping its key lending rate on hold at 5.5%, said governor Gill Marcus on Thursday. The repo rate is the interest rate at which the Sarb lends money to commercial ba... (“Reserve Bank keeps repo rate stable” *City Press*, 2014-03-27)

UIP co-ordinating manager Clare Swithenbank-Bowman said all residential and commercial property owners in the UIP nodes would be encouraged to come on board for the project to be a success. She sai... (“Ballito debates UIP model” *Ballito Fever*, 2014-01-17)

Narrow

"I am happy that our party is growing stronger on a daily basis and this serves as a boost for us as we move closer towards national elections," she said in a statement. "Contrary to what the proph... ("NFP welcomes new ECape members" *The Citizen*, 2014-04-17)

"You don't go boxing and expect to lose. You only get into the boxing ring when you can win," he said at the Election Result Operation Centre in Pretoria yesterday. Malema took a tour of the 11 000... ("Juju and party ready for an upset" *The Citizen*, 2014-06-05)

COSATU general secretary Zwelinzima Vavi abandoned the ANC's "good story" script at Cosatu's May Day rally and instead spoke more uncomfortable truths. He talked about a troubled tripartite allianc... ("Vavi dumps speech to speak out openly" *Daily Dispatch*, 2014-05-02)

Johannesburg - An application to compel IEC chairwoman Pansy Tlakula to resign should be dealt with before the May 7 elections, the Electoral Court sitting in Johannesburg heard on Tuesday. "Based ... ("Tlakula resignation bid urgent: lawyer" *Independent Online*, 2014-04-29)

Johannesburg - Two Ocean View residents were handed the keys to their innovative new homes by City of Cape Town Mayor Patricia de Lille on Tuesday. The two stone-clad houses are the first in a proj... ("Cape Town mayor hands over keys to new homes" *News24*, 2014-04-22)

MyNews24 is a user-generated section of News24.com . The stories here come from users. So the IEC say voting is free and fair. One must wonder how they come to this conclusion, when there are voter... (“IEC fair?” *News24*, 2014-03-26)

Pres. Jacob Zuma’s now infamous architect has its own mini-Nkandla building - presumably due to the R16 million he earned from his work for the president. Minenhle Makhanya his pig farm in the lush... (“Zuma se argitek bou sy eie mini-Nkandla” *Rapport*, 2014-03-22)

RUSTENBURG A so-called Marikana ”hit-list” is allegedly being used by rival unions as an underhanded tactic to smear the Association of Mineworkers and Construction Union (Amcu) and paint it as a... (“Marikana hit-list a smear campaign?” *Herald’s Bonus*, 2014-05-19)

THE Umzimkhulu Local Municipality mayor, Mphuthumi Mpabanga has described the Economic Freedom Fighter as ”ill-informed opportunists”. This comes after party members marched to the municipality off... (“Opposition party hands over list of demands” *East Griqualand Fever*, 2014-04-03)

Thousands of DA supporters are expected to take to the streets of Joburg to march against corruption and unemployment. ”We will march for support for entrepreneurs, internships for young people, ti... (“” *Sandton Chronicle*, 2014-04-23)

Non-election

Cape Town - The Cheetahs will soon be able to call on the services of Springbok flank Heinrich Brssow.According to the website , Brssow joined up with his team-mates in Bloemfontein on Monday aft... ("Brssow to boost Cheetahs" *News24*, 2014-04-03)

Christina Makhanani Baloyi, 20, would remain in police custody until her second appearance on Friday, Brigadier Hangwani Mulaudzi said. The matter was postponed for further investigation. Police fo... ("Limpopo woman appears over baby" *The Citizen*, 2014-05-19)

Creating your profile will enable you to submit photos and stories to get published on News24. This username must be unique, cannot be edited and will be used in the URL to your profile page across... ("8 primary school pupils stabbed in China" *News24*, 2014-05-20)

Did you know that there are children who never drink the opportunity to see rain or fresh water? Fresh water is very valuable for the planet and yet we are wasting so much of this precious liquid b... ("Vinnige feite om water te bespaar" *Potchefstroom Herald*, 2014-03-18)

Jessica Alba may be a good businesswoman, but her husband was not crazy about the idea. Radaronline.com reports The Honest Company, Jessica's eco-friendly company does better than she ever imagined... ("Jessica se Cash word deur haar sakevernuf bedreig" *Tame Times*, 2014-04-07)

MORE than 450 pupils at Forestview Primary in Waterfall signed up for the "My Little Fingers" project last Tuesday, The project involved taking pupils' fingerprints and personal information to make... ("Little Fingers drive starts" *Hillcrest Fever*, 2014-05-25)

OOM HERRIE read last week from an Aussie who woke up from earache. It turned out to be a cockroach that crawled the night in the landing. "Typically Australian," said Uncle Harry's cousin, Looking ... ("Draadjies tussen ore maar bedenklik" *Vrystaat News*, 2014-01-23)

The case against a girl who was arrested for allegedly stabbing a pupil was closed in court on March 26. Last month, the EXPRESS reported on a Kensington school pupil (15) who allegedly stabbed a "... ("" *Joburg East Express*, 2014-04-11)

THE sounds of family members cheering each other reverberated across the Empangeni Sports Club last weekend, as children from Oompaloompa Nursery School and their supportive family members took to ... ("Fun (walk) in the sun" *Zululand Observer*, 2014-03-26)

Thursday, 20 February: Ryan O'Connor, a PhD student at the University of Pretoria, will speak on Temperature Adaptations of the Rufous-cheeked Nightjar in an arid environment at a public meeting of... ("Dagboek: 19 Februarie 2014" *Noordkaap*, 2014-02-19)

Table 5: For each catgory, 10 randomly selected documents.

Topic	Top words in topic
1	FREX: commuter, oncoming, taxi, robber, metrorail, paramedic, intersection
	Lift: pnw, rusa, emrs, wiid, tvnw, sbw, unroadworthy
	Score: driver, robbery, motorist, taxi, accident, traffic, paramedic
	prob driver, steal, accident, traffic, taxi, bus, injure
2	FREX: advert, sanral, icasa, apps, android, allenby street, smartphones
	Lift: touchcard, wechat, vpc, cluedapp, vusi mona, john clarke, mark zuckerberg
	Score: advert, sabc, user, e, icasa, sanral, online
	prob e, user, online, facebook, twitter, toll, sabc
3	FREX: sharks, stormers, super rugby, brumbies, crusaders, flyhalf, waratahs
	Lift: cheika, wallabies, johan goosen, wallaby, flanker, tim swiel, elton jantjies
	Score: sharks, stormers, rugby, cheetahs, super rugby, lions, bulls
	prob rugby, sharks, score, stormers, coach, lions, cheetahs
4	FREX: pajero, torque, almera, juke, outlander, alloy, rpm
	Lift: vgt, dci, awd, rpm, scubi, ecosport, winckelmann
	Score: pajero, engine, wheel, corolla, diesel, almera, outlander
	prob model, wheel, engine, speed, seat, rear, fuel
5	FREX: krejcir, msimango, lukhele, parole, nyaope, derby-lewis, siboniso miya
	Lift: lamoer, kgopa, masunga, vollgraaff, marthella, mthabela, treurnicht
	Score: sentence, magistrate, drug, krejcir, rape, rap, bail
	prob drug, sentence, victim, assault, trial, rape, prison
6	FREX: zuma, madonsela, mantashe, nkandla, protector, thuli, gigaba
	Lift: scopa, ncop, mahlobo, ndala, mabuyane, trevor manuel, chabane
	Score: zuma, nkandla, madonsela, parliament, jacob zuma, protector, democracy
	prob zuma, nkandla, parliament, deputy, madonsela, jacob zuma, upgrade
7	FREX: hiv, virus, vaccine, ebola, diabetes, epilepsy, sanbs
	Lift: resveratrol, hospicewits, raju, piot, hypertension, homeopathy, ovarian
	Score: disease, cancer, patient, hiv, medical, clinic, treatment
	prob medical, patient, treatment, blood, cancer, disease, doctor
8	FREX: news24, mynews24, disclaimer, news24.com, pet, animal, owl
	Lift: waste group, gerhard van rooyen, tekkie tax day, laurence kingston, colleen downs, 24.com, url

Score: news24, animal, dog, mynews24, spca, bird, article

prob dog, animal, news24, article, reserve, editor, letter

9 FREX: though, certainly, hardly, probably, idea, obvious, st dunstan's

Lift: frighteningly, st dunstan's, brownell, auntie muriel's, rb7, sabine schmitz, blithering

Score: holiday, bit, college easter rugby festival, the_citizen_reporter, idea, st dunstan's, certainly

prob idea, quite, probably, history, bit, certainly, holiday

10 FREX: actress, liberace, siba, roxmouth, disney, bieber, artwork

Lift: qf1, bananagrams, set!jake, siba, liberace, disney junior play, menswear

Score: film, exhibition, art, movie, artist, actor, comedy

prob film, art, movie, exhibition, design, artist, fashion

11 FREX: concert, jazz, guitar, orchestra, songwriter, vocalist, guitarist

Lift: germiston simmer rugby club, kagiso rugby club, ub40, tamia, armin, cortes, innibos

Score: music, festival, album, ticket, song, concert, artist

prob music, ticket, festival, song, band, dance, album

12 FREX: vavi, gay, cosatu, sadtu, allegation, disciplinary, ruling

Lift: ntola, abland, mase, lyc, soobrayan, seakhoa, aboobaker

Score: cosatu, vavi, allegation, legal, numsa, corruption, commission

prob legal, allegation, commission, cosatu, letter, committee, application

13 FREX: pope, jesus, disciple, heaven, god, snake, holy

Lift: colton, ku, nga, deity, krishna, psalm, esv

Score: god, jesus, church, lord, prayer, pray, snake

prob god, church, remember, actually, wrong, jesus, truth

14 FREX: solar, shedding, geyser, flood, rainfall, flooding, turbine

Lift: sherpas, malca-amit, turbine, oberholtzer, cng holdings, heaviside, peaker

Score: eskom, electricity, rain, gas, plant, energy, river

prob energy, rain, eskom, plant, river, electricity, damage

15 FREX: auntie, grandmother, claus, son, daily sun, father, marry

Lift: claus, nolwazi, chanika, lovey, yolanda, vekneshan, sunteam

Score: girl, father, daughter, son, boy, husband, baby

prob girl, son, father, boy, daughter, couple, husband

16 FREX: sars, financial, insurance, payment, regulation, proposal, tax

Lift: mayman, hillen, surapure, eea2, estina, taxable, policyholder

Score: financial, tax, debt, budget, payment, income, consumer

prob financial, tax, property, account, budget, grant, credit

17 FREX: boko haram, kiev, ukrainian, obama, islamist, putin, separatist

Lift: makaburi, renamo, kidal, musharraf, yingluck, thaksin, putin

Score: ukraine, russia, russian, military, kiev, crimea, moscow

prob ukraine, russia, military, us, russian, war, nigeria

18 FREX: bursary, classroom, pupil, scholarship, university, campus, academic

Lift: bursary, gofpep, icollege, deneil, ssip, pedagogy, toastmasters

Score: student, pupil, education, teacher, learner, university, grade

prob student, pupil, education, teacher, study, learner, skill

19 FREX: rider, championships, karate, agn, paddler, dusi, javelin

Lift: afs, hooters, kruiskamp, stithians, vettel, jbay, javelin

Score: race, athlete, medal, km, championship, tournament, rider

prob race, km, prize, compete, athlete, category, junior

20 FREX: worcester, vredenburg, fundraising, weslander, langebaan, hermanus, west coast

Lift: john fry, worcester, luxolo, christmas fund, pioneer house, tcc, koerantskenking

Score: donation, donate, worcester, church, stellenbosch, sponsor, farm

prob donation, donate, sponsor, church, farm, stellenbosch, hall

21 FREX: plane, malaysia, malaysian, mh370, kuala lumpur, aviation, malaysia airlines

Lift: najib, faa, capsize, airbus, jeju, incheon, snohomish county

Score: plane, aircraft, malaysian, search, malaysia, mh370, flight

prob search, plane, flight, rescue, ship, aircraft, pilot

22 FREX: vaal weekly, msisdn, vaal triangle, vdbp, ni, meyerton, vereeniging

Lift: wiseguy, vaal weekly, tunny, danki, sunbabe, vaal mall, portholes

Score: thank, msisdn, vaal weekly, vanderbijlpark, lady, vaal, vereeniging

prob thank, lady, thanks, message, please, mr., sms

23 FREX: yen, s&p, pfizer, astrazeneca, economist, alibaba, currency

Lift: pfizer, repo, rbs, astrapak, eurobond, avi, premarket

Score: percent, investor, growth, index, economy, price, investment

prob percent, price, growth, rate, africa, bank, economy

24 FREX: ochakvaite, mayoral, blog't, bet365.gr, cms, manhole, skoro

Lift: uip, mfunda, mqwathi, iversen, clr da rocha, orlando ekhaya, mohapi

Score: municipality, councillor, municipal, ward, mayor, council, metro

prob municipality, council, municipal, ward, mayor, building, land

25 FREX: poverty, inequality, unemployment, empowerment, udm, socio, rural

Lift: soes, underdevelopment, manufacturing indaba, depoliticise, audits, ignoble, partite

Score: udm, sector, economy, policy, unemployment, infrastructure, poverty

prob sector, economy, policy, poverty, land, rural, infrastructure

26 FREX: wicket, bowler, batsman, t20, sri lanka, inning, twenty20

Lift: rafe, hore, duminy, wisden, lasith malinga, highveld lions, cricketing

Score: cricket, wicket, bowler, proteas, england, india, batsman

prob cricket, india, over, australia, england, wicket, ball

27 FREX: voting, eff, ballot, malema, ifp, nfp, bekkersdal

Lift: iec., peter mutharika, patricia kopane, kamanya, www.elections.org.za, mamoepa, sy mamabolo

Score: da, voter, iec, eff, voting, ballot, malema

prob da, voter, voting, eff, iec, ballot, candidate

28 FREX: ml, cheese, flour, tomato, salad, lemon, compost

Lift: storytime, ratatouille, gerasperde, versoeter, tues, eggplant, pinot

Score: ml, wine, garden, egg, cheese, sugar, bake

prob garden, wine, eat, egg, plant, restaurant, ml

29 FREX: birdie, bogey, nadal, djokovic, rodger, pga, federer

Lift: birdie, flores, cro, pga, roger federer, djokovic, federer

Score: birdie, hair, hole, golf, tournament, bogey, round

prob round, hole, hair, golf, shot, champion, tournament

30 FREX: pistorius, steenkamp, reeva, oscar pistorius, gerrie nel, dixon, barry roux

Lift: reeva, gudmundstottir, oldwage, henke, masipa, michelle burger, silverwoods

Score: pistorius, nel, reeva, steenkamp, oscar, roux, reeva steenkamp

prob pistorius, nel, witness, steenkamp, oscar, reeva, trial

31 FREX: liverpool, chelsea, sundowns, midfielder, goalkeeper, psl, bafana

Lift: rodgers, bartlett, valcke, sevilla, forresters, repucom, novians

Score: league, chiefs, liverpool, coach, football, goal, chelsea

prob league, goal, coach, title, football, score, united

32 FREX: amcu, miner, lonmin, amplats, association of mineworkers, impala platinum, r12,500

Lift: r12 500, mark cutifani, goldrich holdings, william mpembe, anglo american platinum and, anglo american platinum (, csr zhuzhou electric locomotive

Score: amcu, lonmin, strike, platinum, miner, rhino, mine

prob	strike, mine, union, employee, amcu, platinum, protest

Table 6: Topics and their labels.