

Measuring Populism in Context: A Supervised Approach with Word Embedding Models^{*}

YAOYAO DAI[†]

University of North Carolina at Charlotte

ABSTRACT

Populism is a worldwide phenomenon. However, much of the academic literature has focused on just a few regions, notably Western Europe and Latin America. One obstacle to the study of populism from a global perspective is coming up with a measure of populism that is comparable across both space and time. In this paper, I clarify the theoretical concept of populism and introduce a novel computer-assisted supervised method combined with word embedding models for identifying populism in political texts. My method is highly generalizable and is not constrained by language. I demonstrate the usefulness of my method by measuring the degree of populism in English-language party manifestos and Chinese news articles. My method is able to identify populist texts with a high degree of accuracy in out-of-sample predictions, and therefore opens up new research opportunities to examine populism from a truly comparative perspective.

^{*}Note: I thank Jonathan Abel, Matt Golder, Sona Golder, Kirk Hawkins, Boliangzhu for their helpful comments on this paper. The data, codebook, and computer code necessary to replicate the results in this analysis will be made publicly available on my dataverse site and on my personal webpage at <https://yaoyaodai.github.io> on publication.

[†]Assistant Professor, Department of Political Science and Public Administration, University of North Carolina at Charlotte. (yaoyao.dai@uncc.edu).

1 Introduction

Populism is growing across the globe. One need only look at the 2016 Brexit referendum campaign in the UK, the rapid growth of the radical left Podemos and Syriza in Spain and Greece, the emergence of radical right governments in East European countries such as Poland, Hungary, Bulgaria, and Slovakia, the presidential election of Donald Trump in the United States, the leftist charismatic presidencies of Hugo Chavez, Evo Morales, and Rafael Correa in Latin America, the successful anti-establishment campaign of Muhammadu Buhari in Nigeria, and the persistent emphasis on ‘Mass Line’ leadership in China, among many other examples, to see that populism is a truly global political phenomenon (Golder and Golder, 2016). Populism and its adherents pose a direct challenge to pluralism and liberal democracy (Mudde and Rovira Kaltwasser, 2013). When politics is portrayed as a moral conflict between good and evil, there is no space for opposition or diverse points of view. As such, it is important to understand what drives the recent rise in populism.

Although populism is a global phenomenon, much of the existing literature focuses on just a few regions, notably Europe and Latin America. Significantly, scholars of populism who work on the different regions of the world rarely speak or interact with each other. To a large extent, this geographic segregation of research on populism is not because regional scholars differ in fundamental ways over how to conceptualize and think about populism. Most scholars, for example, agree that populism can usefully be understood as a thin ideology or discursive style that portrays the world as being divided into two homogenous and antagonistic groups, the ‘good people’ and the corrupt few (Mudde, 2004, 2007). Instead, the problem comes when scholars attempt to measure the concept of populism. The actors who comprise the ‘corrupt few’ and the words used to describe the conflict between the two antagonistic groups are often context dependent, both in terms of space and time. This context dependency has made it difficult to conduct empirical research to test theories about the causes and consequences of populism from a truly cross-regional or global perspective.

Historically, scholars have adopted one of two approaches for measuring populism. The first, and most common, approach involves using humans to code populist discourse. This is problematic because it is costly and time-consuming. This helps to explain why the human coding of populist discourse has been limited to a handful of case studies of individual countries (Jagers and Walgrave, 2007; Hawkins, 2009). The second, and more recent, approach involves using dictionary-based

methods of automated content analysis to code populist discourse (Rooduijn and Pauwels, 2011; Bonikowski and Gidron, 2016b). These dictionary-based methods, which count the proportion or the presence of words associated with populism, are more efficient. However, they are less generalizable and have less content validity as the relevant words in the dictionary, often related to the identity of the corrupt few, are specific to particular national and temporal contexts. Not only are dictionaries hard to generalize to contexts outside of the domain in which they were constructed (Grimmer and Stewart, 2013), but building a valid dictionary also requires a deep knowledge of the populist politics in a given country, something that hinders the application of dictionary-based methods to new and less familiar cases. Using different dictionaries for different cases also makes the comparability of any populism measure questionable, because we are using different coding rules in different cases.

In this paper, I introduce a computer-assisted supervised method for measuring populism that combines the high content validity and generalizability associated with human coding with the efficiency associated with automated text analysis methods. Intuitively, we can consider the supervised approach as training the computer to learn from human coded examples and finally to code like a human coder itself. To address the fact that populist rhetoric is not embedded in specific words but rather in a particular context, and that words, such as ‘people’, that appear often in populist texts are also common in other forms of text, I incorporate a word-embedding (doc2vec) model from machine learning to process political texts. The doc2vec model, which is new to political science, loosens the ‘bag of words’ assumption underlying the typical automated text analysis methods used in political science (Le and Mikolov, 2014).¹ Rather than representing political texts based on the counts of each unique word, the doc2vec model learns the meaning or similarities between words and documents by taking account of their surrounding words and context. An appealing feature of my supervised method is that it is unconstrained by language.

To evaluate the usefulness and accuracy of my method, I examine the level of populist discourse in English-language party manifestos from seven different countries since 1945 and in reports from China Central Television’s (CCTV) Xinwen Lianbo (Daily News Program) from 2003 to 2015. As I demonstrate, my method is able to classify populist documents with 95% accuracy

¹The doc2vec model is an extension of the more foundational word2vec model (Mikolov et al., 2013). To my knowledge, Radford (2016) provides the only published application of a word2vec model in a political science setting; I am unaware of any political science applications using a doc2vec model.

and 89% AUC (area under the curve) in out-of-sample predictions in the English corpus and 88% AUC in the Chinese corpus. Given that there is considerable variation in how even humans code political texts for populist discourse (Hawkins, 2009; Hawkins and Silva, 2018), this performance is quite remarkable. By measuring populist discourse in context, my supervised method opens up new avenues for a truly cross-regional and global approach to the study of populism.

In the next section, I discuss and clarify the concept of populism. While there is broad agreement that populism portrays the world as being divided into two homogenous and antagonistic groups, the ‘good people’ and the corrupt few, scholars have yet to agree on the precise conceptual components of populist discourse that distinguish it from other forms of discourse. I argue that the necessary and sufficient components for identifying populist discourse are (i) people-centrism, (ii) anti-pluralism, and (iii) moralized politics. With the concept of populism in hand, I then present the different approaches that scholars have historically employed to measure populism and discuss their strengths and weaknesses. As I indicate, a key problem in the literature is that we do not currently have a method for measuring populism that is both generalizable and efficient. I address this problem in section four where I outline my new computer-assisted supervised learning method for measuring populism in political texts. Among other things, I justify adopting a supervised approach to measuring populism and explain how one can measure populism in context by incorporating word-embedding models. In the fifth section, I evaluate my supervised method by using it to measure populism in both English- and Chinese-language texts. In addition to having strong predictive accuracy, I demonstrate that my method is easy to validate and apply to a broad set of cases.

2 Populism as a Theoretical Concept

Scholars have conceptualized populism in at least four different ways. According to the *political-strategic conceptualization*, populism is a mobilization strategy used by personalistic leaders who seek to govern based on a direct, unmediated, and uninstitutionalized relationship with their supporters (Weyland, 2001, 2017; Roberts, 2006; Levitsky and Roberts, 2011). According to the *sociocultural conceptualization*, populism is performative and involves political actors flaunting an improper or ‘low’ style of politics that is intended to shock and build a close rapport with their

supporters (Moffitt, 2016; Ostiguy, 2009, 2017). According to the *economic conceptualization*, populism involves adopting unsustainable economic policies that are popular but ultimately hurt the interests of the majority (Dornbusch and Edwards, 1991; Acemoglu, Egorov and Sonin, 2013). While these different approaches to conceptualizing populism offer valuable insights into different aspects of populism, they have been criticized for their lack of clarity and for the fact that they tend to emphasize features that are not shared by all forms of populism around the globe (Gidron and Bonikowski, 2013; Bonikowski and Gidron, 2016a; Mudde and Rovira Kaltwasser, 2017, Forthcoming; Hawkins and Kaltwasser, 2018). Partly in response to these critiques, there has been a growing consensus in the literature around adopting a more minimalist ideational conceptualization of populism that focuses on populism’s core features (Mudde, 2004, 2007; Mudde and Rovira Kaltwasser, 2017; Hawkins and Rovira Kaltwasser, 2017; Hawkins et al., 2018). This is the concept of populism employed in this paper.

According to the *ideational conceptualization*, populism is a thin-centered ideology or discursive style that portrays society as being divided into two homogenous and antagonistic groups, the good people and the corrupt few, and that emphasizes that politics should reflect the general will of the people (Mudde, 2004; Hawkins et al., 2018).² Populists divide society into two groups based purely on morality — good versus evil — rather than in terms of socioeconomic status, class, ethnicity, or some other form of political identity. This means that members of a country’s socioeconomic elite, such as Silvio Berlusconi in Italy and Donald Trump in the United States, can act as authentic representatives of the ‘people’. While these political actors do not necessarily provide descriptive representation, they provide symbolic representation by ‘standing for’ the people (Pitkin, 1967). As emphasized by constructivist scholars of political representation (Saward, 2006, 2014), the key to populist representation rests simply on being accepted by those one claims to represent. As populist politicians are often part of the conventional elite in many countries, the group opposed to the people is best thought of as the ‘enemy’ rather than the ‘elite’. Although the specific identity of the enemy varies across countries and over time, the enemy is always portrayed as corrupt and engaged in subverting the interests and general will of the people. As a thin-centered ideology or discursive style, populism has little specific ideological content. This is why populism tends to

²While some scholars differentiate between populism as a thin-centered ideology and a discursive style (Gidron and Bonikowski, 2013; Bonikowski and Gidron, 2016b), the differences are, in practice, minor and often irrelevant for most research questions (Mudde and Rovira Kaltwasser, Forthcoming; Hawkins and Kaltwasser, 2018).

be combined with “other ideologies such as communism, ecologism, nationalism, and socialism” (Mudde, 2004, 544). It is these other ideologies that provide the bulk of the policy content of populist parties. This helps to explain why populism is associated with parties across the traditional left-right spectrum, and why it often takes such varied political forms across countries and regions (Mudde and Rovira Kaltwasser, 2013).³

Populism is directly opposed to elitism and pluralism (Mudde and Rovira Kaltwasser, 2017). Like populism, elitism separates society into two groups on moral terms. The difference, though, is that while populism views the people as morally superior, elitism views them as corrupt and dangerous. Pluralism acknowledges that society is composed of multiple groups with differing interests, and appreciates that compromise and bargaining are necessary to reconcile these differences. In contrast, populism considers the people to be a homogenous group that shares a general will. Since there are no meaningful divisions among the people, populism does not recognize the need for political compromise or constitutional checks and balances to protect minority and individual rights. Populism calls for more direct forms of government such as referenda and popular initiatives that provide unmediated access to the people, and criticizes group-based politics as ‘special interest’ politics. It is populism’s distinctly anti-pluralist stance that puts it at odds with liberal democracy, despite the fact that both populism and liberal democracy claim to view the people as the legitimate source of all political power.

While there is a growing consensus in the literature around this minimalist ideational conceptualization of populism, there remains considerable disagreement as to exactly what the necessary components are when it comes to measuring populism. For example, Rooduijn and Pauwels (2011) focus on people-centrism and anti-elitism as the two core theoretical components of populism, whereas Jagers and Walgrave (2007) focus on people-centrism and anti-establishmentarianism. In their study, Bonikowski and Gidron (2016b) emphasize references to a morally corrupt elite and the defense of ordinary people. Scholars also disagree on whether discourse can be considered populist if it has only some, but not all, of the elements associated with populism. Jagers and Walgrave (2007) refer to discourse that frequently mentions the people as ‘thin’ populism and discourse that frequently criticizes the establishment as ‘thick’ populism. Similarly, Hawkins (2009)

³The lack of specific ideological content associated with populism means that it is inappropriate to use topic models (Blei, Ng and Jordan, 2003; Grimmer and Stewart, 2013; Roberts et al., 2013) as a form of automated content analysis to identify populist parties or topics — populist discourse can be associated with any topic.

considers discourse with none of the components of populism as non-populist, discourse with some of the components as somewhat populist, and discourse with all of the components as extremely populist. Although [Rooduijn and Pauwels \(2011\)](#) argue that people-centrism and anti-elitism are both components of populism as a theoretical concept, they only use anti-elitism as an indicator when they measure populism. Many scholars have argued, though, that anti-elitism or anti-establishmentarianism is not sufficient for identifying populist discourse. Indeed, [Müller \(2016\)](#) argues that anti-pluralism is much more important for populism than anti-elitism.

In what follows, I argue that populism has three theoretical components — *people-centrism*, *anti-pluralism*, and *moralized politics*. These components are jointly necessary and sufficient to identify populism. In other words, discourse is populist if and only if it is people-centric, anti-pluralist, *and* it uses morality to distinguish between societal groups. Since it recognizes the people as the only legitimate source of power, populist discourse must be people-centric. People-centrism, though, is not sufficient to distinguish populism from other forms of political ideology. Liberal democracy, for example, also believes that political power should be derived from the people and it is not unusual for democratic leaders to speak positively about the people.

What separates populism from liberal democracy is its emphasis on anti-pluralism. While there may be differences among the people, such as gender, occupation, and ethnicity, populists believe that the people have a general will and that politics is about finding and representing this ([Müller, 2016](#); [Mudde and Rovira Kaltwasser, Forthcoming](#)). Ultimately, populists reject the idea that there are any politically meaningful divisions among the people. Although necessary, anti-pluralism, like people-centrism, is insufficient on its own to distinguish populism from other political ideologies. Elitism, for example, is anti-pluralist in viewing the elite and the people as two homogenous groups. Nationalism is also anti-pluralist in its demand for cultural or ethnic congruence between the state and the nation ([Mudde, 2007](#)).

Moralized politics is what separates populism from other anti-pluralist and people-centric ideologies. Populism structures politics primarily along moral lines. Although ideologies such as nationalism and socialism also emphasize the importance of ordinary people and create a separation between us and them, the divide in nationalism is primarily cultural or ethnic and the divide in socialism is primarily class. Like people-centrism and anti-pluralism, moralized politics is insufficient on its own to identify populism. Elitism, for example, is also a form of moralized politics in

which the elite are depicted as morally superior to the ordinary people.

It is common for scholars to associate two additional, but unnecessary, features — anti-elitism and anti-establishmentarianism — with populism. Although these two features are often used interchangeably, they are, in fact, distinct. Anti-elitism is an antipathy against all elites and rejects the claim that the elite is morally superior to the people. In some cases, the constructed elites are the establishment. In these cases, anti-elite rhetoric is also anti-establishment rhetoric. However, it is possible for the constructed elites to not be part of the establishment. Anti-establishmentarianism is, thus, neither necessary nor sufficient for identifying populism. More importantly, treating anti-establishmentarianism as a necessary component of populism rules out the possibility that the establishment could be populist as well.

Anti-elitism is very similar to people-centrism in the way that it considers the people, rather than the elite, to be the legitimate source of political power. A focus on people-centrism is preferable, though, as anti-elitism suggests that populism cannot be used by elites against their opponents. This would seem to rule out the possibility that authoritarian leaders, for example, would use populism against their enemies (Dai and Shao, 2016). Moreover, an emphasis on anti-elitism often confuses the reader into thinking that populism is against some conventional elite based on its socioeconomic or ethnic status rather than its moral status. Note that much of what is thought important about anti-elitism by populism scholars is captured, at least implicitly, by people-centrism. Discourse can be populist without explicitly criticizing the corrupt few. By depicting a virtuous and homogenous people with a general will (people-centrism), any voice that disagrees with the ‘representative’ of the general will will automatically be considered the corrupt few who are subverting the people’s interests. Consider Hugo Chávez’s campaign speech in 2006 when he said,

“You the people are the giant that awoke, I your humble soldier will only do what you say. I am at your orders to continue clearing the way to the greater Fatherland ... Because you are not going to reelect Chávez really, you are going to reelect yourselves, the people will reelect the people. Chávez is nothing but an instrument of the people.”

Although there is no mention of a corrupt enemy in this quote, it is clear that this discourse is populist. By depicting the people as one giant and his actions as following the people’s will, Chávez is clearly indicating that anyone who stands against him should be considered the people’s enemy.

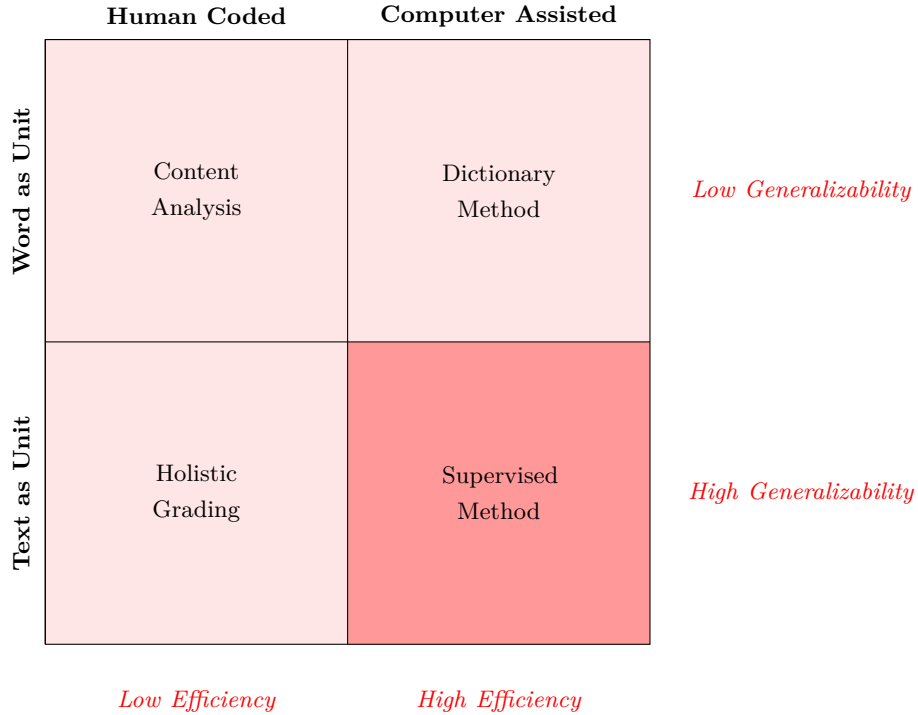
Conceptualizing populism in terms of three necessary and sufficient components — *people-centrism*, *anti-pluralism*, and *moralized politics* — has several advantages for comparative and empirical research on the causes and consequences of populism. First, this conceptualization is theoretically clear and empirically sufficient to distinguish populism from other political ideologies. Using only one or some subset of these three components makes it impossible to distinguish populism from other related ideas such as liberal democracy or nationalism. Second, the three components I identify contribute to a minimalist definition of populism, thereby facilitating a truly comparative approach to the study of populism. All definitions of populism must include at least the three components identified in my conceptualization. Third, my conceptualization focuses on the procedural aspects of populism as opposed to the consequences of populism, and therefore allows us to avoid a tautological study of the effects of populism. Although some studies argue that populist politicians tend to simplify politics, that they are charismatic, and that they introduce unsustainable redistributive policies, these are only possible outcomes of populism and may not be observed in all contexts (Acemoglu, Egorov and Sonin, 2013; Mudde and Rovira Kaltwasser, Forthcoming; Hawkins et al., 2018).

3 Different Approaches to Measuring Populism

With a clear theoretical concept of populism in hand, we can now turn to measuring populism. Many scholars have attempted to measure populist discourse among political actors. Approaches to measuring populism can be categorized along two different dimensions. The first dimension distinguishes between whether scholars conduct human-coded content analysis or whether they conduct computer-assisted content analysis. Computer-assisted methods are more efficient than human-coded methods. The second dimension focuses on whether the word or the text as a whole is treated as the unit of analysis. Approaches that use the text as the unit of analysis are more generalizable than those that use words as the unit of analysis. In Figure 1, I use these two dimensions to identify four distinct approaches to measuring populism.

Human-coded methods comprise either standard *content analysis* or *holistic grading*. Standard content analysis involves breaking up texts into units of a certain length, such as sentences or paragraphs, and then using humans to code whether each unit should be considered populist based

Figure 1: Different Approaches to Measuring Populist Discourse in Political Texts



on the presence of certain words associated with things like people-centrism and anti-pluralism. Different aggregation rules are then used to measure the level of populism in the text as a whole. [Jagers and Walgrave \(2007\)](#) use this type of content analysis to examine the level of populism in the communication styles of political parties in Belgium, while [Rooduijn and Pauwels \(2011\)](#) deploy it to evaluate populist discourse in four European countries. With holistic grading, human coders use whole texts as the unit of analysis. Instead of being given lists of words associated with things like people-centrism and anti-pluralism, holistic graders are given coding rubrics for identifying populist, non-populist, and possibly mixed texts. These human coders are also provided with ‘anchor texts’ that exemplify each of the possible categories. [Hawkins \(2009\)](#) uses holistic grading to examine the level of populism in the speeches of political chief executives around the world.

Holistic grading is more generalizable than standard content analysis. Standard content analysis relies on identifying specific words or phrases that are associated with things like people-centrism and anti-pluralism. A single list of words is unlikely to travel well across different countries and time periods. One could use a different list of words for each new context, but this raises

concerns about the extent to which the resulting populism scores are comparable across different cases. Since holistic grading uses a common coding rubric and does not rely on identifying the presence of specific words and phrases, it is more generalizable.

All human-coded methods are time consuming and inefficient. With standard content analysis, human coders must break documents up at the sentence or quasi-sentence level and then read and label every sentence. This is time and labor intensive. While holistic grading is faster, because it does not require coders to read every word and label every sentence, it still requires the human coder to read the full document, which could be several pages long in the case of a speech or dozens (if not more) pages long for other political texts such as party manifestos. This helps to explain why scholars using human-coded methods often only sample a handful of texts out of a much larger collection of political texts in their empirical analyses (Jagers and Walgrave, 2007; Hawkins, 2009). Political actors can vary their level of populism based on their audience, the stage of an election campaign, or other circumstances. As a result, it can be very misleading to infer a political actor’s overall level of populism based on just a small number of texts.

Computer-assisted methods are much more efficient as they enable scholars to quickly measure populism in a large numbers of political texts. *Dictionary methods* are the only computer-assisted method that have currently been used to measure populism (Rooduijn and Pauwels, 2011; Bonikowski and Gidron, 2016b). To some extent, the dictionary method mirrors the standard content analysis conducted by human coders. Researchers first construct a list of words — a dictionary — associated with populism. They then measure populism by counting the proportion of ‘populist’ words in a given text. However, unlike standard content analysis, which asks the human coder to take account of the context in which words appear, the dictionary method assumes that the individual words are all that matter. By treating a text as a ‘bag of words’, dictionary methods ignore both the order and context of individual words.

The ‘bag of words’ assumption underpinning dictionary methods is especially problematic when it comes to measuring populism because many of the words that reflect people-centrism, such as ‘people’, ‘majority’, ‘we’, or ‘us’, commonly appear in both populist and non-populist texts. Without taking account of the *context* in which these words appear, it is all but impossible to determine whether they signal populism or not. Indeed, this is precisely why Jagers and Walgrave (2007) trained their human coders to interpret the context when coding positive references to the

people. To avoid false positives (coding non-populist texts as populist), scholars utilizing the dictionary method to measure populism have tended to ignore words associated with people-centrism and focus instead on words associated with anti-elitism as these words are less commonly observed in non-populist texts. Ignoring indicators of people-centrism, though, gives these measures less content validity than those produced by standard content analysis where people-centrism and context are taken into account (Rooduijn and Pauwels, 2011). Even if one were to ignore these issues, dictionary methods, like standard content analysis, suffer from low generalizability as dictionaries built for one setting do not necessarily travel to other settings. As an example, the dictionary used by Bonikowski and Gidron (2016b) includes American-specific words such as ‘Wall Street’, ‘Main Street’, and ‘Washington elite’. Building good dictionaries for different settings is difficult as it requires significant language skills and a deep understanding of a country’s politics and history. Even if one had the necessary contextual knowledge to build multiple dictionaries, it is unclear whether the resulting measures of populism would be comparable due to the use of different dictionaries.

As Figure 1 indicates, holistic grading scores high on generalizability but low on efficiency, while dictionary methods score high on efficiency but low on generalizability. What would be ideal is if we were able to combine holistic grading with a computer-assisted method of text analysis. I introduce such a method for measuring populism in political discourse in the next section.

4 A Supervised Method with Word Embedding

In this section, I propose a computer-assisted supervised method that learns to classify populist and non-populist texts from human-coded examples based on holistic grading. To address the challenge that populist discourse is not situated in individual words but in the broader context, I relax the common ‘bag of words’ assumption underlying the typical automated text analysis methods used in political science.⁴ I do this by introducing a word-embedding model from machine learning that takes account of both the order and context in which words appear when processing political texts.

There are three steps to my supervised method. Recall that the goal is to identify populist

⁴Hawkins and Silva (2018) have also recently employed a supervised method for measuring populism. Unlike my method, though, their method does not incorporate a word-embedding model and instead retains the traditional ‘bag of words’ representation of text. Possibly due to the complexity of populist discourse and the difficulty of capturing the various components of populism in individual words, the performance of their supervised method suffers from, among other things, a high false-negative rate; that is, it frequently fails to identify populist texts.

and non-populist rhetoric in political texts. After collecting the corpus of relevant political texts, the first step is to hand-code a small sample of these texts into populist and non-populist categories using holistic grading. Following my earlier conceptual discussion of populism, these texts are coded according to a rubric in which texts are identified as populist if and only if they display people-centrism, anti-pluralism, and a moralized politics, and non-populist otherwise. This small sample of texts now becomes the ‘training set’. The second step is to ‘train’ a machine (a computer) to ‘learn’ the ‘rules’ (textual features) that allow it to classify the texts in the training set as closely as possible into the same populist and non-populist categories produced by the human coder. Intuitively, the training set texts perform a similar role to the anchor texts that human coders use in holistic grading. Once the machine has ‘learned’ to code populism ‘like a human’, the third step involves using it to identify populist discourse in the remaining texts in the full corpus.⁵ An appealing feature of my supervised method is that it is unconstrained by language. The only requirement is to have a small amount of hand-coded training documents in each relevant language. Since the hand-coding in each language follows the same coding rubric, the measurement of populism across languages should be comparable.

The second step in my supervised method has two distinct parts: representation and classification. While human coders can directly interpret words by looking at their meaning and context, words first need to be transformed into numbers for a machine to be able to ‘read’ them. Once the training texts have been represented numerically, a classification algorithm can then be used to separate populist and non-populist texts along similar lines to human coders. It is clear from this that the machine does not code political texts in the exact same way as a human coder does. Instead, the machine attempts to mimic a human coder who is using holistic grading by producing the same, or similar, classification of populist and non-populist texts in the training set. In what

⁵In general, *supervised* methods involve giving a computer a training set with some input features, such as the frequencies of certain words in a document, and labeled output classes for each document, such as populist and non-populist. Once the computer learns how the input features are related to the labeled output classes, it can be used to classify unseen texts into the target output classes based on their input features. An alternative approach to classification is to use an *unsupervised* method. With *unsupervised* methods, there is no training set and the computer attempts to learn the underlying structure of the data without using any explicit output labels. A common unsupervised method used in political science is the topic model (Blei, Ng and Jordan, 2003; Grimmer and Stewart, 2013; Roberts et al., 2013). Supervised methods are preferable whenever we want to retrieve something specific, such as when we want to measure a clearly-defined theoretical concept such as populism. Put differently, supervised methods are better when we know what we want, or at least we know it when we see it. In a measurement context similar to the one here, supervised methods should have better face and content validity than unsupervised methods because they train the computer to sort similarly to a human. Unsupervised methods can only produce machine-sorted categories that may or may not accurately reflect specific human-defined concepts such as populism.

follows, I discuss the benefits of using word-embedding models to represent text as numbers and briefly describe the classification algorithm used to separate texts into different categories.

4.1 Word Embedding Models: A Better Way to Represent Text

There are many ways to represent words and texts numerically. In political science, scholars typically use a ‘bag-of-words’ approach, which assumes that a text is composed of a ‘bag’ of independent words and can be represented by the counts or frequencies of each unique word (Grimmer and Stewart, 2013). The order of, and relationships between, words, such as their semantic and syntactic similarities, are lost when text is represented in this way.

Consider a simplified example in which we have three documents that each contain two words. The first document contains the text “Citizen Party”. The second document contains the text “People Party”. The third document contains the text “Elite Party”. There are four unique words in this corpus: “Citizen”, “People”, “Elite”, and “Party”. A document-feature matrix that numerically represents the text as counts of words in our three documents is shown in Table 1. Document 1 mentions “Citizen” one time, “People” zero times, “Elite” zero times, and “Party” one time. Document 2 mentions “Citizen” zero times, “People” one time, “Elite” zero times, and “Party” one time. Document 3 mentions “Citizen” zero times, “People” zero time, “Elite” one times, and “Party” one time. Each textual feature, or word in this case, can be treated as a dimension along which to map our documents. For example, if we were to treat “Citizen” as our x -axis, “People” as our y -axis, “Elite” as our z -axis, and “Party” as our w -axis, then Document 1 can be represented as vector $(1, 0, 0, 1)$, Document 2 can be represented as vector $(0, 1, 0, 1)$, Document 3 can be represented as vector $(0, 0, 1, 1)$. The vectorization of our three documents is graphically illustrated in Figure 2a. Because all three documents have the same value in the “Party” dimension, the “Party” dimension is omitted in Figure 2a. In a real corpus, we can have many thousands of unique words, and a vector space with many thousands of dimensions is needed to represent all of the documents.

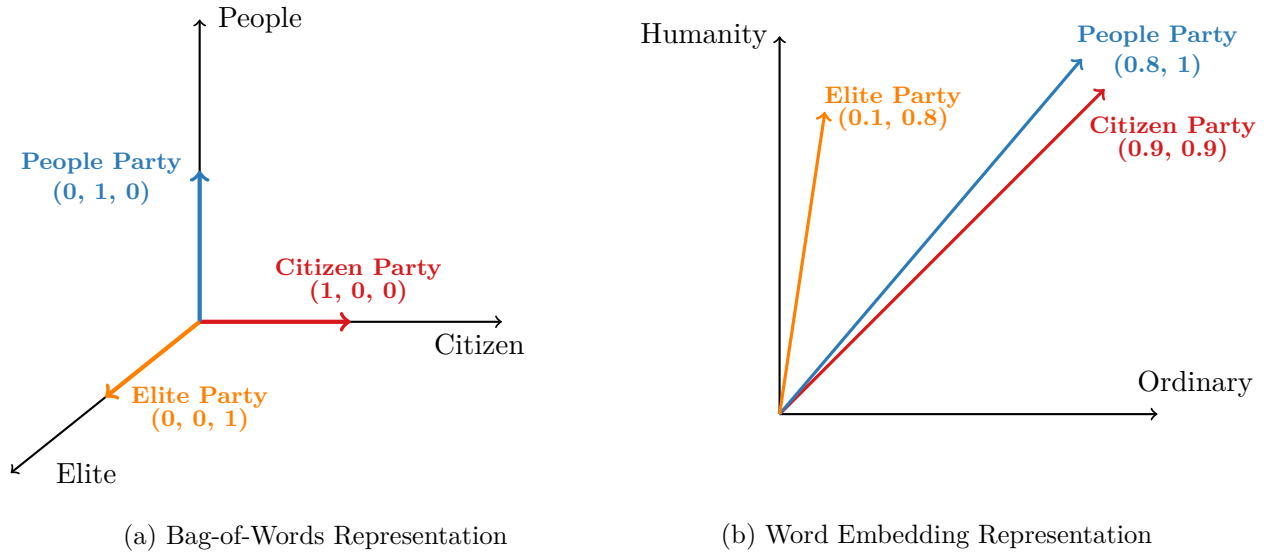
Because a vector is just the numerical representation of a distribution of words in a document, it is possible to compare how similar documents are by calculating the distance between their corresponding vectors. The vectors representing similar documents will be close to each other in

Table 1: A Simplified Example of a Document-feature Matrix

	Citizen	People	Elite	Party
Document 1	1	0	0	1
Document 2	0	1	0	1
Document 3	0	0	1	1

the multidimensional space, while the vectors representing two different documents will be far apart. Based on this “bag-of-words” vectorization of the three documents, the three documents have the same distance from each other. In other words, the three documents are considered equally similar to each other. However, in reality, some of these documents are more similar than others. In populist discourse, the words “People” and “Citizen” are often used interchangeably. In contrast, the word “Elite” signifies something quite different from “People” or “Citizen”. A human coder of our three documents would recognize this and conclude that Document 1 (“Citizen Party”) and Document 2 (“People Party”) are pretty much identical, but that Document 3 (“Elite Party”) is quite different and may even be the opposite of the other two documents.

Figure 2: Simplified Examples of Document Vectorization



Note: Figure 2 shows two different ways of vectorizing our three documents, “Citizen Party”, “People Party”, and “Elite Party”.

The problems with this “bag of words” representation here is that each word is treated as an arbitrary dimension that is independent from all of the other dimensions. This ignores the

semantic and syntactic relationship between words, such as the similarity between people and citizen in populist discourse. To incorporate the semantic and syntactic relationship between words, I use the word embedding method from computational linguistics. Intuitively, word embedding is a way to represent words or texts based on their meaning or the information they convey. Continuing with our previous example, the three words “Citizen” “People” and “Elite” can be captured by just two dimensions: ordinary and humanity. The three words all convey similar information on the humanity dimension. However, the words “People” and “Citizen” are on the higher end of the ordinary dimension, while the word “Elite” is on the lower end of the ordinary dimension. The words “People” and “Citizen” signify something more ordinary than “Elite”. A hypothetical embedding matrix is shown in Table 2. As before, we can visualize these dimensions and our documents in a vector space. Again, party dimension is omitted to make things clearer, because all three documents have the same value on that dimension. As illustrated in Figure 2b, word embedding recognizes that the documents “People Party” and “Citizen Party” are similar to each other (close in the vector space), while the document “Elite Party” is dissimilar (far away in the vector space).

Table 2: A Simplified Example of a Embedding Matrix

	Ordinary	Humanity
Citizen	0.9	0.9
People	0.8	1
Elite	0.1	0.8

There are many different algorithms to train a word embedding representation. Instead of treating each word as an unique and independent dimensions (bag-of-words), those methods assume that a word’s meaning is given by the context in which it is used, that is, from its surrounding words (distributional hypothesis). If two words tend to be surrounded by similar words, then the model learns to position the two words close to each other in the vector space. In this paper, I use the doc2vec model which learns the vector representation of words and documents the same time so that semantically and syntactically similar words and documents (those that share common contexts) are represented by similar numerical vectors (Le and Mikolov, 2014). Although this method has been applied frequently in machine learning applications, it has yet to be employed in a political

science application. A detailed and more accurate explanation of doc2vec model is included in [Online Appendix A](#).

Once we have vectorized the documents in the training set using the doc2vec model, we can use a classification algorithm to learn how to best separate them into populist and non-populist categories to match the classification choices made by the human coders. There are many classification algorithms one could use to perform this supervised learning, including support vector machines (Boser, Guyon and Vapnik, 1992; D’Orazio et al., 2014), neural networks (Peterson and Spirling, 2018), random forests (Breiman, 2001; Hill and Jones, 2014), and boosting methods (Mason et al., 2000; Freund and Schapire, 1997). Each algorithm has its comparative advantage, and none outperforms all of the others across all possible evaluative criteria. I use random forests to classify texts into populist and non-populist categories for three reasons (Breiman, 2001).⁶ First, random forests allow for complex non-linear relationships between textual features during the classification process. Second, random forests suffer less from overfitting problems than other classification algorithms. Third, random forests not only classify texts as populist or non-populist, they also provide a measure of uncertainty around their classification predictions. Such measures of uncertainty are not produced by human-coded methods or computer-assisted dictionary methods.

5 Applications

To demonstrate how my supervised method can be employed to measure populist discourse, I use it in two applications: (i) English-language party manifestos from 1945-2017 from the Manifesto Project (Lehmann et al., 2017) and (ii) daily news reports from China Central Television’s (CCTV) Xinwen Lianbo (Daily News Program) from 2003 to 2015.

5.1 Measuring Populist Discourse in English-language Party Manifestos

Party manifestos are arguably the most important form of campaign message used by political parties. This is because they contain a party’s official platform. Previous research has shown that parties spend considerable time deciding which issues to include in their manifestos and how much space to give them (Janda et al., 1995; Green and Hobolt, 2008; Adams, Ezrow and Somer-

⁶For a detailed explanation of how random forests learn to classify documents that have been labeled populist and non-populist, see [Online Appendix B](#).

Topcu, 2011). There is also evidence that parties think strategically about how they talk about the issues that appear in their manifestos (Crabtree et al., 2018). Scholars often determine a party's membership in a particular party family based on the content of their manifesto (Mair and Mudde, 1998). Party manifestos are clearly an important source of political discourse.

An appealing feature of manifestos is that they are highly comparable across countries because they tend to be produced in a similar way and serve a similar purpose. Party manifestos are also a very accessible form of political discourse as the Manifesto Project (Lehmann et al., 2017) provides original manifestos and coded policy positions for over 1,000 parties from 1945 until today in over 50 countries. Significantly, manifestos have already been used in a wide range of research projects looking at political discourse, including populist discourse (Rooduijn and Pauwels, 2011; Rooduijn, de Lange and van der Brug, 2014; Akkerman and Rooduijn, 2017). This means that we could easily use existing measures of populism based on manifestos to validate my new measure of populist discourse. The Manifesto Project also provides additional variables that are important to studies of populism. For example, it provides hand-coded issue areas and policy positions for parties at the level of the quasi-sentence. This opens up the possibility of looking at whether populist discourse is attached to some issue areas more than others. It also provides election results that could be used to examine the relationship between populist rhetoric and election success.

In this paper, I utilize all of the party manifestos written in English from the Manifesto Project. These manifestos are written by 54 parties in seven countries: the United States, Canada, the United Kingdom, Ireland, South Africa, Australia, and New Zealand. In total, there are 119 party manifestos and 103,800 quasi-sentences in the corpus. A typical party manifesto is somewhere between 20 and 50 pages in length and covers many issue areas. Although Party Manifesto project codes issue areas at the quasi-sentence level so that each quasi-sentence addresses only one issue area, quasi-sentences are too short to convey all three of the necessary components of populism. As a result, I treat each *section* of a manifesto as a document. By section, I mean the content between two subheadings. Based on my reading of the manifestos, a section usually expresses a cohesive and complete message, which is a better unit of analysis for measuring populist rhetoric than a single sentence or a full manifesto. In my set of manifestos, there are 11,801 sections, which means I have 11,801 'documents' in my corpus.

5.1.1 Hand-Coding the Training Set: Holistic Grading

The first step of my supervised method involves hand-coding a random sample of manifestos as a training set using holistic grading. To build my training set, I randomly sampled one manifesto from each of the United Kingdom, United States, South Africa, Australia, and New Zealand. There are 439 sections and 4,700 quasi-sentences in the five manifestos in the training set.⁷ Unlike standard content analysis techniques, which asks the human coder to code the counts of keywords within a window of adjacent words or sentences, holistic grading asks the coder to code the text into predefined categories based on her interpretation of the full text. The holistic grading method requires researchers to develop a clear rubric and to provide several anchor texts that exemplify each category in the rubric. Based on my earlier conceptualization of populism, my coding rubric classifies a document as populist (1) if it refers to people-centrism, anti-pluralism, *and* a moralized politics, and non-populist (0) otherwise. Sixteen (4%) of the 439 documents in the training set contained all three of the necessary components of populism and were coded as populist. Populist discourse appears to be quite rare in party manifestos.

To illustrate how the holistic grading works, I show four documents in Figure 3, two of which are coded as populist and two of which are coded as non-populist. The two documents on the left bring up people-centrism and anti-pluralism, referring to “our people”, “all of us”, “the majority”, the “whole society”, “special interests”, and “a select few.” They also refer to a moralized politics, speaking, among other things, of betrayal, fairness, illegality, and suffering. The two documents on the right, while they mention the people, do not talk in moralized terms, nor do they exhibit anti-pluralism. One thing to note from these four documents is how difficult it would be to construct a dictionary capable of functioning well across these different cases.

⁷Each document in my corpus was assigned a unique ID. I then used the random number generator in R to generate five different numbers to get my random sample of manifestos.

Figure 3: Examples of Populist and Non-Populist Sections in Party Manifestos

Populist

“It is not fair for a young person, with two jobs, stacking supermarket shelves all night to pay secondary tax while a very rich person with millions pays no tax. It’s not fair that you pay tax on savings interest less than the inflation rate and then get slammed with tax on top of that ... The multi-millionaires must pay their share ... Much of the foreshore and seabed can now be privatised, in the interest of a select few who will make money out of it. We’ve been betrayed – all of us – whatever our race ... Secret deals can, and are being done, between people whose only thought is for their own special interests - and their pockets. You have no say. It’s not a fair go.”

New Zealand First, 2011 (New Zealand)

“Our programme is socialist ... we believe that it is only through a socialist transformation programme, that we will end the suffering of our people ... When in government, the EFF will mobilise the whole society to play a meaningful role in transforming South Africa for the betterment of all ... There will be individuals, trapped by narrow neo-liberal thinking who will claim that only capitalism works. Yet it is evident that capitalism has failed to deliver to our people ... Those who currently own the share of South Africa’s wealth acquired it illegally through colonial wars of dispossession and violent defeat of the majority of South Africa’s people. EFF plans to use political power to realise economic justice and such can only happen through maximum implementation of this Manifesto.”

Economic Freedom Fighter, 2014 (South Africa)

Non-populist

“We will reform government to give more power to people. People who live in this country know that too much power is concentrated in too few hands ... Our governing mission is to break out of the traditional top-down, ‘Westminster knows best approach’, and devolve power and decision-making to people and their local communities ... We will reform institutions and devolve power to deal with the causes of our economic problems, and we will encourage local authorities to innovate to better serve their communities ... We will promote and encourage a model of citizenship based on participation and shared responsibility.”

Labor Party, 2015 (United Kingdom)

“Indigenous Australians have the same right to quality housing as other Australians. Progress in re-housing residents of the appalling town camps in Alice Springs is too slow. Housing is also desperately needed by Indigenous Australians in other remote locations. The Nationals maintain that adequate Indigenous housing and infrastructure is the foundation for increasing self-respect and social cohesion. These lead to improvements in living standards, health and education. They will be high priorities for an incoming federal government. The Nationals support the goal of personal home ownership for all Australians and Indigenous Australians are no exception.”

Nationals, 2013 (Australia)

5.1.2 Training the Word Embedding Model and the Classifier

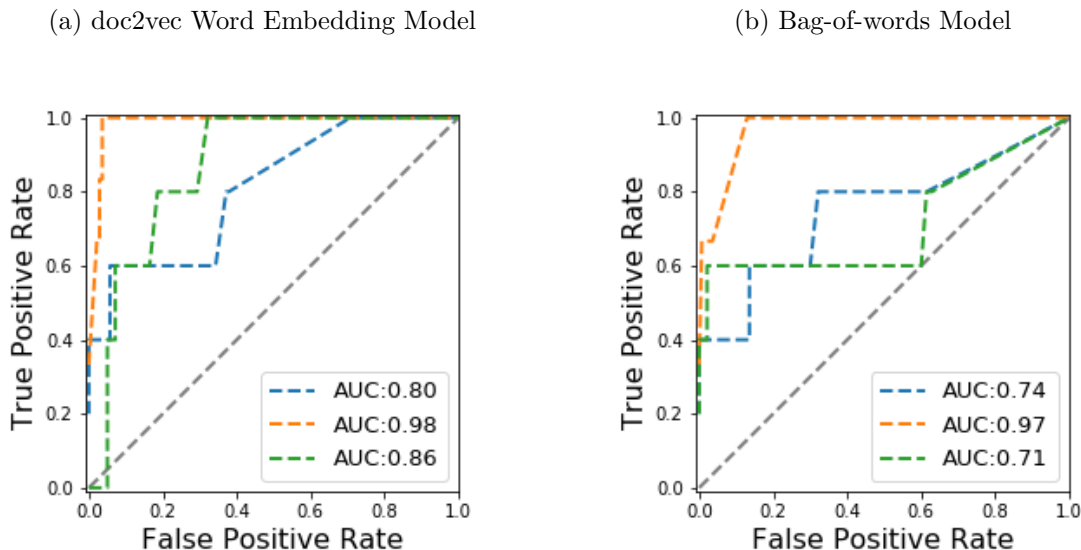
The second step of my supervised method involves training the doc2vec word embedding model to position words and documents (sections) in a 100 dimensional vector space, so that similar words and documents — those that share similar contexts — are positioned close to each other.⁸ However, documents can be similar on many different dimensions. For example, a populist document could be similar to a non-populist document because they both address the same topic, perhaps immigration, foreign policy, or the state of the economy. As a result, it is still necessary to take additional steps to separate populist and non-populist documents.

As I noted earlier, I do this using a random forests classifier. In order to be able to assess the performance of the classifier, I train the model on just two thirds of the hand-coded documents in the training set. I then test its generalizability in classifying the remaining ‘out-of-sample’ documents in the training set. I repeat this process multiple times. On average, my model achieves an accuracy rate of 95%. In other words, my model classifies the out-of-sample documents as populist or non-populist in the same way as the human coder 95% of the time. While this predictive accuracy is high, it is important to remember that there are very few populist documents in the training set. A naïve coder who simply classified all of the out-of-sample documents in the training set as non-populist would achieve an accuracy rate greater than 90%.

To address this point, I also examined the Receiver Operating Characteristic (ROC) curve and calculated the area under the ROC curve (AUC) score to evaluate the model performance (Bradley, 1997). In Figure 4a, I show the ROC curves from three iterations of the random forests classifier. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) in out-of-sample prediction. The TPR captures the proportion of out-of-sample documents correctly classified as populist, while the FPR captures the proportion of out-of-sample documents that were incorrectly classified as populist. A ROC curve for a classifier that perfectly predicts class membership (populist and non-populist) would form a 90 degree angle in the upper left corner of the plot — the TPR would be 1 and the FPR would be 0. A classifier with no predictive power, one that performs randomly, would have a ROC curve that follows the diagonal dashed gray line in Figure 4a. Correspondingly, an AUC of 1 indicates perfect classification and an AUC of 0.5

⁸For more information on why I use a 100 dimensional vector space, see [Online Appendix A](#).

Figure 4: Comparing the Out-of-Sample Performances of the doc2vec Word Embedding Model and the Bag-of-Words Model



Note: Although the models are tested on multiple training-test splits, for a clearer visualization, I plot only three of the random splits: a high performance split, an average performance split, and a poor performance split. Models trained on the same training split are plot in the same color.

indicates random classification. The AUC can also be interpreted as the probability of assigning a higher probability to a true positive case than to a negative case if we were to randomly sample a positive case and a negative case. On average, over all the iterations, my model achieves an AUC score of 0.89, meaning that there is an 89% chance that it predicts a higher probability of a populist case being classified as a populist case than a non-populist case.

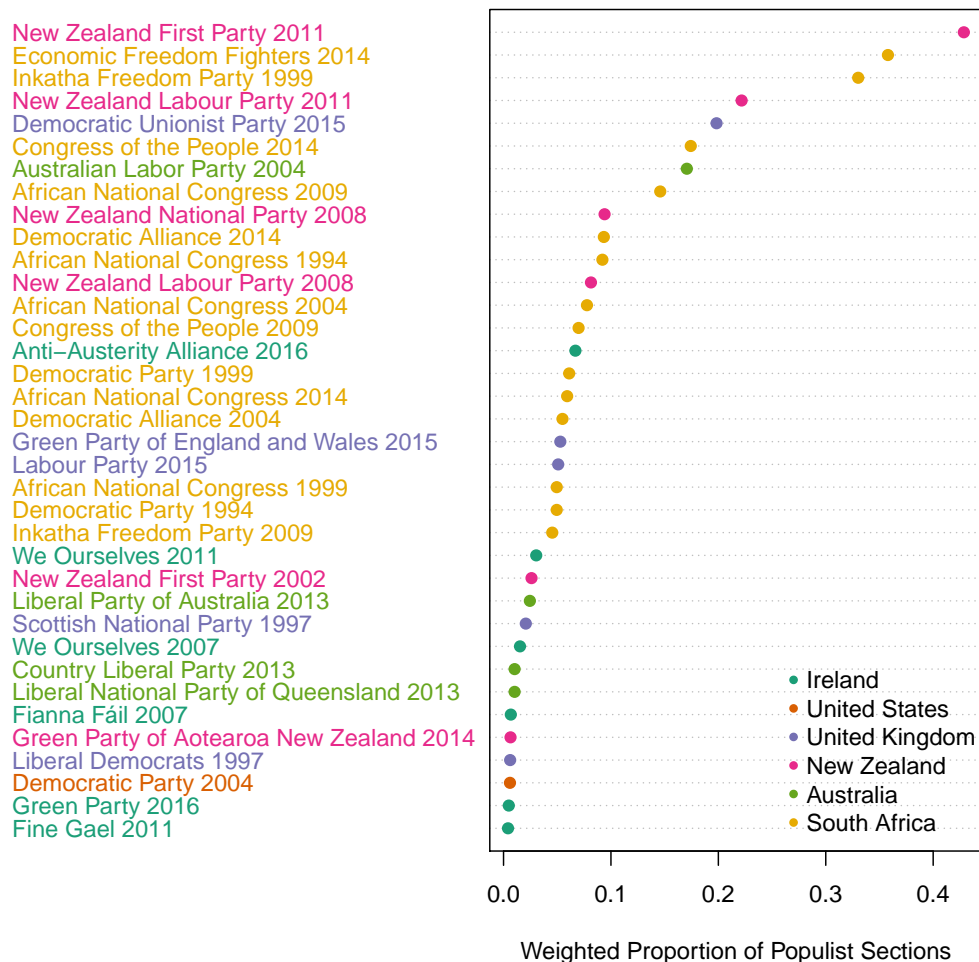
To examine how much the use of the doc2vec word embedding model improves the supervised method, I also trained the random forests classifier on the bag-of-words representation of the manifestos using the same training data. The ROC curves for three iterations are included in Figure 4b. On average, the random forests classifier trained on the bag-of-words representation of the texts achieves an out-of-sample AUC score of 0.8, which is about 0.1 less than the classifier trained on the doc2vec word embedding model. The classifier trained on the word embedding model consistently outperforms the classifier trained using the bag-of-words representation. In addition to having higher AUCs, the classifier trained using word embedding tends to suffer a smaller trade-off between true positives and false positives. In other words, we can get a much higher true positive

rate when we tolerate a relatively small false positive rate in the model using word embedding.

5.1.3 Applying the Model to the Full Corpus

The final step in the supervised method is to apply the model trained to identify populist and non-populist documents in the hand-coded training set to the full manifesto corpus. One thing we might want to know is the extent to which manifestos use populist discourse. Because I treat each section in a manifesto as a single document, I need to aggregate the measure of populism at the section level up to the manifesto. To identify the degree of populism in a manifesto, I simply use the proportion of populist sections in a manifesto weighted by the length of the sections. In Figure 5, I show the extent to which manifestos use populist discourse for those manifestos with at least one populist section. Of the 119 party manifestos in the full corpus, 36 (30%) contain

Figure 5: Party Manifestos with at least One Populist Section



at least one populist section. These manifestos are written by 24 distinct parties. Each color in Figure 5 represents a different country. According to my measure, the most populist manifesto among English-language manifestos is New Zealand First’s 2011 manifesto. As Figure 5 indicates, there is lot of variation in the percentage of populist sections across manifestos. Almost half of the sections in New Zealand First’s 2011 manifesto were coded as populist, whereas less than 2% of the Fine Gael’s 2011 manifesto were coded as populist. There is also evidence that the degree of populist discourse varies over time within the same party. For example, New Zealand First’s 2011 manifesto is much more populist than its 2002 manifesto. The ability to measure populist discourse across and within parties over time opens up new opportunities for understanding the causes and consequences of populism around the world. Interestingly, the preliminary results shown in Figure 5 show that half of the top ten most populist manifestos are written by political parties in South Africa. With a few exceptions (Cheeseman, Ford and Simutanyi, 2014; Cheeseman and Larmer, 2009; Resnick, 2010, 2016), there has been relatively little research on populism in Africa.

5.2 Measuring Populist Discourse in China’s State-controlled Media

Although populism has been studied in many different parts of the world, there is almost no research on populism in China. This is surprising as populism forms the basis for Mao’s famous ‘Mass Line’ leadership in which the Chinese leadership should be “for the masses, relying on the masses, from the masses, and to the masses” (Dai and Shao, 2016). Recently, however, Tang (2016) provided an innovative theoretical discussion of populism’s role in regime sustainability in China, and offered empirical evidence to suggest that populism can have a significant effect on public opinion. In his analysis, though, Tang (2016) focuses primarily on the people-centrism aspect of populism, and does not take account of populism’s emphasis on anti-pluralism and moralized politics.

To further explore the study of populism in China, I measure the populist rhetoric used in the daily news reports from China Central Television’s (CCTV) Xinwen Lianbo (Daily News Program) from 2003 to 2015. CCTV is the predominant state television broadcaster in China as well as the main propaganda machine of the Chinese government. The Xinwen Lianbo program broadcasts news every day from 7 p.m. to 7:30 p.m. All of the local TV stations are required to broadcast Xinwen Lianbo on at least one of their local channels during that time period. In fact, for many years, Xinwen Lianbo was the only TV program Chinese audiences could watch at this time in the

evening. In many ways, my examination of populist discourse in China’s state-controlled media follows the same basic approach adopted by Rooduijn (2014), who examines opinion articles in West European newspapers to measure the diffusion of populism in public debates. In the context of authoritarian regimes, state-controlled media acts as a propaganda machine and the mouthpiece of the government. As a result, news reports from state-controlled media, such as CCTV, should provide a fairly accurate reflection of the level of populist discourse used by the government.

5.2.1 Hand-Coding the Training Set: Holistic Grading

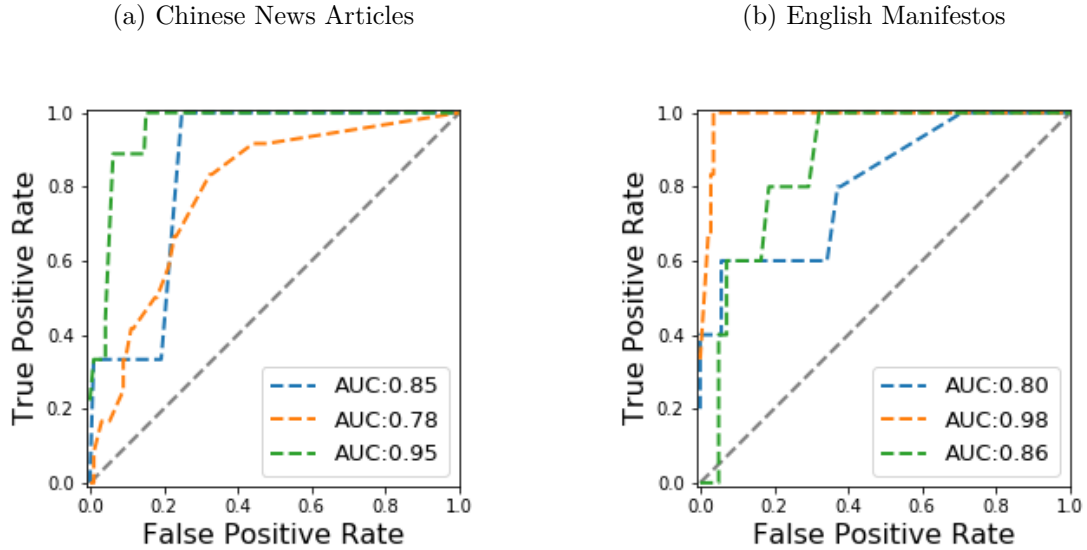
To obtain my training set, I randomly sampled 600 of 115,882 news reports that were broadcast on CCTV’s Xinwen Lianbo from 2003 to 2015.⁹ Each news report was coded using the same holistic grading rubric as before. I treat each news report as one document. I classify a news report as populist (1) if it refers to people-centrism, anti-pluralism, *and* a moralized politics, and non-populist (0) otherwise. Twenty-six (4.3%) of the 600 news reports in the training set contain all three of the necessary components of populism and were coded as populist. As with party manifestos, populist discourse appears to be quite rare in CCTV news reports.

5.2.2 Training the Word Embedding Model and the Classifier

I train the doc2vec word embedding model to position words and documents (news reports) in a 100 dimensional vector space, so that similar words and documents — those that share similar contexts — are positioned close to each other. I then train random forests using a randomly-selected half of the human-coded data from the training set (2-fold cross validation) to learn the relationship between the labeled classes (populist and non-populist) and the document features. I then use the trained model to classify the other half of the hand-coded training data, the test set. I repeated this process ten times and calculated the AUC score in the out-of-sample prediction each time. On average, the model achieves an AUC score of 0.88, which means that there is an 88% chance that it predicts a higher probability of a populist case being classified as a populist case than a non-populist case. The average model performance for the Chinese news reports is almost the same as for the English-language manifestos. In Figure 6, I show the ROC curves from three iterations

⁹Each news report in my corpus was assigned a unique ID. I then used a random number generator in R to generate 600 unique numbers to get my random sample of news reports.

Figure 6: Out-of-Sample Performances in Chinese News Articles and English Manifestos



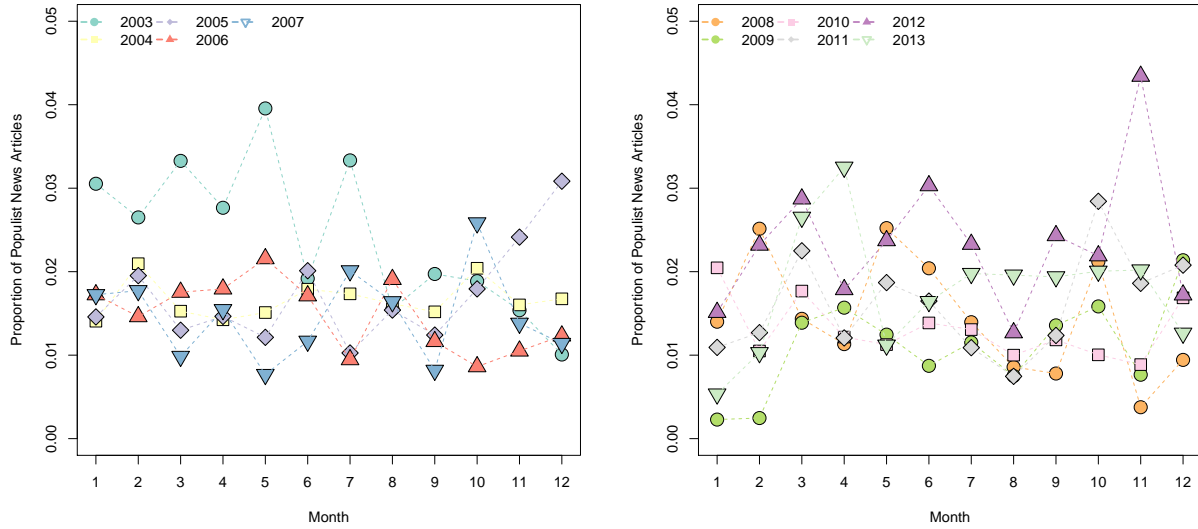
Note: Although the models are tested on multiple training-test splits, for a clearer visualization, I plot only three of the random splits: a high performance split, an average performance split, and a poor performance split. The doc2vec and random forests models in both corpora use the same hyper-parameters: 100 dimensional vector space and a PV-DBOW algorithm with a window size of 10 and a negative sample size of 10 for d2v, 5,000 decision trees, maximum depth of 10 and 2-fold calibration for random forests, training on two thirds of the labeled data.

with relatively high to low performance in Chinese CCTV corpus and English manifestos. As in English manifestos, on average we only need to increase false positive rate to around 0.2 to 0.3 to achieve a true positive rate around 0.9.

5.2.3 Applying the Model to the Full Corpus

In the final step, I apply the model trained on the hand-coded data set to the full CCTV news corpus. Of the 115,882 news articles, 1,880 (1.6%) are predicted to be populist. One advantage of using this highly effective, transparent, and reproducible computer-assisted method is that it is able to produce or restructure the measurement to the level that is appropriate for a particular research purpose. In this application, I use each news article as one document and predict whether each news article is populist or not. We can easily reproduce the measurement to a different level of analysis, though. For example, we can also aggregate the measure to the level of days, months, or years. In Figure 7, I show the proportion of populist news report in each month in the Chinese Central TV's Daily News Program from 2003 to 2013. The proportion of populist news reports in each

Figure 7: Populism in Chinese State Controlled Media



Note: For easier interpretation, I split the proportion of populist news articles from 2003 to 2013 in two graphs in Figure 7. Each dot represents the proportion of populist news articles in a given month. Each line represents the degree of populism in a year.

month ranges from 1.2% to 4.3%. Each line shows the trend of populism in the news every year. The degree of populism varies monthly and yearly. The variation requires further investigation and opens up the opportunity to examine why and how the Chinese government uses populist rhetoric.

6 Evaluation

In this section, I evaluate my method in terms of its validity and reliability, and compare it with other existing methods.

6.1 Validity

Validity refers to the extent to which a measure reflects the concept it is intended to capture (Adcock, 2001; Weber, 1990). In other words, do the coding rules that we have employed capture the concept of populism (construct validity)? Are the populist texts coded by the method actually populist, and are the non-populist texts in the results actually not populist (face validity)? How does the new measure correlate with other established measures of populism (concurrent validity)?

While there is no clear procedure or standard for assessing the validity of a dictionary in the dictionary-based method, we can easily validate a supervised model in all three steps of the process. In the first step, we have clear and transparent written coding rules based on our theoretical concept of populism. In the second and third steps, we can evaluate the validity of our measure by comparing the outputs of the machine coding and the human coding. In the second step where I train the classifier, I used cross-validation to evaluate the possible models based on their out-of-sample prediction. I divided the training set into two subsets and built the model on one subset of the training data. I then used the model to predict the other subset. Because I had already hand coded the training set, I was able to examine whether the model predicts high probabilities for populist texts and low probabilities for the non-populist texts. I repeated the process multiple times. On average, there is an 89% chance that the model predicts a higher probability for a populist document than a non-populist document in two different types of corpus in two different languages. In addition to cross validation in the training process, after measuring populism in the whole corpus, we can also apply a blind test to further evaluate the model's performance.

6.2 Reliability

Reliability refers to the extent to which the measurement process produces the same results when applied multiple times. In the human-coded method, it refers to the inter-coder reliability, which captures the extent to which different human coders code a text the same way when using the same coding rule. The inter-coder reliability is often influenced by the clarity of the coding rules and how well the human coders are trained. In this particular case, I coded the training set myself as a pilot experiment. Thus, I cannot calculate the inter-coder reliability at this current stage. However, when introducing the holistic grading method to the measurement of populism, [Hawkins \(2009\)](#) shows that the correlation between the scores produced by different human coders are quite high: $r = 0.79$ for all the individual texts and 78% of the time human coders produced exactly the same score for the same case. I expect that in the future when I hire human coders for the same cases, I will obtain similar levels of reliability. In the supervised method, once the training set is coded, the classifier trained on the training set should return the same predicted probability for the same case.

Calculating the reliability of a dictionary method, on the other hand, is tricky. Although

using the same dictionary would always give the same results, different researchers are likely to develop quite different dictionaries even when they are following the same definition of populism. It is also hard to find or train qualified coders to independently build dictionaries, as developing a dictionary requires significantly more knowledge and skills than training human coders in holistic grading. To develop a dictionary, one needs to understand populism, to be able to read and write in the language of the country, and also to be familiar with, or even be expert in, a country’s politics and history.

7 Conclusion

In this paper, I clarify the theoretical concept of populism and introduce a supervised approach with word embedding models for measuring populism in political texts. While the study of populism is comparative in nature, a measure of populism that applies over time and space is still lacking. Human-coded methods tend to be more content valid and more generalizable, but are costly and less efficient. Existing automated dictionary-based methods are more efficient but hard to validate and apply to a broader set of cases. However, this does not mean that the existing measures of populism that are based on the dictionary-based method are not valid. In fact, the existing dictionaries are often carefully built by experts on populism and a specific country’s history and politics. Dictionary-based methods have been shown to produce valid measures when carefully conducted. However, we are not just interested in the degree of populism in a single case at a specific point in time. The dictionary-based method becomes problematic when we want to apply the method to different countries or the same country over time.

The supervised approach combined with holistic grading and a word embedding model proposed here is easier to validate and easier to apply to different cases and over time. The coding rule in holistic grading is centered around the concepts of people-centrism, anti-pluralism, and moralized politics, rather than the specific identities of the people and the enemy, which makes it more content valid and more generalizable over space and time. Combining the holistic grading with the supervised method increases the efficiency of the holistic grading, and still has the merit of being easily applied to different contexts.

Another advantage of my supervised approach that makes it more generalizable is that it is

not constrained by a particular language. There are two different ways to expand my measure of populism to other languages without language translation. First, we can hand code a small amount of training data in other languages based on the holistic grading in step one of the supervised method. I have demonstrated this approach in measuring populism in two different languages and two different types of political discourse. The only requirement of this approach is to have a small set of hand-coded documents in each relevant language. Because the hand coding in each language follows the same coding rules, the measurement in each language should be comparable. Second, we can expand the single language doc2vec model to multilingual word embedding (Dai and Radford, 2018). In the doc2vec model, similar words and documents are mapped closer to each other in the vector space. Similarly, multilingual word embedding seeks to map the similar words in different languages close to each other in the vector space. Once we map different languages into the same vector space, we can use the model trained in one language to measure populism in other languages.

Moreover, because the supervised method classifies documents by learning the relationship between the labeled categories and the documents, it can also potentially be used to classify different sub-types of populism, such as the inclusive type and the exclusive type (Mudde and Rovira Kaltwasser, 2013). The process of measuring the sub-types of populism would involve the same three steps I outlined in the paper: first, we need to develop coding rules for the sub-types that we are interested in and hand code a small sample of documents as a training set; second, we need to train a classifier that learns the relationship between the categories and the documents in the training set; third, we can then apply the trained model in the second step to the full corpus.

References

- Acemoglu, Daron, Georgy Egorov and Konstantin Sonin. 2013. “A political theory of populism.” *Quarterly Journal of Economics* 128(2):771–805.
- Adams, James, Lawrence Ezrow and Zeynep Somer-Topcu. 2011. “Is anybody listening? Evidence that voters do not respond to European parties’ policy statement during elections.” *American Journal of Political Science* 55(2):370–382.
- Adcock, Robert. 2001. “Measurement validity: A shared standard for qualitative and quantitative research.” *American Political Science Review* 95(3):529–546.
- Akkerman, Tjitske and Matthijs Rooduijn. 2017. “Flank attacks: Populism and left-right radicalism in Western Europe.” *Party Politics* 23(3):193–204.
- Blei, David, Andrew Y. Ng and Michael Jordan. 2003. “Latent Dirichlet allocation.” *Journal of Machine Learning* 3:993–1022.
- Bonikowski, Bart and Noam Gidron. 2016a. “Multiple traditions in populism research: Toward a theoretical synthesis.” *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 26(2):7–15.
- Bonikowski, Bart and Noam Gidron. 2016b. “The populist style in American politics: Presidential campaign discourse, 1952–1996.” *Social Forces* 94(4):1593–1621.
- Boser, Bernhard E., Isabelle M. Guyon and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM pp. 144–152.
- Bradley, Andrew P. 1997. “The use of the area under the ROC curve in the evaluation of machine learning algorithms.” *Pattern recognition* 30(7):1145–1159.
- Breiman, Leo. 2001. “Random forests.” *Machine learning* 45(1):5–32.
- Cheeseman, Nic and Miles Larmer. 2009. “Ethnopolitism in Africa: Opposition mobilization in diverse and unequal societies.” *Democratization* 22(1):22–50.
- Cheeseman, Nic, Robert Ford and Neo Simutanyi. 2014. Is there a ‘populist threat’ in Zambia? In *Zambia: Building Prosperity from Resource Wealth*, ed. Christopher Adam, Paul Collier and Michael Gondwe. New York: Oxford University Press.
- Crabtree, Charles, Matt Golder, Thomas Gschwend and Indriði H. Indriðason. 2018. “It’s not only what you say, it’s also how you say it: The strategic use of campaign sentiment.” Unpublished manuscript, Pennsylvania State University.
- Dai, Yaoyao and Benjamin Radford. 2018. “Multilingual word embedding for zero-shot text classification.” Paper presented at the Annual Meeting of the American Political Science Association, Boston.
- Dai, Yaoyao and Zijie Shao. 2016. “Populism as a strategy of authoritarian survival in China: Concept and measurement.” *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 26(2):31–40.

- D’Orazio, Vito, Steven T Landis, Glenn Palmer and Philip Schrodtt. 2014. “Separating the wheat from the chaff: Applications of automated document classification using support vector machines.” *Political Analysis* 22(2):224–242.
- Dornbusch, Rudiger and Sebastian Edwards. 1991. The macroeconomics of populism. In *The Macroeconomics of Populism in Latin America*, ed. Rudiger Dornbusch and Sebastian Edwards. Chicago: University of Chicago Press pp. 7–13.
- Freund, Yoav and Robert E. Schapire. 1997. “A decision-theoretic generalization of on-line learning and an application to boosting.” *Journal of Computer and System Sciences* 55(1):119–139.
- Gidron, Noam and Bart Bonikowski. 2013. “Varieties of populism: Literature review and research agenda.” Paper Series 13-0004, Weatherhead Center for International Affairs, Harvard University.
- Golder, Matt and Sona Nadenichek Golder. 2016. “Populism in a comparative perspective.” *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 26(2):1–131.
- Green, Jane and Sara B. Hobolt. 2008. “Owning the issue agenda: Party strategies and vote choices in British elections.” *Electoral Studies* 27(3):460–476.
- Grimmer, Justin and Brandon M. Stewart. 2013. “Text as data: The promise and pitfalls of automatic content analysis methods for political texts.” *Political Analysis* 21(3):267–297.
- Hawkins, Kirk A. 2009. “Is Chávez populist? Measuring populist discourse in comparative perspective.” *Comparative Political Studies* 42(8):1040–1067.
- Hawkins, Kirk A. and Bruno Castanho Silva. 2018. Textual analysis: Big data approaches. In *The Ideational Approach to Populism: Concept, Theory, and Analysis*, ed. Kirk A. Hawkins, Ryan E. Carlin, Levente Littvay and Cristóbal Rovira Kaltwasser. New York: Routledge pp. 44–74.
- Hawkins, Kirk A and Cristóbal Rovira Kaltwasser. 2017. “The ideational approach to populism.” *Latin American Research Review* 52(4):513–528.
- Hawkins, Kirk A. and Cristóbal Rovira Kaltwasser. 2018. Introduction: Concept, theory, and method. In *The Ideational Approach to Populism: Concept, Theory, and Analysis*, ed. Kirk A. Hawkins, Ryan E. Carlin, Levente Littvay and Cristóbal Rovira Kaltwasser. New York: Routledge pp. 1–74.
- Hawkins, Kirk A., Ryan E. Carlin, Levente Littvay and Cristóbal Rovira Kaltwasser, eds. 2018. *The Ideational Approach to Populism: Concept, Theory, and Analysis*. New York: Routledge.
- Hill, Daniel W and Zachary M. Jones. 2014. “An empirical evaluation of explanations for state repression.” *American Political Science Review* 108(3):661–687.
- Jagers, Jan and Stefaan Walgrave. 2007. “Populism as political communication style: An empirical study of political parties’ discourse in Belgium.” *European Journal of Political Research* 46(3):319–345.
- Janda, Kenneth, Robert Harmel, Christine Edens and Patricia Goff. 1995. “Changes in party identity: Evidence from party manifestos.” *Party Politics* 1(2):171–196.

- Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pp. 1188–1196.
- Lehmann, Pola, Jirka Lewandowski, Theres Matthieß, Nicolas Merz, Sven Regel and Annika Werner. 2017. “Manifesto Corpus.” Berlin: WZB Berlin Social Science Center.
- Levitsky, Steven and Kenneth M. Roberts. 2011. Conclusion: Democracy, development, and the left. In *The Resurgence of the Latin American Left*. Baltimore: Johns Hopkins University Press pp. 399–428.
- Mair, Peter and Cas Mudde. 1998. “The party family and its study.” *Annual Review of Political Science* 1:211–229.
- Mason, Llew, Jonathan Baxter, Peter L. Bartlett and Marcus R. Frean. 2000. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems*. pp. 512–518.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, ed. C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger. pp. 3111–3119.
- Moffitt, Benjamin. 2016. *The Global Rise of Populism: Performance, Political Style, and Representation*. Stanford: Stanford University Press.
- Mudde, Cas. 2004. “The populist zeitgeist.” *Government and Opposition* 39(4):541–563.
- Mudde, Cas. 2007. *Populist Radical Right Parties in Europe*. New York: Cambridge University Press.
- Mudde, Cas and Cristóbal Rovira Kaltwasser. 2013. “Exclusionary vs. inclusionary populism: Comparing contemporary Europe and Latin America.” *Government and Opposition* 48(2):147–174.
- Mudde, Cas and Cristóbal Rovira Kaltwasser. 2017. *Populism: A Very Short Introduction*. New York: Oxford University Press.
- Mudde, Cas and Cristóbal Rovira Kaltwasser. Forthcoming. “Studying populism in comparative perspective: Reflections on the contemporary and future research agenda.” *Comparative Political Studies*.
- Müller, Jan-Werner. 2016. *What is Populism?* Philadelphia: University of Pennsylvania Press.
- Ostiguy, Pierre. 2009. “The high-low political divide: Rethinking populism and anti-populism.” Political Concepts: Committee on Concepts and Methods Working Paper Series, 35.
- Ostiguy, Pierre. 2017. Populism: A socio-cultural approach. In *The Oxford Handbook of Populism*, ed. Cristobal Rovira Kaltwasser, Paul A. Taggart, Paulina Ochoa Espejo and Pierre Ostiguy. New York: Oxford University Press pp. 73–97.
- Peterson, Andrew and Arthur Spirling. 2018. “Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems.” *Political Analysis* 26(1):120–128.
- Pitkin, Hanna F. 1967. *The Concept of Representation*. University of California Press.

- Radford, Benjamin James. 2016. Automated Learning of Event Coding Dictionaries for Novel Domains with an Application to Cyberspace PhD thesis Duke University.
- Resnick, Danielle. 2010. *Urban Poverty and Party Populism in African Democracies*. New York: Cambridge University Press.
- Resnick, Danielle. 2016. “Two generations of African populism: Applying a cumulative conceptual approach.” *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 26(2):62–69.
- Roberts, Kenneth M. 2006. “Populism, political conflict, and grass-roots organization in Latin America.” *Comparative politics* 38(2):127–148.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley and Edoardo M. Airoidi. 2013. “The structural topic model and applied social science.” *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation* .
- Rooduijn, Matthijs. 2014. “The mesmerising message: The diffusion of populism in public debates in Western European media.” *Political Studies* 62(4):726–744.
- Rooduijn, Matthijs, Sarah L. de Lange and Wouter van der Brug. 2014. “A populist zeitgeist? Programmatic contagion by populist parties in Western Europe.” *Party Politics* 20(4):563–575.
- Rooduijn, Matthijs and Teun Pauwels. 2011. “Measuring populism: Comparing two methods of content analysis.” *West European Politics* 34(6):1272–1283.
- Saward, Michael. 2006. “The representative claim.” *Contemporary Political Theory* 5(3):297–318.
- Saward, Michael. 2014. “Shape-shifting representation.” *American Political Science Review* 108(4):723–736.
- Tang, Wenfang. 2016. *Populist Authoritarianism: Chinese Political Culture and Regime Sustainability*. New York: Oxford University Press.
- Weber, Robert Philip. 1990. *Basic Content Analysis*. London: Sage.
- Weyland, Kurt. 2001. “Clarifying a contested concept: Populism in the study of Latin American politics.” *Comparative politics* 34(1):1–22.
- Weyland, Kurt. 2017. Populism: A political-strategic approach. In *The Oxford Handbook of Populism*. New York: Oxford University Press.

Online Appendix A: Word Embedding Models

Word embedding is a type of language model that maps words or sentences and documents into vectors of real numbers. A more common way of vectorizing words and documents is the ‘bag-of-words’ approach, which represents each word as a unique dimension (one-hot vector with 1 in that word’s dimension, 0 in all other word dimensions) and represents documents as counts of each unique word (sum of one-hot vectors). The order of words is assumed to be irrelevant to the analysis. The semantic and syntactic meaning of words are also lost in ‘bag-of-words’ approach.

Unlike the ‘bag-of-words’ method of vectorization, in which one word is one dimension, word embedding represents words and documents in a dense continuous vector space with many fewer dimensions and positions semantically and syntactically similar words close to each other in this vector space. The method of word embedding is based on a distributional hypothesis in linguistics theory, which states that the meaning of a word is a function of its contexts or surrounding words. Unlike the ‘bag-of-words’ assumption, which treats words as independent atomic units, the distributional hypothesis aims to model the meaning of a word and assumes that the meaning of a word is given, and can be approximated, by the sets of contexts in which the word appears. In effect, the underlying idea is that words that frequently appear in same contexts are likely to have a similar meaning.

There are several different ways to train word embedding. In this paper, I use a doc2vec model (Le and Mikolov, 2014), which is based on the more foundational word2vec model (Mikolov et al., 2013). I begin by describing the word2vec model. The word2vec model is a neural network based model that takes each unique word in the vocabulary of a corpus as an input. The input word, represented as a one-hot vector, is then multiplied by a dense, real-valued weights matrix of size $V \times d$, where V is the length of the vocabulary in the corpus and d is the chosen size of the hidden layer or ‘embedding’.¹ By multiplying the $1 \times V$ input vector for a word with the $V \times d$ weights matrix, a $1 \times d$ vector is generated; this is the word’s vector representation, v_{word} . The model then uses this vector representation of the input target word as the input to a softmax classifier to predict which of the V words in the vocabulary are likely to be the context words of the input word. Context words are those that appear in a certain range of words before and after

¹I choose $d = 100$, in keeping with standard practice.

the current/target word. The model learns the embedding or the parameters in the hidden layer by finding the parameters that maximize the predicted probability of true context words. In other words, the word2vec model seeks to set parameters θ to maximize the conditional probability of contexts C when observing the target word T : $p(C|T; \theta)$ for all words in the vocabulary (Mikolov et al., 2013; Mikolov, Chen, Corrado and Dean, 2013; Goldberg and Levy, 2014).² Therefore, mathematically, the model assigns similar parameters to words that are used interchangeably in the same contexts.

Because maximizing $p(C|T; \theta)$ for all target and possible contexts is expensive to compute and there are more words that do not appear together than words that often appear together, I adopt negative sampling skip-gram in training the model. In negative sampling skip-gram, the input layer contains target-context word pairs. The target-context pairs are generated by taking the target word at index i and pairing it with all context words from $i - k$ to $i + k$ given a window size k .³ For every true target-context word pair, we generate s negative samples; these are target-context word pairs that are not observed in the actual text corpus.⁴ The output layer contains dummy values 1 and 0 indicating whether the input pair is a true target-context pair that co-locates in the texts (1) or a negative/fake pair that does not appear together in the texts (0). The predicted value given an input pair is computed by taking the dot product of the target word vector (target embedding) and the context word vector (context embedding) and then applying the logistic function, $\sigma(\cdot)$. The model uses small non-zero random values as the initial parameters in the hidden layer to produce the embedding/word vector. Stochastic gradient descent is then used to optimize the parameters through back-propagation to minimize the logarithmic loss between $\sigma(v_{\text{target word}} \cdot v_{\text{context word}})$ and the true value $[0, 1]$.

Expanding the word2vec model to the document level is simple; each document is labeled with an ID, and treated as one unit (like a word). This document ID is positioned within the text in the document. For example, suppose we have a one-sentence document labeled as Doc1: “Our Party should serve the people whole-heartedly.” The document ID is treated as one unit and positioned

²Word2vec encompasses two different, related neural-network based models, including the continuous bag-of-words (CBOW) and skip-gram (SG) models (Mikolov, Chen, Corrado and Dean, 2013). The SG model, which is used and explained here, inputs a target word from a text and attempts to predict the target word’s likely context words. The CBOW model does the reverse. Given a set of context words, CBOW attempts to predict the context’s target word.

³The actual window size is randomly sampled, per target word, from 1 to k . This effectively causes more distant context words to be sampled with less frequency than nearer context words. In my model I use $k = 10$.

⁴In my model, I used $s = 10$.

within its text: “Our Party should serve *Doc1* the people whole-heartedly.”⁵ The negative sampling algorithm can now be applied to both the target word and the document, which is treated as a target word. In this way, the documents sharing similar texts or content are positioned close to each other in the vector space (Le and Mikolov, 2014).

⁵In practice, the model is adjusted so that the *Doc1* token occurs in all of document 1’s words’ contexts and all of the document 1’s words appear in the *Doc1* token’s context.

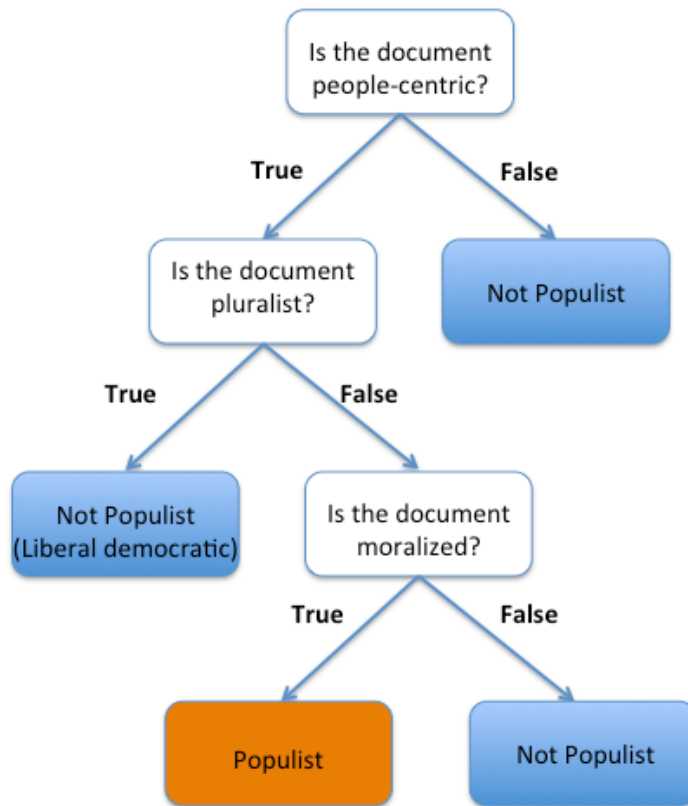
Online Appendix B: Random Forests Classification Algorithm

In this paper, I treat the problem of measuring populism as a classification problem, thereby making it a supervised learning problem.⁶ Some discourses are populist and others are not. To measure populism, therefore, we simply need to sort the populist discourses from the non-populist discourses. In the main text, I briefly explain why supervised learning is more suitable for this type of measurement problem (see footnote 5), and I outline the general steps of the supervised method. Here, I explain the algorithm used to perform the supervised classification tasks in more detail. There are many classification algorithms, such as support vector machines (SVM) (Boser, Guyon and Vapnik, 1992; D’Orazio et al., 2014), naïve Bayes, decision trees, an ensemble of decision trees (random forests and boosted decision trees) (Breiman, 2001; Hill and Jones, 2014; Mason et al., 2000; Freund and Schapire, 1997), and neural networks (Peterson and Spirling, 2018). Each algorithm has its strengths and weaknesses, and no single algorithm definitely outperforms all of the others across all of the possible evaluative criteria. With neural networks, for example, we can build very complex models that are particularly good when we have a large dataset with many features. However, neural networks are sensitive to the scaling of data and to the choice of hyperparameters. Another classification algorithm, random forests, is robust and scaling insensitive, but tends to perform worse when we have large datasets with many features (very high dimensional sparse data). In what follows, I focus on the random forests used in my supervised method.

The random forests algorithm is a kind of ensemble of decision trees. To understand how the random forests algorithm works, we first need to understand decision trees. Intuitively, we can consider the decision tree algorithm as mimicking the human decision-making process. For example, when deciding whether a certain document is populist or not, a human coder’s decision process might follow a decision tree similar to the one shown in Figure 8. First, the human coder decides whether the document depicts the people as the only legitimate source of power. If this is the case, then the human coder needs to decide whether the people are pluralist or not. If the people are pluralist, then the document would be classified as liberal democratic. If the people are not pluralist, then the human coder needs to identify whether the document structures politics along moralized lines. If politics is structured along moralized lines, then the human coder will

⁶In addition to classification, the other supervised learning problem is regression.

Figure 8: An Example of a Decision Tree used by a Human Coder



classify the document as populist. The decision tree algorithm attempts to learn a hierarchy of if/else questions, like the one shown in Figure 8, that leads it to classification decisions that match those made by a human coder. The researcher does not need to identify the rule used in each node of the hierarchy for the machine. Instead, the decision tree algorithm learns those rules from the labeled data in the training set, which in our case comprises the documents that have been hand-coded into populist and non-populist categories using holistic grading.

But how does the decision tree algorithm learn to make the rules? To grow a single decision tree in the random forest, the algorithm first randomly bootstraps a sub-sample of the training data. Second, it randomly selects a sample of features. The features are the 100 dimensions used to represent documents in the doc2vec word embedding approach or the unique words in the bag-of-words approach. Third, it finds the best pair {feature, feature threshold} to split the data into two subsets, in our case populist and non-populist, that are the ‘purest’ among all the possible

splits by minimizing the cost function of the ‘impurity’ measure shown in Equation 1.

$$J(k, t_k) = \frac{m_a}{m} G_a + \frac{m_b}{m} G_b \quad (1)$$

where $\begin{cases} G_a = 1 - \sum_{t=1}^n (p_{a,k})^2 \text{ is the gini impurity measure at subsample/node } a \\ m_a \text{ is the number of instances/cases at subsample } a \text{ after the split.} \end{cases}$

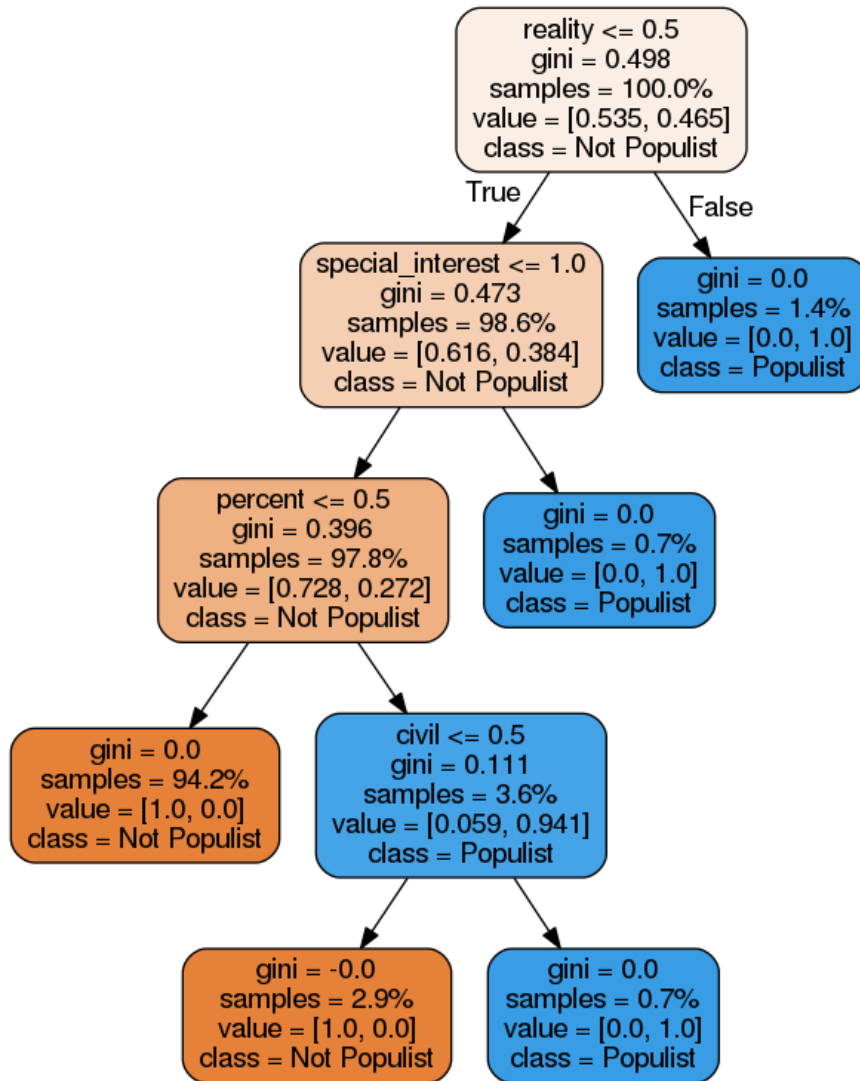
After a split is conducted, the data (documents) are sorted into two different nodes. In each ‘daughter’ node, the algorithm repeats steps two and three to continue splitting the data until a stop-point criteria is met, such as when the training set is perfectly sorted (into populist or non-populist classes) or a pre-determined maximum depth of the tree is reached (Strobl, Malley and Tutz, 2009; Hill and Jones, 2014).

At the end of this process, we have a decision tree in which each document has been classified into one of the possible classes, in our case populist or non-populist. This whole process is then repeated many, many times, resulting in a large number of decision trees.⁷ Together these trees make a ‘random forest’. Because each decision tree is only trained on a random sample of the training set and each split feature node is selected from a random sample of the features, this algorithm is less likely to overfit the training data. Each decision tree is a ‘weak learner’ meaning that it is only slightly better than a random guess. However, the ensemble of them results in powerful classifiers. Once we build the random forest using the training data, we can use it to classify new documents. To classify a new document, each decision tree in the random forest makes a prediction based on its hierarchical rules and the document’s features. The final prediction on the class of the new document is made using a majority rule, which is the most commonly predicted class among all the decision trees. Thus, if the new document ends up in the populist class in more than half of the trees in the random forest, then it is predicted to be a populist document by the random forest. Because the predicted class membership for a given observation is an ensemble of a large number of decision trees or weak classifiers, we can also calculate the uncertainty of the prediction based on the distribution of the predictions from all the decision trees.

One additional advantage of the random forests algorithm is that it can be visualized and is

⁷In my model, I used 5,000 trees.

Figure 9: An Example of a Decision Tree in a Random Forest



easier to interpret than other supervised learning algorithms. However, since my model represents the words and texts in my training set using word embedding, which has 100 dimensional real value vectors, it is hard to interpret. Therefore, to demonstrate the random forests algorithm, I present an example of a decision tree in a random forest trained using the ‘bag-of-words’ approach. The decision tree in Figure 9 is trained using a random sample of hand-coded training data. In the beginning, 54% documents in the bootstrapped training sample are not populist. The task is to sort the populist documents into the populist class and the non-populist documents into the non-populist class. The features that the algorithm uses to do the sorting are the words/terms

frequencies in each document. The first node ‘reality’ with a threshold of 0.5 is selected to split the data, because among the random sample of features and possible thresholds, ‘reality’ with a threshold of 0.5 results in the purest sub-samples. The sub-sample of documents with a value on ‘reality’ above the 0.5 threshold are all populist in the training sample, which is a pure populism group. Therefore, there is no need for further splitting. For the documents that have a value on ‘reality’ that is below the threshold, there is still a mixture of populist and non-populist documents. A daughter node, ‘special_interest’ with a threshold of 1 gives the purest sub-samples and is selected to split the data. The documents that have values on ‘special_interest’ above the threshold of 1 are all populist documents. The process is repeated two more times, and a stopping criteria is reached. In this case, after splitting at the end node, the remaining data splits into a pure non-populist group and a pure populist group; there is no need to split the data further. This tree built using the training sample is then used to predict new documents. If the new document mentions ‘reality’ above the 0.5 threshold, it is classified as populist by this tree. If not, the tree will move the next node until a decision is made. A random forest consists of thousands of decision trees like the one shown in Figure 9. To classify a single case, the random forest algorithm predicts the class using the most commonly predicted class for that case among all of the decision trees (majority vote).

Online Appendices: References

- Boser, Bernhard E., Isabelle M. Guyon and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM pp. 144–152.
- Breiman, Leo. 2001. “Random forests.” *Machine learning* 45(1):5–32.
- D’Orazio, Vito, Steven T Landis, Glenn Palmer and Philip Schrodt. 2014. “Separating the wheat from the chaff: Applications of automated document classification using support vector machines.” *Political Analysis* 22(2):224–242.
- Freund, Yoav and Robert E. Schapire. 1997. “A decision-theoretic generalization of on-line learning and an application to boosting.” *Journal of Computer and System Sciences* 55(1):119–139.
- Goldberg, Yoav and Omer Levy. 2014. “Word2vec explained: Deriving Mikolov et al.’s negative-sampling word-embedding method.” *arXiv preprint arXiv:1402.3722* .
- Hill, Daniel W and Zachary M. Jones. 2014. “An empirical evaluation of explanations for state repression.” *American Political Science Review* 108(3):661–687.
- Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pp. 1188–1196.
- Mason, Llew, Jonathan Baxter, Peter L. Bartlett and Marcus R. Frean. 2000. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems*. pp. 512–518.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, ed. C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger. pp. 3111–3119.
- Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. “Efficient estimation of word representations in vector space.” *arXiv preprint arXiv:1301.3781* .
- Peterson, Andrew and Arthur Spirling. 2018. “Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems.” *Political Analysis* 26(1):120–128.
- Strobl, Carolin, James Malley and Gerhard Tutz. 2009. “An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.” *Psychological Methods* 14(4):323.