Distributed representations of politicians

Bobbie Macdonald
Department of Political Science
Stanford University
bmacdon@stanford.edu

Abstract

Methods for generating dense embeddings of words and sentences have grown rapidly in prominence over the past few years. However, social scientists are often more interested in understanding the authors of text – such as politicians, journalists, and thinktanks – rather than individual words, sentences, and paragraphs. While existing methods such as `word2vec` and `doc2vec` can be easily aggregated to author representations, it is unclear whether this aggregation step would produce meaningful vector representations. In this study, we examine the performance of several of these methods at generating dense vector representations of Kenyan politicians based on a corpus of legislative speeches between 1998 and 2012. We evaluate the resulting vectors on several tasks, with disappointing results. We conclude that despite the recent enthusiam about distributed representations for text, more work is needed to effectively extend these methods to representations of the individuals and organizations that author these texts. This work is especially needed in situations involving small and homogneous corpora – precisely the datasets that social scientists often find themselves working with.

1 Introduction

Elected representatives vary a great deal in how they choose to spend their time in office – differing in how they prioritize societal problems, divide time between legislative and constituency work, allocate government spending within their constituency, incite identity cleavages, et cetera. Collectively, these choices represent a politician's *behavioral style*, having important implications for who benefits from government spending, how constituent preferences are represented, the exercise of executive oversight, and the evolution of national policies.

However, the many observable actions of politicians – such as legislative speeches, local constituency spending, campaign speeches, scandals, and legislative votes – produce sparse high-dimensional representations from which it is difficult to infer underlying differences in behavioral styles. Existing research relies largely on the use of item-response theory (IRT) models to measure the policy preferences – or ideal points – of political actors from roll call votes and political donations (see Carroll et al., 2013; Bonica, 2013; Rosenthal and Poole, 1997). However, in most countries, roll call votes and political donation histories are not publicly available. Moreover, these models overlook a great deal of variation in politicians that occurs in speech and actions outside of the narrow window of legislative voting.

In this study, we attempt to address these shortcomings by constructing dense representations of politicians based on the text of legislative speeches. Specifically, we apply `word2vec`, `doc2vec`, and other methods for inferring dense vector representations to a novel dataset of legislative speeches in Kenya between 1998 and 2012. Overall, our results are disappointing, standing in stark contrast to widespread enthusiasm about distributed vector representations for a wide range of natural language

processing tasks (e.g. Baroni et al., 2014). We find little evidence that word2vec, doc2vec, or latent dirichlet allocation (LDA) out-perform a simple bag-of-words representation in several tasks.

That said, we believe that this topic is in need of far greater attention. Meaningful vector representations of politicians would be immensely useful for forecasting future behavior and events (e.g. scandals, policies, campaigns), detecting aberrations in individual behavior, and deepening our understanding of political cleavages and conflict. Hence, we plan to run additional experiments using deeper neural models with applications beyond Kenyan politicians.

2 Existing work

Methods for inferring dense vector representations of characters, words, and sentences – such as word2vec and GloVe – have received a great deal of attention over the past few years. In particular, a growing number of models provide tools for generating fixed-length dense vector representations from variable-length texts, such as sentences and paragraphs. For instance, Le and Mikolov (2014) introduce doc2vec, a slight variation on word2vec that allows for direct inference of "paragraph vectors", rather than summing/averaging word vectors contained within a paragraph or sentence. Le and Mikolov (2014) describe two variations of doc2vec: distributed memory model of paragraph vectors (doc2vec-dm) and distributed bag of words (doc2vec-dbow). The doc2vec-dm is identical to the continuous bag of words implementation of word2vec, with the exception that a fixed-length paragraph vector is concatenated (or averaged) to the bag of context words before predicting a center word. In this set up, the context words are drawn from a sample window from wtihin the paragraph, where the center word to be predicted is another word within the window. Conversely, in the doc2vec-dbow model, the only input is a fixed length paragraph vector, which is tasked with predicting a randomly sampled word from within the paragraph. Figure 1 illustrates each of these models.

Figure 1: Doc2Vec

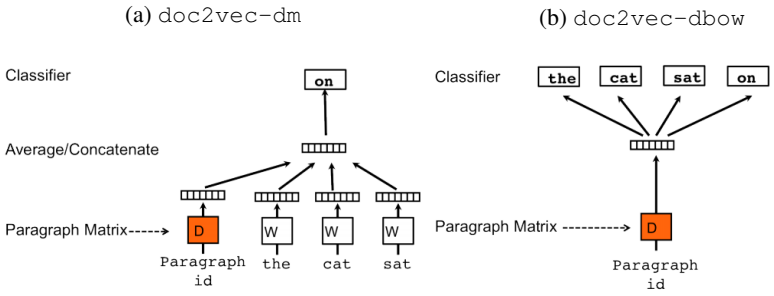

Fig. 1. Figure from Le and Mikolov (2014) displaying the doc2vec-dm and doc2vec-dbow models.

doc2vec is just one model among a rapidly growing number of modeling architectures for inferring sentence and document embeddings. For instance, Kiros et al. (2015) introduce skip-thought vectors trained through a encoder-decoder structure, while Kenter et al. (2016) use a Siamese neural network architecture to construct sentence representations from word embeddings.

However, social scientists often care about higher levels of aggregation – such as "person embeddings" and dense representation of companies or organizations. Very little attention to date has focused on constructing meaningful embeddings at the level of authors and organizations.[1] While methods such as *doc2vec* can be easily extended to infer dense vectors for "mega-paragraphs" representing the collection of, say, all speeches by a politician, all articles written by a journalist, or all reports produced by a thinktank, it is not clear *ex ante* whether existing methods for word- and paragraph-level vector representations will perform well when individuals and/or organizations are the target of inerence.

[1] As an exception, Ganesh et al. (2016) introduce author2vec, combining information from scholars' co-authorship network and the content of abstracts in order to represent authors as dense vectors. However, network data not available in many contexts, severely restricting the scope of potential applications.

Moreover, social scientists tend to work with datasets that are more homogenous and orders of magnitudes smaller than the corpora on which `doc2vec`, `word2vec`, and other existing methods are often trained. For instance, legislative speeches, election campaign materials, and judicial rulings are of great substantive interest, yet contain few observations and little variation relative to online datasets commonly used in NLP applications. As a result, it is unclear whether existing methods for generating word- or document-level representations would perform well when asked to make subtle distinctions betweens individuals in relatively homogenous corpora.

3 Data

Our analyses are based on a new dataset of legislative activity in Kenya between 1998 and 2012, consisting of: (a) legislative speeches; (b) election results; and (c) a set of covariates (e.g. cabinet positions, local development spending). While a growing number of emerging democracies are making unstructured legislative transcripts publicly available, the dataset we present here is the first of its kind for an emerging democracy that is parsed and formatted in a way that can be readily used for statistical analysis. The dataset will be made publicly available soon.[2]

3.1 Speech processing

We began by extracting legislative speeches from transcripts of legislative debates in Kenya's National Assembly between 1998 and 2012, covering Kenya's 8th (1998-2002), 9th (2003-2007), and 10th (2008-2012) parliaments. Legislative debates are recorded and transcribed in the Kenya Hansard (the official record of legislative proceedings) for each day that the legislature is in session.[3] Each transcript contains a sequence of alternating speeches, indicating who is speaking (e.g. "*Mr. Munyao*"), what item/topic is under discussion (e.g. "*Anti-corruption and economic crimes bill, second reading*"), and what was said. We define a single *speech* as the period between when a Member of Parliament (MP) begins to speak and when she is interrupted by the next speaker (or by the end of the transcript). Speeches range from one to several thousand words in length.

In the 8th-10th parliaments, the Kenyan National Assembly[4] consisted of 210 MPs elected in single member constituencies, along with 41 nominated MPs. In this period, the National Assembly sat for 70-80 sessions per year[5] and operated similar to the Westminster system which provides back-benchers and opposition MPs with the opportunity to scrutinize Ministers and Assistant Ministers through questions by private notice (in addition to the usual legislative business of debates over motions, bills, and petitions).

Filtering. The entire speech dataset contains a total of $412,582$ speeches, which we filter by two criteria. First, we remove all speeches that are not in English.[6] This leaves $391,082$ speeches in English. Second, we exclude all speeches from the Speaker or Deputy Speaker of the National Assembly. Third, we exclude all speeches that are less than 25 words in length, since these short speeches do not convey any substantive content related to particularistic or national orientation.[7] This leaves a total of $237,125$ speeches for analysis across 530 unique MPs. Finally, we tokenize each of the $237,125$ speeches and remove words that occur in less than 100 of the $237,125$ speeches, leaving a total of $7,271$ unique words across $237,125$ speeches. Throughout the paper, we use the terms "speech", "document", and "paragraph" interchangably to refer to a single speech by an MP on a given day.

[2] The code repository for this project can be found at github.com/bnjmacdonald/kenya-hansardlytics. Documentation can be found at kenya-hansardlytics.readthedocs.io. The raw data will be made available soon.

[3] Transcripts of the Kenyan legislative debates can be viewed on Google Books here.

[4] The Kenyan parliament was unicameral for all years under consideration in this study. The Senate was abolished in 1965 shortly after independence and was reinstated in 2013 under the promulgation of a new constitution.

[5] Each session is 4 hours in length.

[6] We define a speech as English if more than 50% of words in the speech are English words. The Kenyan parliament has two official languages: English and Kiswahili. Most speeches are in English, but the Standing Orders state that if a member begins a speech in Kiswahili, she must continue in Kiswahili for the remainder of the speech. I would like to include these Kiswahili speeches in the analysis, but have not yet decided what approach to take.

[7] Examples of short speeches: "Excuse me, Mr. Speaker", "On a point of order", "Who was on the Floor?"

Name-matching. Next, in order to associate each speech with a unique MP, we matched each of the $237,125$ speeches to a single MP by extracting a master list of MP names from the Kenya Elections Database 2.0 and computing a similarity ratio between each unique speaker name from the $237,125$ speeches and every MP on the master list.[8] If no MP name in the master list exceeded a minimum threshold of 0.925 on the similarity ratio, no match was made.[9] Otherwise, the name with the highest similarity ratio was selected as a match. Overall, 77% of the $237,125$ speeches were matched to a unique speaker.[10]

We used the Ratcliff/Obershelp algorithm to compute the similarity ratio between two strings, defined as double the number of matching characters divided by the total number of characters in the two strings. This is equivalent to the total number of characters in the two strings minus the levenshtein distance, all divided by the total number of characters in the two strings. This similarity ratio ranges from 0 to 1. Finally, we concatenate speeches at the politician-day, politician-week, politician-month, politician-year, politician-parliament, and politician levels for use in some of the experiments described below.[11]

4 Approach

Doc2vec. As described above, the `doc2vec` model has two variations: `doc2vec-dm` and `doc2vec-dbow`. Both variations directly infer a fixed-length vector for each paragraph. We experiment with both model variations. We also experiment with alternative document identifiers when training the `doc2vec` models. In the original `doc2vec` implementation (Le and Mikolov, 2014), each paragraph is given its own identifier (and thus its own vector), and identifiers are not shared across paragraphs. In our experiments, we allow identifiers to be shared across documents. Specifically, in addition to estimating the standard `doc2vec` model in which each paragraph is given its own identifier, we estimate a slight variation in which we restrict the document identifiers to be shared within a politician. In other words, we assign the same document identifier to all speeches by politician, such that the same paragraph vector is shared across all speeches by that politician. The input for a single training example is still a single speech, *not* the concatenation of all speeches by the MP. We also experiment with shared document identifiers at alternative levels aggregation, such as the politician-parliament, politician-year, politician-month, and politician-day levels. We expect the `doc2vec` models with shared identifiers to produce more coherent politician embeddings, since the same document vector is shared across a politician's speeches.

Word2vec. We compare the performance of `doc2vec` against document vectors constructed from the sum of word vectors traing via `word2vec`. Specifically, we train word vectors on the corpus of Kenyan legislative speeches using both the skip-gram (`word2vec-sg`) and continuous bag of words (`word2vec-cbow`) implementations, producing a k-dimensional vector for each of the $7,271$ words in the vocabulary. At test time, we construct document vectors by summing the word vectors for all words in a document and then normalizing vector lengths to one.

LDA. Latent dirichlet allocation is growing rapidly in prominence in political science (see Lucas et al., 2015; Grimmer, 2013; Quinn et al., 2010), providing a simple and intuitive representation of text as a probability distribution over topics. Dai et al. (2015) show that `doc2vec` outperforms LDA and a bag of words baseline on several common NLP tasks – such as similarity and vector operations (see also Lau and Baldwin, 2016). However, it is unclear how LDA would perform against `doc2vec` and `word2vec` at the level of authors. LDA document vectors are inferred at test time from the trained mapping of words to topics.

Bag of words baseline. Finally, we compare all results to a simple bag of words baseline, where each document is represented as a vector of word counts.

[8] To create a time series of electoral performance for each candidate, we use the same name matching system described here to match MP names to speeches.

[9] In addition, a match could only be made between a speaker name and MP from the master list if the MP was an active member of parliament on the date that the speech was made.

[10] We conducted random checks for false positives and false negatives, but have not yet implemented a more rigorous assessment of the precision/recall of the name matching system. We plan to train a classifier in the coming months to improve performance of the name matching system.

[11] For instance, at the MP-month level, every speech by an MP in a particular month is concatenated together into a single mega-speech.

5 Experiments

Pre-estimation aggregation. We train all models with varying levels of pre-estimation aggregation. First, we estimate all models where an input document is defined as a single speech. We then train all models at higher levels of aggregation, where we concatenate speeches at the politician-day, politician-month, politician-year, politician-parliament, and politician level. At the highest level of aggregation (politician level), the corpus contains only 530 documents. This pre-estimation aggregation is distinct from the experiments with shared document identifiers described above. Specifically, under pre-estimation aggregation, we concatenate speeches *before* feeding them into the model, whereas in the `doc2vec` experiments described above we merely share document vectors across speeches without concatenating the speeches prior to estimation.

Hyperparameters. We train all models with embedding sizes of 50, 100, and 200. We also vary the length of training over a maximum of 1, 10, 50, 100, 200, 400, or 800 epochs.

Evaluation. We evaluate the performance of each model on three tasks: (1) intruder detection; (2) classification of nationally-oriented speeches; and (3) prediction of constituency spending on education projects. These tasks are described in greater detail below.

5.1 Results

Figure 2 displays a T-SNE visualization of the inferred document vectors from the `doc2vec-dm`, with colors corresponding to the political party of the speaker. The left panel displays 15,000 randomly sampled speeches inferred from the `doc2vec-dm` model in which we did not allow identifiers to be shared across speeches by the same politician. The right panel displays inferred vectors for the 530politicians, taken from the `doc2vec-dm` model in which we allowed document identifiers to be shared across all speeches by the same politician. In both cases, we see little evidence of separation between political parties, suggesting that the document vectors are failing to represent a crucial dimension of political conflict in Kenya.

Figure 2: Doc2Vec result

(a) No document identifier sharing

(b) Document identifiers shared

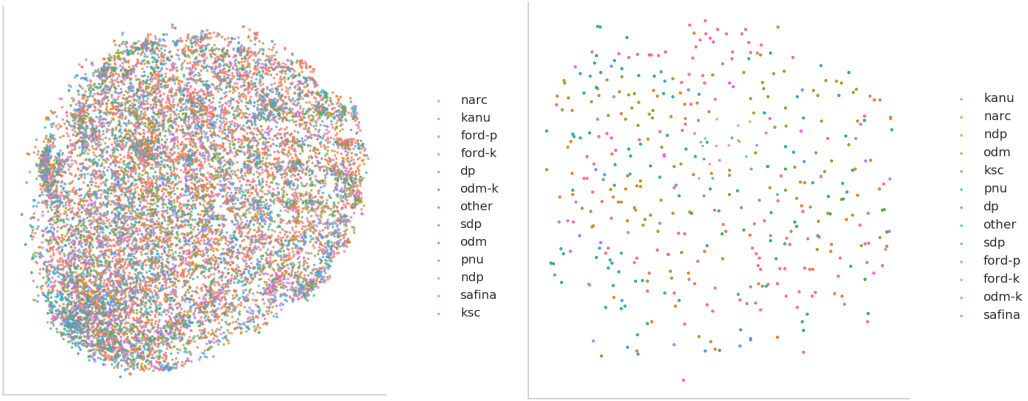

Fig. 2. T-SNE visualization of document vectors trained in the `doc2vec-dm` model. The left panel displays results when identifiers are not shared across documents (such that each point is a single speech). The right panel displays results when document identifiers are shared across all speeches by the same politician, such that a single point is a single politician. In both cases, we see little separation between political parties.

Intruder detection. Next, for each model, we randomly sample 500 groups of speeches, where each group contains four speeches by the same politician and one "intruder" speech (i.e. a speech from another politician). For each model, we use the inferred document vectors to compute the cosine similarity between all pairs of the five speeches, where the predicted intruder is the speech with the lowest average similarity amongst the five speeches.

Figure 3 displays the f1-score of each method on this task at varying epochs and embedding sizes. Here, we see that, in general, performance improves with the number of training epochs. On the other hand, there is no clear relationship between the embedding size and detection of intruders. LDA performs best overall, reaching an upper f1-score of approximately 0.4. Yet, since random guessing would result in an f1-score of approximately 0.2, these results are hardly impressive. Moreover, the bag of words baseline outperforms all methods except for LDA, raising concerns about information loss due to the naive application of word- and document-level embedding models.

Figure 3: Intruder detection

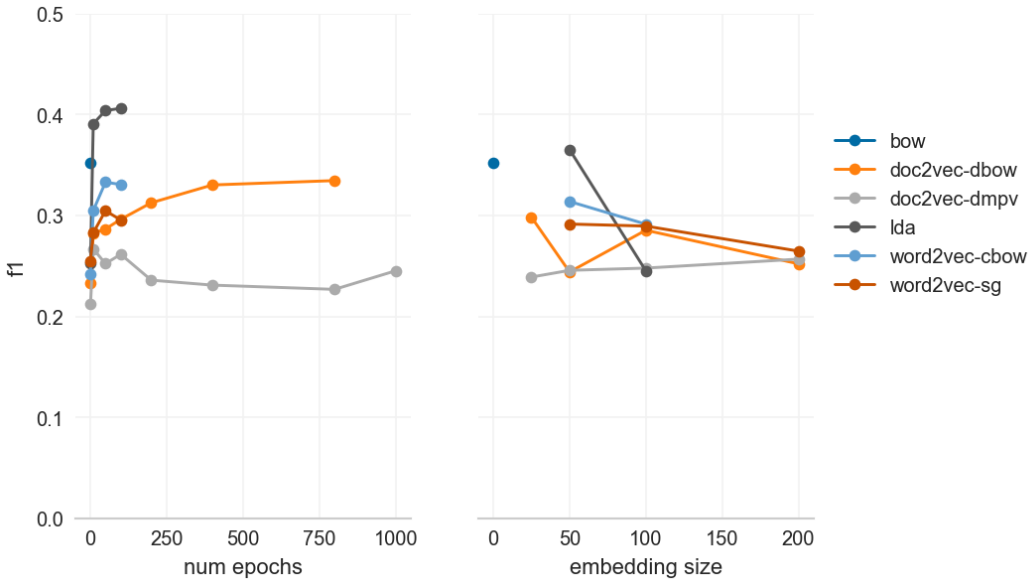

Fig. 3. This Figure displays the f1-score on the intruder detection task with four randomly sampled speeches from a politician and one "intruder" from another randomly sampled politician.

Classifying nationally-oriented speech. Next, we examine whether the document vectors capture variation in the degree to which politicians focus on *particularistic* concerns that primarily benefit their own constituents versus *national* concerns of broader importance to the country. To assess performance, we randomly sampled $1,000$ speeches and manually assigned each speech one of four labels: nationally-oriented, particularistic, procedural, or other. *nationally-oriented* speeches include speeches and debates which are national in scope. In contrast, *particularistic* speeches refer to issues and projects in specific localities. Examples of particularistic and nationally-oriented speeches are provided in the Appendix. *Procedural* speeches are in reference to legislative business and procedures rather than substantive issues. For instance, speeches in which a member raises a point of order or moves an order of business are procedural in nature. Finally, *other* speeches capture all other speeches that could not be easily categorized into the three preceding labels. We designed a simple `Django` application for the purposes of sampling and annotating these speeches.

Using the inferred document vectors from each method, we classify speeches as nationally-oriented, particularistic, procedural, or other using multinomial logistic regression with 5-fold cross-validation. Figure 4 displays the performance of each method on this task, at varying epochs and embedding sizes. Again, we see that the document embeddings fail to capture an important dimension of political speech in Kenya, with f1-scores hovering just below 0.3. In contrast, the bag of words baseline significantly outperforms the models, with an f1-score of approximately 0.53.

Classifying constituency spending. Finally, using a dataset of project-level spending in each parliamentary constituency between 2003 and 2010, we examine whether the inferred document vectors from a politician's speeches are predictive of the amount they spend on education projects in the following year. Specifically, we aggregate the total amount spent on education projects by each politician in the following year and then bin their spending into 10 equally sized categories. We then

Figure 4: Classification of nationally-oriented attention

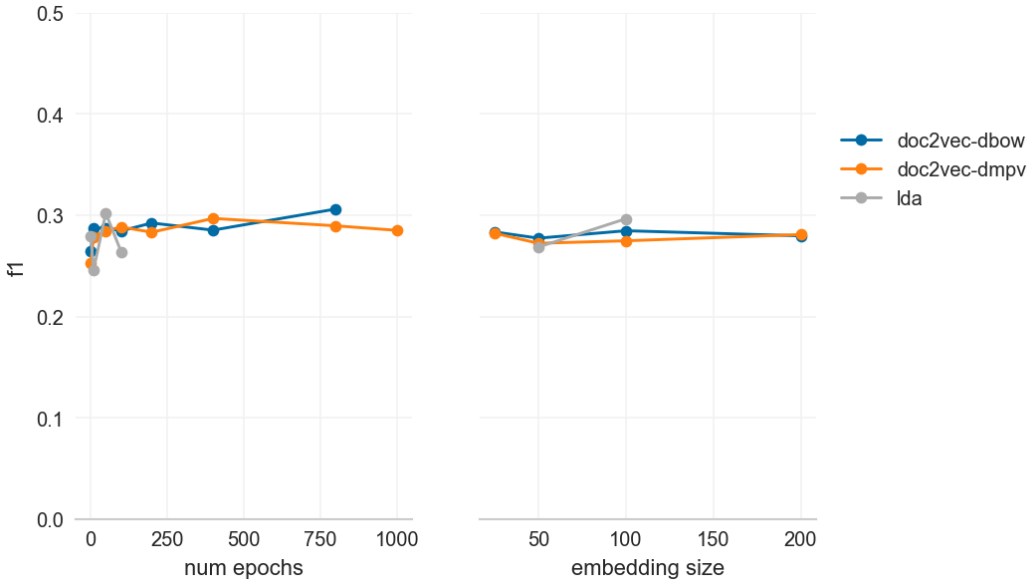

Fig. 4. This figure displays the f1-score from multinomial logistic regressions of nationally-oriented speech on the inferred document vectors. F1-scores are averaged over 5-fold cross-validation.

estimate a multinomial logistic regression of the spending bin on the politicians' inferred document vectors. The results of this exercise are shown in Figure 5. Consistent with the results presented so far, all of the methods perform poorly, hovering around an f1-score of 0.1. In contrast, bag of words achieves an f1-score of more than 0.2.

Figure 5: Classification of constituency spending on education project

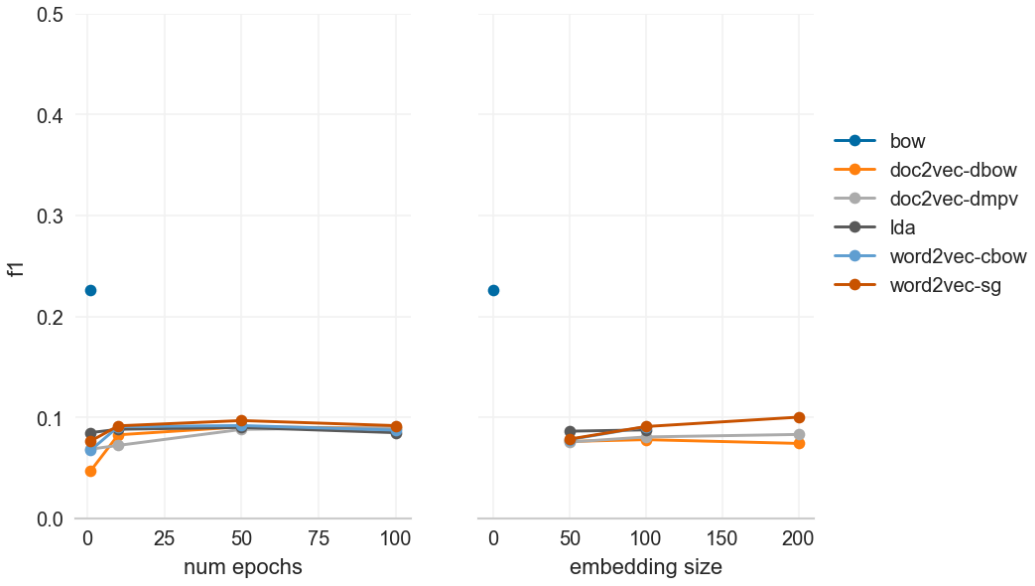

Fig. 5. This figure displays the f1-score from multinomial logistic regressions of education spending on the inferred document vectors. F1-scores are averaged over 5-fold cross-validation. Education spending is binned into 10 equally sized categories.

Finally, from qualitatively sampling speeches and their nearest neighbors in the inferred document vector space, it is clear that the `doc2vec`, `word2vec`, and LDA models are effectively clustering together documents with similar topics, such as documents debating similar policies or raising similar economic issues. As a result, it is *not* simply the case that the methods were not trained long enough or that they were incorrectly specified. Instead, our results suggest that it is not straightforward to apply these methods to capturing meaningful differences among politicians. In short, despite the ability of these methods to cluster together *similar speeches*, they ostensibly fail to produce meaningful dense representations of the *politicians making these speeches*.

6 Conclusion

Dense vector representations of words and sentences have proven themselves to be immensely useful in a wide range of natural language processing tasks. However, social scientists are often more interested in the *authors* of text, such as politicians, journalists, and thinktanks. In this study, we apply `word2vec`, `doc2vec`, and other methods to the task of generating dense vector representations of Kenyan politicians and their speeches, finding that these methods perform no better than a bag of words baseline.

There are several next steps we plan to take in this project. First, we plan to implement deeper neural models that can better account for the structured nature of legislative debates, such as convolutional neural networks and sequence-to-sequence models. By representing speech at multiple levels, these models may be better suited for capturing the differences between politicians in structured interactions such as legislative sessions. Second, we plan to apply these methods to at least two other datasets: (1) a dataset of all newspaper articles from African publishers since 1996, where the publisher and author of each article is known; and (2) transcripts of US Senate and Congressional debates.

References

Baroni, M., G. Dinu, and G. Kruszewski (2014). dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors.

Bonica, A. (2013). mapping the ideological marketplace.

Carroll, R., H. Rosenthal, K. T. Poole, J. Lo, and J. B. Lewis (2013). the structure of utility in spatial models of voting.

Dai, A. M., C. Olah, and Q. V. Le (2015). document embedding with paragraph vectors.

Ganesh, J., S. Ganguly, M. Gupta, V. Varma, and V. Pudi (2016). author2vec: learning author representations by combining content and link information.

Grimmer, J. (2013). *Representational Style in Congress: What Legislators Say and Why it Matters*. New York, NY: Cambridge University Press.

Kenter, T., A. Borisov, and M. D. Rijke (2016). siamese cbow: optimizing word embeddings for sentence representations.

Kiros, R., Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler (2015). skip-thought vectors.

Lau, J. H. and T. Baldwin (2016). an empirical evaluation of doc2vec with practical insights into document embedding generation.

Le, Q. and T. Mikolov (2014). distributed representations of sentences and documents.

Lucas, C., R. a. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 1–24.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), 209–228.

Rosenthal, H. and K. T. Poole (1997). *congress: a political-economic history of roll call voting*. Oxford University Press.

Appendices

Examples of speeches

Particularistic speeches:

> "Mr. Speaker, Sir, I request for a Ministerial Statement from the Minister for Environment and Natural Resources. I would like him to inform the House what his Ministry has done about the hyacinth which is choking Kisumu and its environment. What does he intend to do to clear the hyacinth within a very short time."

> "Mr. Speaker, Sir, I beg to ask the Minister for Co-operative Development and Marketing the following Question by Private Notice. (a) Is the Minister aware that Ol Kalou Farmers Sacco Society has suffered serious cash-flow problems primarily due to misappropriation of members' funds? (b) Is he further aware that despite the mater having been taken up by the Anti-Corruption Police Unit nine months ago, no tangible action has been taken to recover the funds or to bring the culprits to book? (c) What measures has the Minister put in place to avoid total collapse of the society?"

Nationally-oriented speeches:

> "Mr. Speaker, Sir, since the funds are not a lot, they have been monitored through the offices of District Development Officers, the International Fund for Agricultural Development (IFAD) programmes and the Millennium Development Goals Commission. I can say before this House that the funds allocated and distributed in the manner they have been distributed cannot have a big impact on poverty eradication. The onus is on this House to ensure that more funds are allocated to the Poverty Eradication Commission, because, as things are today, if we allocate Kshs300 million to that Commission to be shared out to the original 71 districts, we are talking of barely Kshs4 million per district. We all agree that at this time and era, Kshs4 million for projects and poverty eradication is a meagre amount. So, I want to appeal to the House that we need to think about this Commission, and see how we can allocate more money to it."

> "Sometime back! He is sitting here but I do not want to mention his name. That is the truth. Mr. Temporary Deputy Speaker, Sir, our lecturers' remuneration is unbelievable compared to what their equals in other countries get. Even within the country, a lecturer or a professor who leaves the University of Nairobi and goes to USIU gets much more money. What is in USIU that we do not have at the University of Nairobi or Kenyatta University?"

Procedural speeches:

> "On a point of order, Mr. Temporary Deputy Speaker, Sir. Is it in order for the "Attorney-General" to quote the Bible like Satan did, by misdirecting this House?"

> "On a point of order, Mr. Temporary Deputy Speaker, Sir."