algorithms

MDPI

Article

Topic Scaling: A Joint Document Scaling–Topic Model Approach to Learn Time-Specific Topics

Sami Diaf * and Ulrich Fritsche *

Universität Hamburg, Faculty of Business, Economics and Social Sciences, Department Socioeconomics, Welckerstr. 8, 20354 Hamburg, Germany

* Correspondence: sami.diaf@uni-hamburg.de (S.D.); ulrich.fritsche@uni-hamburg.de (U.F.)

Abstract: This paper proposes a new methodology to study sequential corpora by implementing a two-stage algorithm that learns time-based topics with respect to a scale of document positions and introduces the concept of *Topic Scaling*, which ranks learned topics within the same document scale. The first stage ranks documents using *Wordfish*, a Poisson-based document-scaling method, to estimate document positions that serve, in the second stage, as a dependent variable to learn relevant topics via a supervised Latent Dirichlet Allocation. This novelty brings two innovations in text mining as it explains document positions, whose scale is a latent variable, and ranks the inferred topics on the document scale to match their occurrences within the corpus and track their evolution. Tested on the U.S. State Of The Union two-party addresses, this inductive approach reveals that each party dominates one end of the learned scale with interchangeable transitions that follow the parties' term of office, while it shows for the corpus of German economic forecasting reports a shift in the narrative style adopted by economic institutions following the 2008 financial crisis. Besides a demonstrated high accuracy in predicting in-sample document positions from topic scores, this method unfolds further hidden topics that differentiate similar documents by increasing the number of learned topics to expand potential nested hierarchical topic structures. Compared to other popular topic models, *Topic Scaling* learns topics with respect to document similarities without specifying a time frequency to learn topic evolution, thus capturing broader topic patterns than dynamic topic models and yielding more interpretable outputs than a plain Latent Dirichlet Allocation.

Keywords: document scaling; topic models; supervised learning

Citation: Diaf, S.; Fritsche, U. Topic Scaling: A Joint Document Scaling–Topic Model Approach to Learn Time-Specific Topics. *Algorithms* **2022**, *15*, 430. https://doi.org/10.3390/a15110430

Academic Editor: Frank Werner

Received: 13 September 2022
Accepted: 11 November 2022
Published: 16 November 2022

1. Introduction

The digitization wave triggered an unprecedented interest in the use of textual sources, either at the academic or at the professional level, with the goal of extracting more hidden information and latent patterns from the available corpora. Such tasks could not be performed without the use of machine-learning techniques that offer different approaches to uncover hidden information from text, depending on the corpus characteristics and the model conceptualization. Today's *text-as-data* trend has prompted a growing interest by social scientists who are often confronted to large and complex collections of documents that require sophisticated tools to obtain convincing insights as outputs. Algorithms used by this category of studies could be ranked into two distinct but complementary families of methods, whose application depends on the hypothesis to test and the nature of the corpus.

1.1. Document Scaling

Document scaling refers to a popular class of methods (mostly unsupervised models) used to study political manifestos and other corpora in the social sciences, resulting in a low-dimensional metric or scale used to compare documents based on a set of assumptions regarding word distributions. Earlier approaches used classic machine-learning algorithms, such as Naïve Bayes, to build scales [1]. Laver et al. [2] used *Wordscores* to estimate political

party positions based on pre-established reference scores, or wordlists, for texts. Slapin and Proksch [3] designed *Wordfish*, a parametric approach that uses a Poisson distribution model to infer a unidimensional document scale from the distribution of word frequencies, considered to be a proxy for (latent) ideological positions. *Wordshoal* [4] is a two-stage document-scaling method that applies *Wordfish* on each debate in the corpus and then uses a Bayesian factor aggregation to uncover further dimensions related to the corpus metadata.

Scaling techniques have been widely used in political science [5], particularly to study party manifestos [4] but also as a dimensionality reduction technique. They come with the constraint of recovering just one of many possible hidden dimensions [6], in addition to being unable to properly define its meaningfulness [5], suffering from possible word variations when handling corpora spanning large periods [6] and are sensitive to chosen pre-processing steps when cleaning texts [7]. Goet [6] suggested the use of supervised approaches to have meaningful polarization results.

1.2. Topic Models

Unsupervised topic models have been the usual choice for researchers working with text data, with the aim of unveiling latent features or hidden structures that explain the word-document occurrences. Latent Dirichlet Allocation (LDA) [8] is the usual go-to algorithm for such exercises. This generative model assumes documents are mixtures of independent latent features called *topics*, which are in turn mixtures of words drawn from a Dirichlet distribution. Several variants of topic models were later proposed to deal with specific cases [9] to consider sequences, hierarchies and sentiments when learning topics. Blei and Lafferty [10] used a Dynamic Topic Model (DTM) to study sequential corpora through a discrete-time variant of plain vanilla LDA whose architecture infers time-related topics, based on priors drawn from a Markov process, suitable for large corpora spanning over a long timeframe.

McAuliffe and Blei [11] proposed a supervised Latent Dirichlet Allocation (sLDA) model that builds a generalized linear model (GLM) on top of a classic LDA to help infer topics when documents are paired with labels. Boyd-Graber and Resnik [12] built a multilingual model based on sLDA to capture how multilingual concepts are clustered into thematically coherent topics and how topics associated with text connect to an observed regression variable. Several other tools were designed on top of plain LDA, such as hierarchical topic models [13], Pachinko allocation models [14] and sentence labeling [15] with the aim to uncover further latent structures in corpora based on hierarchies and specific structures. The number of topics to be learned, usually a hyperparameter, is set arbitrarily by users, although a variety of methods were proposed to estimate it such as hierarchical Dirichlet process (HDP) [16] which uses variational inference to uncover the number of topics in the collection of documents. Greene and Cross [17] used a two-layer non-negative matrix factorization model to explore topic dynamics in the corpus of the European Parliament and found substantive niche topics not captured by standard dynamic topic models.

Another class of topic models, the Structural Topic Model, has been used to link exogenous variables with learned topics [18,19] in order to investigate the impact of potential covariates on learned topics, by using document metadata in the estimation step, to facilitate tracking of the effect that variables could have on the learned topics in an ex-ante evaluation.

Our approach builds a two-stage learning process: a document scaling (*Wordfish*) on top of a supervised Latent Dirichlet Allocation, to allow the uncovering of time-dependent topics based on the position of each document in the corpus. Therefore, even if the Dirichlet distribution is not sequential [10], topics could be learned with respect to the evolving distribution of word frequencies that served to estimate the latent scale, which refers to document scores used as a time-based, dependent variable. We used *Wordfish* for document scaling as it builds a unique scale for all documents, rather than Wordshoal, which learns a

distinct scale for each debate in the corpus [4], endowing the learned scale with the ability to uncover potential shifts in word distributions and exploiting it at the topic-level.

To the best of our knowledge, there was no attempt at text mining to use scaling techniques beyond our estimated document positions or explore further extensions using topic models. Our method is noticeably suited to studying the dynamic structure of the corpus by uncovering potentially time-dependent, nested topics with the use of a one-dimensional measurement model (*Wordfish*) for text data, rather than evolving topics that need further hyperparameter tuning such as setting the frequency (time stamp) for the analysis [10]. The optimal number of topics in our method, still a hyperparameter, could be learned by maximizing a metric of choice, such as the root mean squared error (RMSE) between the estimated and predicted document positions or the log-likelihood of the estimated sLDA model. We notice that increasing the number of topics, despite slowing down the execution time for large corpora and not necessarily improving the RMSE accuracy when using regularization, helps the uncovering of the breadth of hidden topic structures similar to Pachinko Allocation Model [14] that are highly informative.

Our findings contribute to two distinctive fields of text mining. In document scaling, we explain document positions based on groups of words occurring together (topics) rather than individual words. In topic modelling, we learn these topics from a collection of documents that are, a priori, time-scaled regarding their word frequencies, so that two documents with similar scores will tend to have similar topic distributions. Lastly, the use of regularization allows the increase to the number of topics without altering the predictive properties of the model, and helps the unfolding of potential hierarchical structures in the learned topics.

In parallel to document scaling, this paper introduces the concept of *Topic Scaling*, which refers to a supervised method used to learn topics with respect to a labelled scale. This approach associates topic scores with document positions and ranks topics with their most occurrences in the corpus to help tracking their distributions over time.

The remainder of the paper describes the two components upon which *Topic Scaling* is built (Section 2), then details the application results over the State Of The Union two-party corpus (Section 3) and later compares the results with a plain LDA model, as well as a dynamic topic model. The corpus of German economic forecasting reports [20] is used to learn coherent topics from a collection of documents that targeted monetary policy practices in Germany and the European Union throughout multiple developments from 1999 to 2017.

1.3. State Of The Union Addresses

State Of The Union (SOTU) addresses bear an importance in politics [21] as the US Constitution (Article II, Section 3) requires the president to provide information to the Congress about measures as he shall judge necessary and expedient, i.e., a description of the current situation, priorities, and the legislative intents [22]. Petrocik et al. [23] assumed that some party-related relationships between a president's party affiliation and topics could be seen in SOTU, as Democratic tenures are more likely to deal with topics such as education, family, welfare, and healthcare, while a Republican presidency is frequently tied to free enterprise and business, reduction of expenses, or support for the military sector. An alternative hypothesis would consider each president to have his own priorities, independently from its predecessors, leading to a distinct vocabulary choice in his addresses [22].

As a popular dataset in linguistics and text-mining applications, the SOTU has been studied by researchers to investigate rhetoric or uncover distinctive patterns, such as text clustering for presidential style [22], vocabulary-growth model [24], topic models of important words [25] or syntactic complexity [26]. Teten [27] studied the rhetorical changes to the SOTU addresses from George Washington to Bill Clinton and found three distinctive periods—a founding, a traditional and a modern period—and Cummins [28] showed the importance of rhetorical attention to economic policy and foreign relations in modern addresses (1953–2000).

2. Method

This paper's algorithm, named *Topic Scaling*, is a two-stage learning process for time-based topics from document positions, which serve as labels for supervised topic models (see Algorithm 1). Hence, a sequential scale is built with the *Wordfish* model, and used as a label to run a supervised Latent Dirichlet Allocation that renders topic allocation over time for historical corpora. This strategy combines two different machine-learning approaches to give an augmented, automated content analysis method where documents and topics are put on a unique scale.

Algorithm 1 Topic Scaling

1. Estimate document positions $\hat{\psi}$ using *Wordfish*
   (a) Assuming: $w_{ik} = \text{Poisson}(\lambda_{ik})$
   (b) Learn via Expectation Maximization $(\alpha, \nu, \beta, \psi)$ from: $log(\lambda_{ik}) = \alpha_i + \nu_k + \beta_k \times \psi_i$
2. Learn a Supervised LDA with an L2 regularization (shrinkage parameter λ)
   (a) Draw topic proportions $\theta|\alpha \sim Dir(\alpha)$
   (b) For each word:
      i. Draw a topic assignment: $z_n|\theta \sim Mult(\theta)$
      ii. Draw word $w_n|z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$
   (c) Draw document scale $\hat{\psi}|z_{1:N}, \eta, \sigma^2 \sim \mathcal{N}(\eta^\intercal \bar{z}, \sigma^2)$ with $z \sim \mathcal{N}(0, \frac{1}{\lambda})$

2.1. Measurement Model

The measurement model of *Topic Scaling* consists of a parametric estimation of document positions that infers a scale to cluster documents based on word frequency similarities over time. *Wordfish* [3] assumes word frequencies w_{ik} are drawn from a Poisson distribution, to estimate word effects β_k and a latent scale ψ_i as proxy of document positions.

$$w_{ik} = \text{Poisson}(\lambda_{ik}) \tag{1}$$

$$log(\lambda_{ik}) = \alpha_i + \nu_k + \beta_k \times \psi_i \tag{2}$$

where: w_{ik} is the count of the word k in document i, λ_{ik} is the parameter of the Poisson distribution denoting its mean and variance. Parameters to be learned are: α_i as a document-specific fixed effect, ν_k a word-specific fixed effect, β_k is the relationship of word k to the latent document position and ψ_i is the latent position of document i or the measurement scale. The model is estimated via the Expectation Maximization algorithm, consisting of estimating word parameters (ν, β) and document parameters (α, ψ) alternatively, until reaching a convergence [3]. Variational inference (Monte Carlo–Markov Chains) with Bayesian priors could be used as well to estimate *Wordfish* parameters. We notice that the main result of *Wordfish*(ψ) has an undetermined scale (direction) and needs to be restricted to identify the *Wordfish* equation given in (2) [3].

This parametric measurement allows a time series scale to be estimated solely based on the observed word frequencies. Although the model is prone to potential departures from the Poisson hypothesis of conditional independence (expected value being equal to the variance) [29], it remains a robust method for outliers in word usage [4], compared to Correspondence Analysis (CA) [6].

2.2. Supervised LDA

The Latent Dirichlet Allocation [8] could be seen as a Bayesian mixed-membership, unsupervised learning algorithm that learns independent structures, or groups of words, called topics, from a collection of documents. Later, McAuliffe and Blei [11] proposed a supervised extension where the triplet documents–topics–words is kept under a generalized

linear model (GLM) that accommodates a variety of response types, as mentioned in Figure 1, where for a dataset of observed document–response pairs $(w_{d,1:N}, y_d)$, the goal is to learn topic multinomials β_k, the Dirichlet parameter α of the topic proportions θ_d and the GLM parameters η and σ^2, using likelihood estimation based on variational Expectation Maximization [11].

A supervised LDA [11] is used here as a second-stage method to learn topics from the corpus, where the dependent variable is the learned scale from the measurement model. An L2 regularization scheme on learned topic scores is used to allow overlapping topics and prevent hard clustering situations.

For the sake of illustration, we describe our approach as a two-stage model, but it could also be interpreted as one-stage method as the Poisson model for document scaling is indeed a special case of the generalized linear model, used in the sLDA estimation [11].

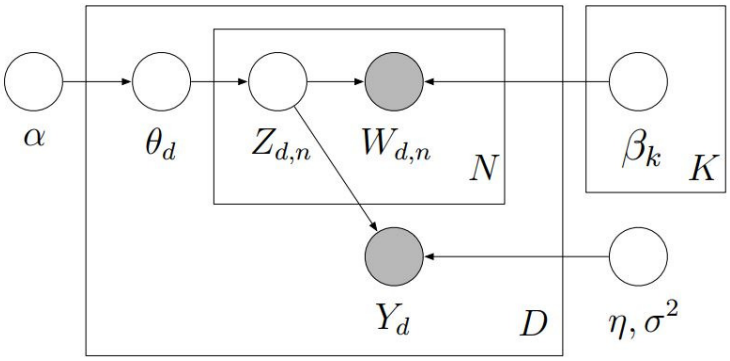

Figure 1. Plate diagram of Supervised Latent Dirichlet allocation [11].

The number of topics to be learned could be approached via maximizing metrics related to sLDA such as R^2 or log-likelihood. Koltcov et al. [30] proposed, based on statistical physics, a combination of the renormalization procedure with Rényi entropy for an efficient and quick estimation of the optimal number of topics in a given corpus. For T topics and N words in the corpus, we set the deformation parameter $q = \frac{1}{T}$ and select N words with a high probability ($\phi_{wt} > \frac{1}{W}$) in the topic-word matrix ϕ to compute the density-of-states function $\rho = \frac{N}{WT}$. The energy can be expressed as:

$$E = -ln(\tilde{P}) = -ln\left(\frac{1}{T}\sum_{w,t}\phi_{wt}1_{\phi_{wt}>\frac{1}{W}}\right) \tag{3}$$

with a partition function $Z_q = e^{-qE+S} = \rho(\tilde{P})^q$. The Rényi entropy is defined as

$$S_q^R = \frac{ln(Z_q)}{q-1} = \frac{q \times ln(q\tilde{P}) + q^{-1}ln(\tilde{\rho})}{q-1} \tag{4}$$

and its minimum corresponds to the optimal number of topics [30].

3. Data and Results

3.1. State of the Union Corpus

We use the corpus of the State Of The Union (SOTU) speeches available in the R package *quanteda* [31], gathering 214 speeches of U.S. presidents from 1790 to 2019. We keep, in our analysis, documents starting from 1853 to ensure a party duality of Democratic–Republican in our corpus that later helps us study document-scaling variations, while words, previously lemmatized with the *spaCy* language model [32], with fewer than 3 occurrences in the selected corpus were excluded to reduce the size of the document term matrix for more efficiency.

At first, document-scaling (*Wordfish*) was run on 176 documents, setting the score of Reagan's address in 1981 as being greater than Carter's of the same year, to identify the scale parameter ψ. Figures 2 and 3 show *Wordfish* results with a clear time-effect for both parties and noticeable similarities in addresses over time (in blue the *Locally Weighted Scatterplot Smoothing loess* curve [33]). Democratic addresses underwent a significant change in wording in the early twentieth century with the exception of Wilson's speech of 1916 (Figure 2), but the modern addresses of Clinton and Obama share close document positions, and hence similarities, in line with the tree-based topical word-term representation of Savoy [22]. On the other hand, Republican addresses show greater variability in terms of document positions (Figure 3) where it is common to have scattered positions of a single president, indicating a potential shift of interest in the addresses.

Figure 4 shows the densities of document positions, by party, to be bimodal. Each party dominates one end of a scale, with interchangeability linked to the presidents' tenure in office, as Democratic addresses have a skewed distribution to the left, while Republican ones have a less skewed one to the right. The document position $\hat{\psi} \approx 0.25$, corresponding to the three addresses of president T.W. Wilson in 1914, could be interpreted as a cutoff or turning point that separates the studied corpus into two dual periods related to the evolution of rhetorical presidency [27]. *Wordfish* scores confirm the results of Savoy [22], who identified a distinctive style for each president since the speech of Roosevelt in 1934, while previous presidents shared many stylistic aspects.

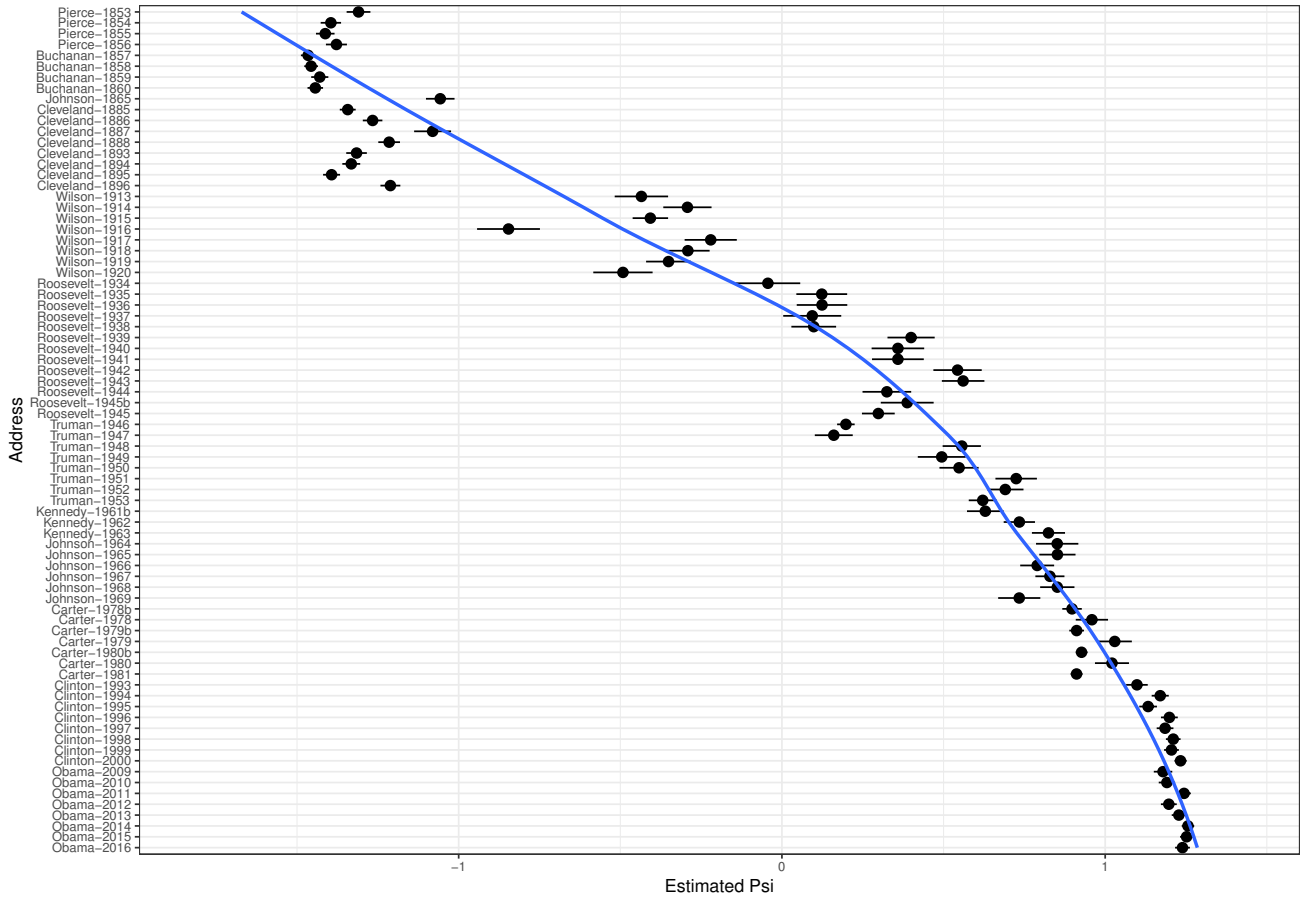

Figure 2. *Wordfish* scores for Democratic presidents (blue line is the smoothed Loess curve).

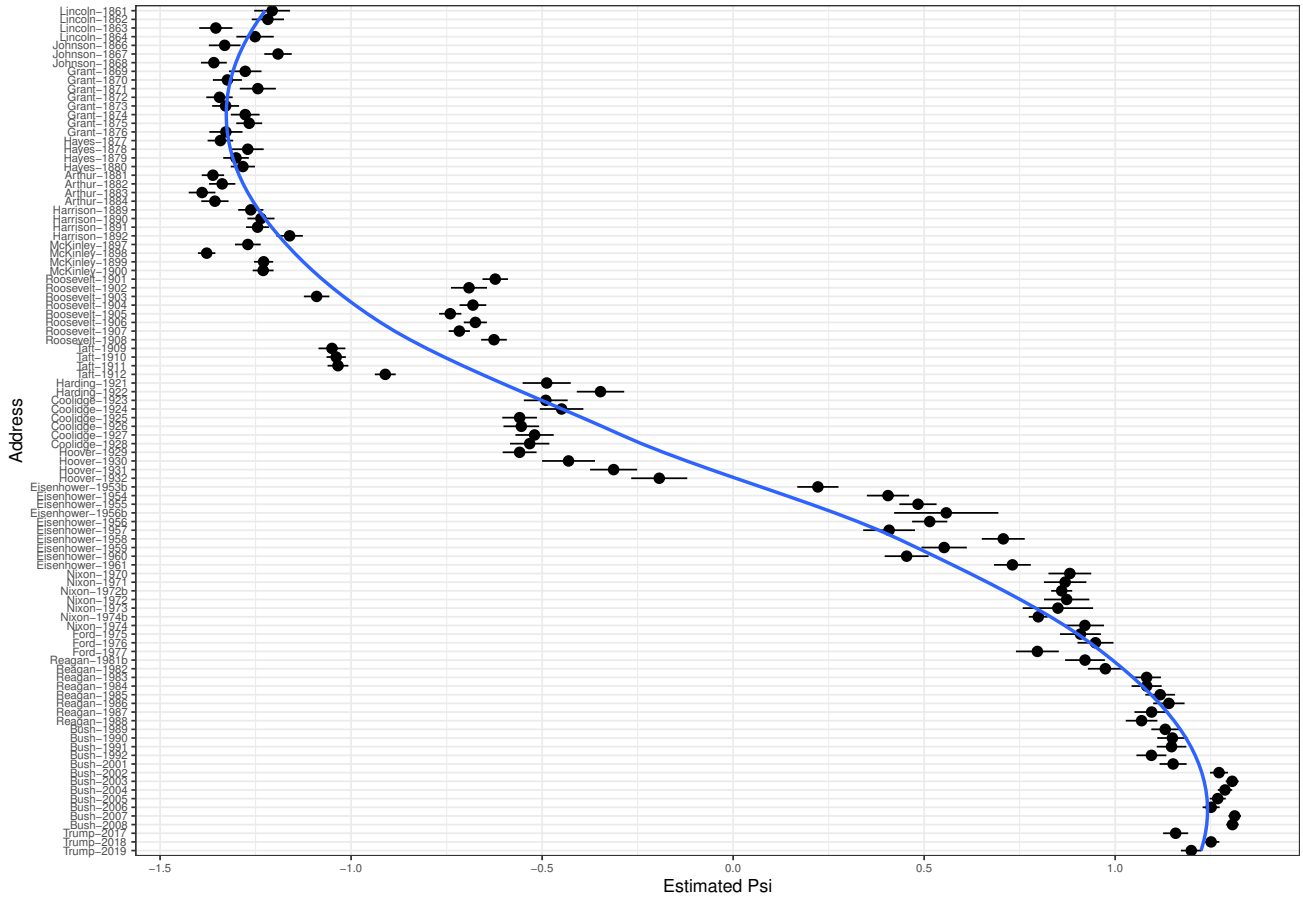

Figure 3. *Wordfish* scores for Republican presidents (blue line is the smoothed Loess curve).

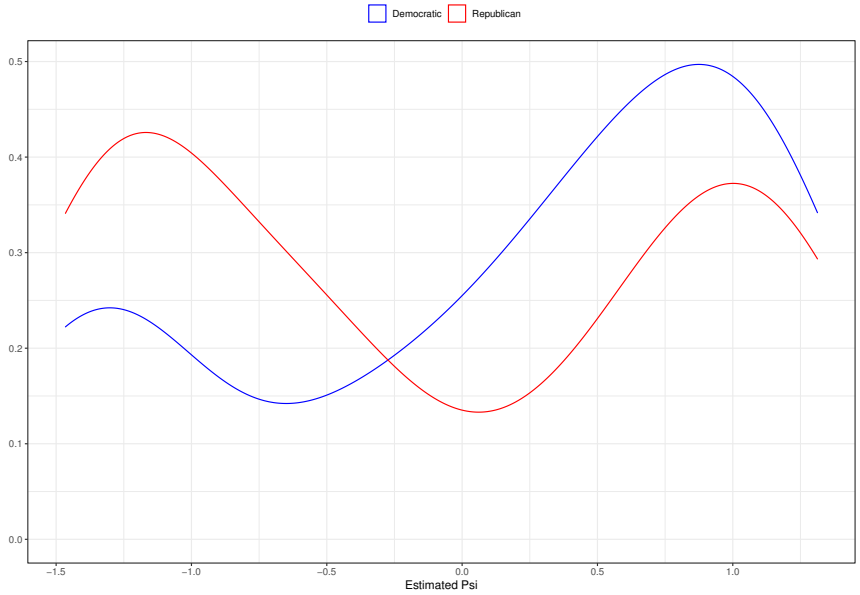

Figure 4. Density plots of *Wordfish* scores by party.

Word positions plot (Figure 5) shows the contribution of specific words to the estimated *Wordfish* scale (words with higher $|\beta_k|$) and their specific effects (ν_k). It appears that words related to security, mostly found in modern speeches, contribute mostly to documents with positive scores, i.e., recent addresses, even if their specific effects are relatively low.

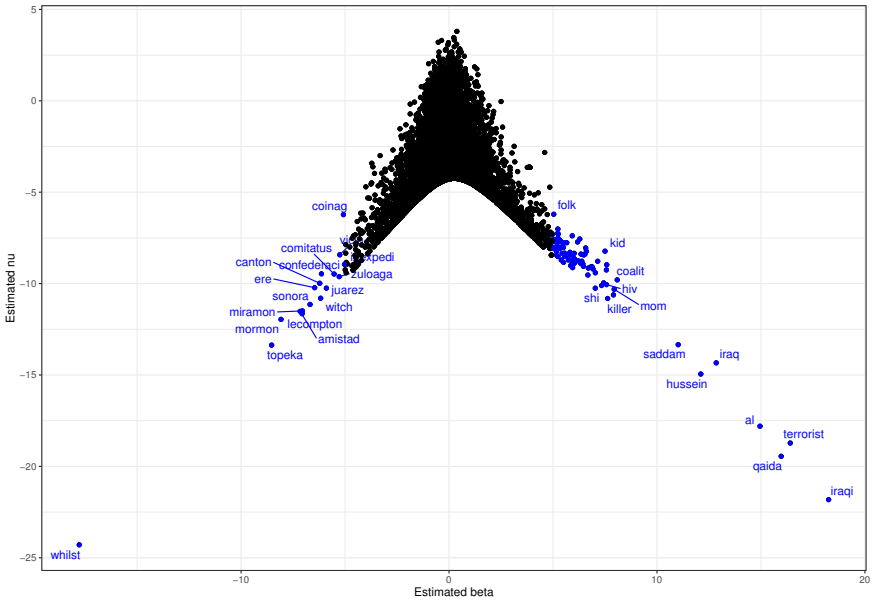

Figure 5. Estimated word positions from *Wordfish*.

In the second stage, supervised LDA was run through an Expectation Maximization algorithm (50 steps expectation and 20 steps maximization), with $\alpha = 1$ (Dirichlet hyperparameter for topic proportions), $\eta = 0.1$ (Dirichlet hyperparameter for topic multinomials) and an L2 regularization scheme to learn topic scores with a shrinkage parameter $\lambda = 0.01$. This setting was tested against a plain LDA with a dynamic topic model by keeping the same hyperparameter setting used for sLDA ($\alpha = 1$ and $\eta = 0.1$).

Figures 6 and 7 show the top 10 words in each learned *Topic-Scaling* model with 10 and 15 topics, respectively, and confirms the existence of nested structures when increasing the number of topics. A model with 15 topics seems to be the winning solution as it maximizes the Rényi entropy (Table 1), while other metrics (R^2 and log-likelihood) were not found to be informative with increasing number of topics.

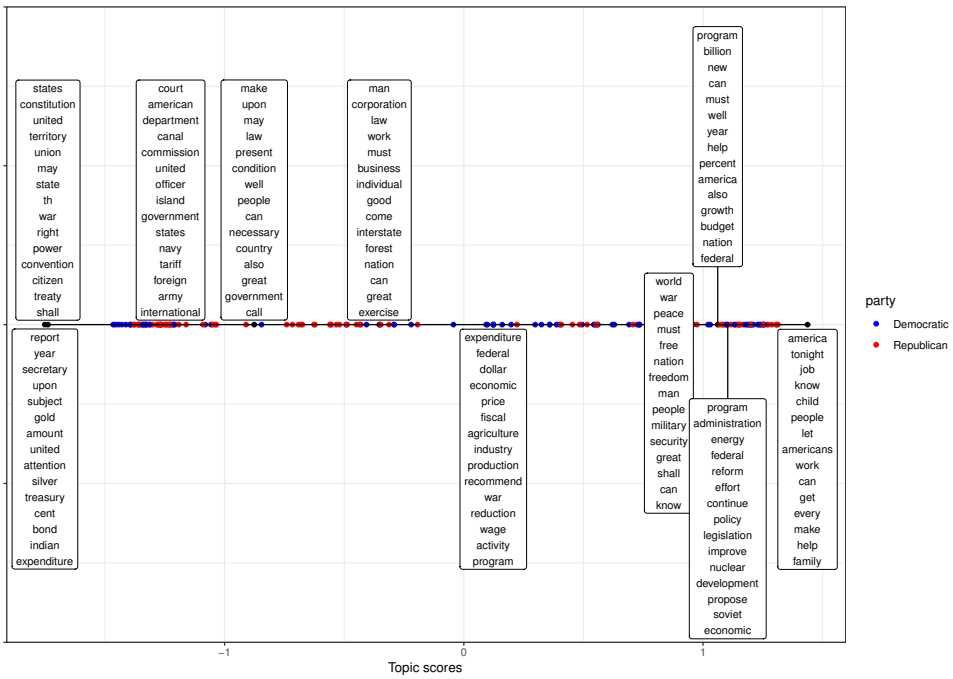

Figure 6. Topic scores learned via *Topic Scaling* (10 topics).

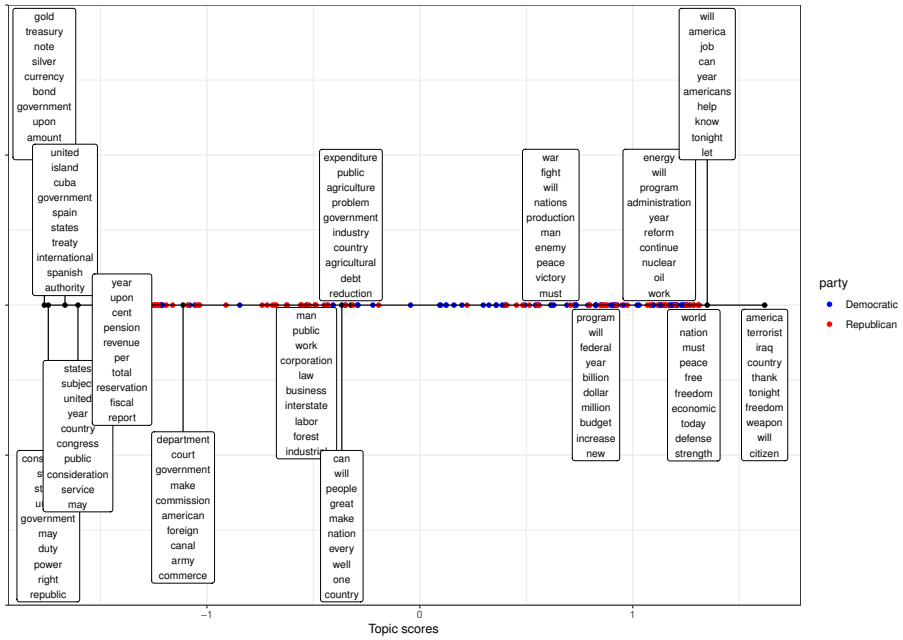

Figure 7. Topic scores learned via *Topic Scaling* (15 topics).

Documents with negative scores are likely to belong to the second half of the nineteenth century and the first three decades of the twentieth century, related to both domestic and international environments with an emphasis on government affairs (regulation and administration), while topics with positive scores are linked to modern addresses, focusing on economic and security issues. These two distinct windows differentiate topics and are separated by addresses given during the 1930s, in line with Figure 4, which indicates two intertwined regimes in document scores.

Recent addresses favour economic welfare, government affairs, security and international environment, corresponding to the four topics on the left of the scale (Figure 6) which seem to be party-related topics [22]. Increasing the number of topics to 15 (Figure 7) reveals further subtopics that provide a better area-specific understanding such as terrorism, foreign policy, internal affairs and labour. A few topics will cluster the topic content so that topic interpretation cannot be easily given, and increasing the topics unveils clearer policy fields of the addresses that help distinguish similarities in the addresses within a moderate timeframe.

Results of our method are displayed in Table 2. As a matter of comparison, results of a plain LDA with 15 topics (Table 3) and a dynamic topic model with three per-decade topics (Table 4) do not yield coherent topic structures, compared to our method. A clear dominance of frequent words used in the rhetoric (such as *government*, *will*, *administration*, *Congress*) is seen in almost all topics, reducing the model's informative content.

The *scale effect*, referring to the use of *Wordfish* scores, helps distinguish topics by periods where documents present similarities in the distribution of word frequencies. One could interpret the learned topics as being the most dominant probability distribution of words over a specific time frame, where addresses exhibit similar word usage.

Finally, *Topic Scaling* displays interesting findings in terms of topic contribution. Two topics (Topics 5 and 6) form the building blocks of modern addresses and are seen as dominant topics in recent speeches (Figures 8 and 9), dealing with family, economic condition and foreign affairs.

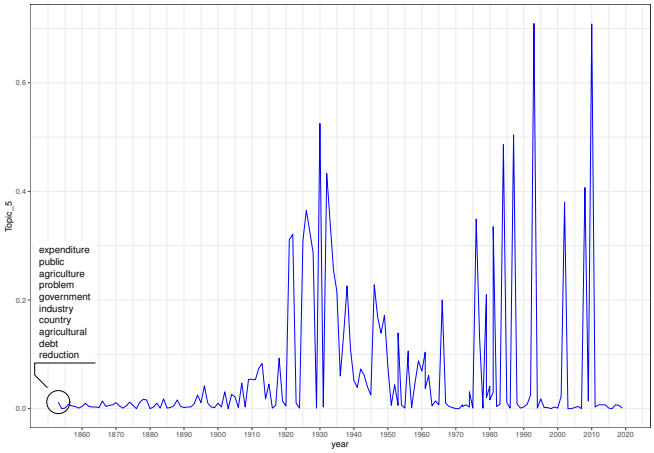

Figure 8. Evolution of *Topic 5* proportions in SOTU speeches (*Topic Scaling*).

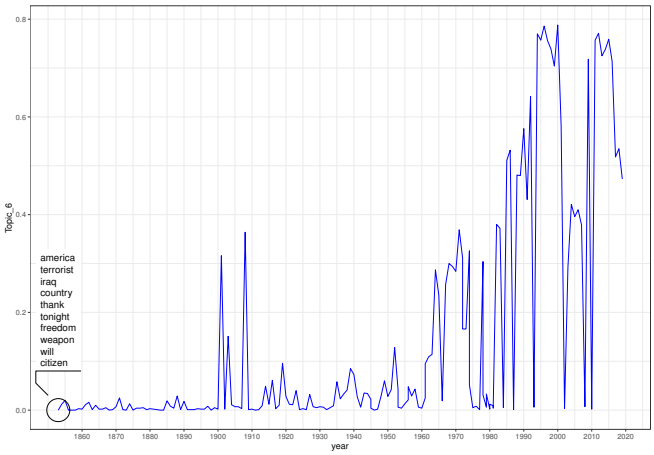

Figure 9. Evolution of *Topic 6* proportions in SOTU speeches (*Topic Scaling*).

Table 1. Results of *Topic Scaling* metrics, per different number of topics.

Number of Topics	R^2	Log-Likelihood	Rényi Entropy	Number of Topics	R^2	Log-Likelihood	Rényi Entropy
4	0.9963	236.5854	−3.7986	15	0.9976	287.9229	−3.3353
5	0.9946	207.1414	−3.6783	16	0.9975	276.7568	−3.4440
6	0.9964	246.9864	−3.6557	17	0.9969	261.3767	−3.3559
7	0.9962	240.2083	−3.6772	18	0.9978	281.0514	−3.3979
8	0.9957	228.3746	−3.5676	19	0.9980	293.1175	−3.4628
9	0.9963	235.5375	−3.5626	20	0.9981	298.9496	−3.5025
10	0.9963	245.5822	−3.3909	21	0.9978	295.9086	−3.5367
11	0.9977	273.5398	−3.4583	22	0.9976	272.1873	−3.5247
12	0.9968	255.8592	−3.4838	23	0.9980	285.0266	−3.6266
13	0.9976	279.4950	−3.3901	24	0.9981	301.2945	−3.5919
14	0.9967	245.8567	−3.4175	25	0.9977	293.4041	−3.6431

Table 2. Top 10 words from the estimated *Topic Scaling* with 15 topics.

Topic	Top 10 Words
1	energy, will, program, administration, year, reform, continue, nuclear, oil, work
2	united, island, cuba ,government, spain, states, treaty, international, spanish, authority
3	program, will, federal, year, billion, dollar, million, budget, increase, new
4	america, terrorist, iraq, country, thank, tonight, freedom, weapon, will, citizen
5	expenditure, public, agriculture, problem, government, industry, country, agricultural, debt, reduction
6	will, america, job, can, year, americans, help, know, tonight, let
7	man, public, work, corporation, law, business, interstate, labor, forest, industrial
8	department, court, government, make, commission, american, foreign, canal, army, commerce
9	war, fight, will, nations, production, man, enemy, peace, victory, must
10	states, subject, united, year, country, congress, public, consideration, service, may
11	constitution, state, states, union, government, may, duty, power, right, republic
12	can, will, people, great, make, nation, every, well, one, country
13	world, nation, must, peace, free, freedom, economic, today, defense, strength
14	year, upon, cent, pension, revenue, per, total, reservation, fiscal, report
15	gold, treasury, note, silver, currency, bond, government, upon, amount, duty

Table 3. Top 10 words from an estimated plain LDA with 15 topics.

Topic	Top 10 Words
1	will, america, must, world, nation, can, year, people, help, freedom
2	will, war, world, nation, can, must, people, great, man, peace
3	will, program, year, must, government, can, nation, congress, federal, new
4	government, will, year, make, war, congress, country, can, federal, public
5	man, law, will, government, make, can, great, nation, people, work
6	will, year, program, congress, federal, administration, new, increase, continue, energy
7	states, united, government, congress, year, may, will, country, upon, make
8	will, upon, make, year, law, people, government, country, public, great
9	government, states, united, year, will, make, upon, congress, american, may
10	states, government, united, state, may, congress, will, power, constitution, upon
11	will, year, must, work, people, can, child, america, new, make
12	will, year, can, america, people, new, american, great, congress, nation
13	government, make, will, states, united, congress, department, american, year, law
14	government, upon, condition, may, present, year, law, make, gold, time
15	will, year, can, job, make, work, america, people, new, american

Table 4. Top 10 words from an estimated Dynamic Topic Model (per decade) with 3 topics.

Decade		1853–1859			1860–1869			1870–1879			1880–1889			1890–1899	
Topics	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
	states	constitution	people	states	constitution	people	states	government	people	states	government	people	government	government	people
	government	would	great	government	would	great	government	would	great	government	would	great	states	would	great
	united	government	nation	united	government	nation	united	people	nation	united	people	nation	united	people	nation
	congress	states	would	congress	states	would	congress	constitution	peace	congress	states	world	congress	public	world
	would	people	peace	would	people	peace	would	states	world	would	public	peace	would	states	peace
	country	country	world	country	country	world	country	country	would	country	country	every	great	country	every
	public	congress	every	public	congress	every	public	public	every	public	congress	would	country	congress	would
	great	public	power	great	public	power	great	congress	government	great	constitution	government	general	present	government
	people	power	government	people	power	government	people	power	nations	people	present	nations	service	national	nations
	citizens	state	nations	citizens	state	nations	general	present	power	general	power	power	people	business	national
Decade		1900–1909			1910–1919			1920–1929			1930–1939			1940–1949	
Topics	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
	government	government	people	government	government	people	government	government	world	government	government	world	government	government	world
	states	would	great	states	would	world	states	congress	people	states	congress	people	congress	federal	people
	united	public	nation	united	public	nation	congress	would	nation	congress	federal	nation	states	congress	nation
	congress	congress	world	congress	congress	great	united	public	great	united	public	peace	united	program	peace
	would	people	peace	would	country	peace	would	federal	peace	would	would	nations	would	public	nations
	great	country	every	american	states	every	american	national	nations	country	national	great	country	legislation	government
	american	states	would	great	people	government	great	country	government	american	legislation	government	great	national	congress
	service	national	government	country	national	would	country	states	would	great	states	congress	service	would	great
	country	business	nations	service	business	nations	service	people	every	service	country	national	american	states	national
	general	present	national	department	federal	national	department	legislation	national	department	program	would	department	administration	economic
Decade		1950–1959			1960–1969			1970–1979			1980–1989				
Topics	1	2	3	1	2	3	1	2	3	1	2	3			
	government	federal	world	government	federal	world	congress	federal	world	congress	administration	world			
	congress	government	people	congress	program	people	government	administration	people	government	federal	people			
	states	congress	nation	states	government	congress	states	program	congress	states	program	congress			
	united	program	nations	united	congress	nation	united	congress	years	energy	congress	years			
	would	legislation	peace	would	administration	years	energy	legislation	nation	united	legislation	america			
	country	administration	congress	energy	legislation	government	administration	government	government	administration	development	american			
	service	public	government	administration	development	peace	would	development	america	would	government	government			
	great	national	years	country	national	nations	legislation	policy	american	legislation	policy	nation			
	american	development	economic	legislation	policy	economic	foreign	national	peace	foreign	national	programs			
	department	would	great	service	public	american	country	states	economic	country	assistance	economic			
Decade		1990–1999			2000–2009			2010–2019							
Topics	1	2	3	1	2	3	1	2	3						
	congress	administration	people	congress	administration	america	congress	administration	america						
	government	federal	america	government	federal	people	government	federal	people						
	states	program	world	states	program	american	states	program	american						
	energy	congress	years	energy	congress	years	energy	legislation	years						
	united	legislation	congress	united	legislation	world	united	congress	world						
	administration	development	american	administration	development	congress	administration	development	americans						
	would	government	government	would	government	americans	would	government	congress						
	legislation	policy	nation	legislation	policy	every	american	policy	every						
	american	national	americans	american	national	government	legislation	national	country						
	country	assistance	every	country	states	nation	country	states	tonight						

3.2. German Macroeconomic Forecasting Reports

As a robustness check, we used the corpus of macroeconomic forecasting reports released by six different German institutions (the German Institute of Economic Research (*DIW*), the Institute for World Economics (*IfW*) in Kiel, the *ifo* institute in Munich, the *RWI* Leibniz Institute for Economic Research in Essen, and the *IWH* Halle Institute for Economic Research in Halle)) during the period 1999–2017 [20], consisting of 292 documents related to monetary policy analyses and recommendations in Germany and the European Union. The *Topic-Scaling* algorithm was applied, leading to an optimal choice of 11 topics, given in Table 5. We notice a regime shift in their respective document positions for all institutions, as demonstrated in Figure 10. Figure 11 shows two distinct groups of learned topics. Topics with scores on the right (positive signs) side of the scale mostly prevailed after the 2008 financial crisis such as the sovereign debt crisis (topic 9), the asset purchase programme (topic 4) and the financial crisis (topic 7), while topics with scores on the left side of the scale (negative sign) contain items usually found in classic monetary policy and economic activity such as interest rate policy (topic 8), inflation rate (topic 11) and international capital markets (topics 2 and 10). This finding is in line with financial market turmoil and successive episodes of instability that continued until 2014, translated with the adoption of *crisis* jargon to describe such episodes. The adoption of such rhetoric concerned all institutions in the corpus.

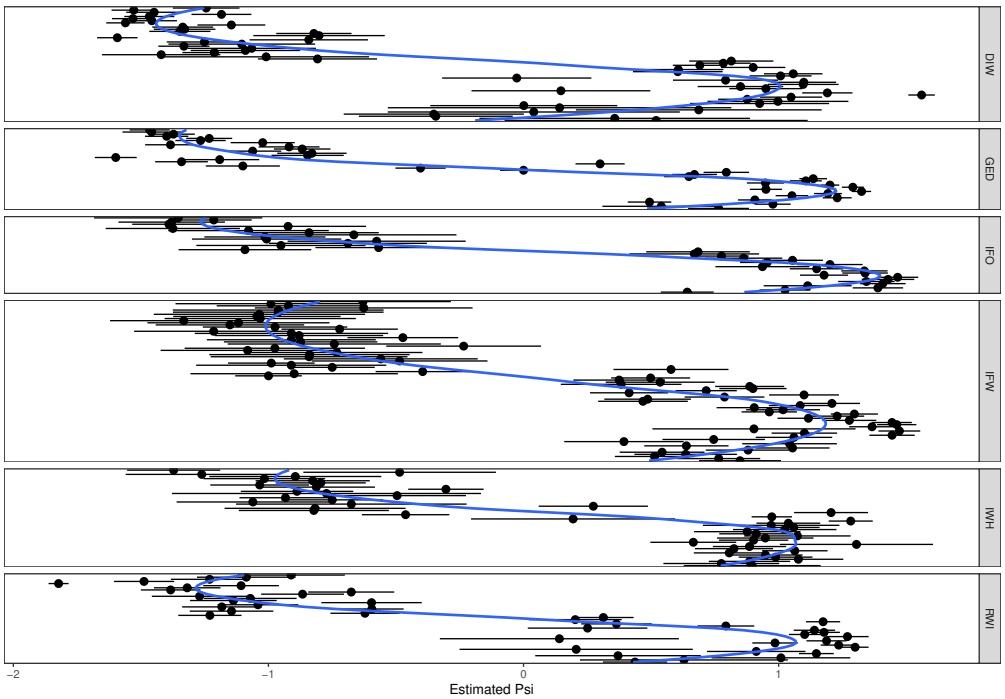

Figure 10. Estimated document positions using Wordfish (blue line is the smoothed Loess curve).

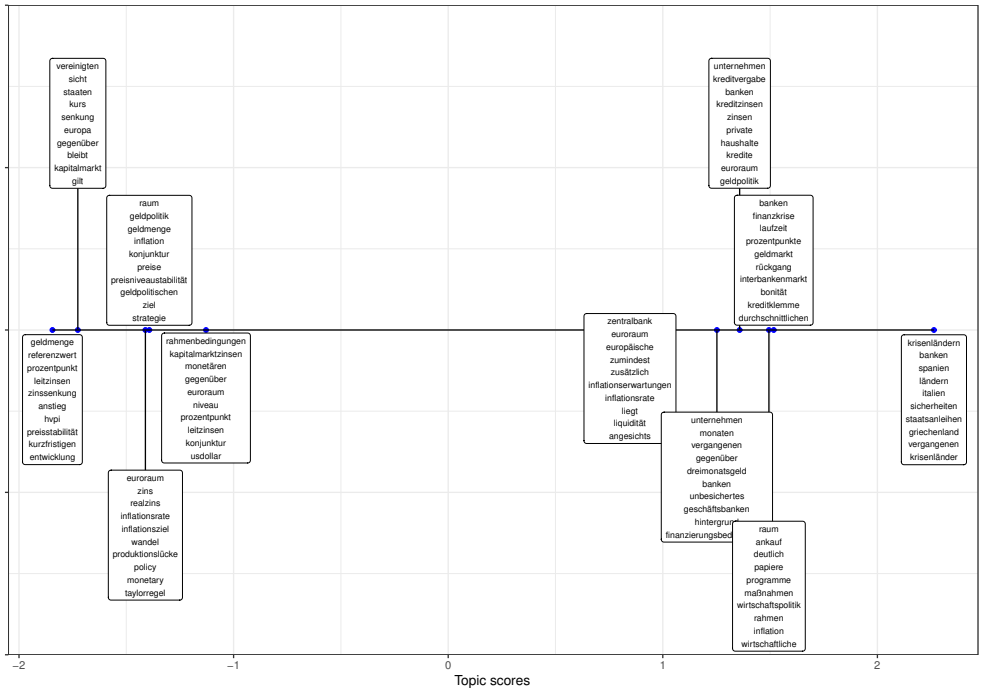

Figure 11. Topic Scores learned via *Topic Scaling* (11 topics).

Table 5. Top 10 words from the estimated *Topic Scaling* with 11 topics.

Topic	Top 10 Words
1	unternehmen, kreditvergabe, banken, kreditzinsen, zinsen, private, haushalte, kredite, euroraum, geldpolitik
2	vereinigten, sicht, staaten, kurs, senkung, europa, gegenüber, bleibt, kapitalmarkt, gilt
3	zentralbank, euroraum, europäische, zumindest, zusätzlich, inflationserwartungen, inflationsrate, liegt, liquidität, angesichts
4	raum, ankauf, deutlich, papiere, programme, maßnahmen, wirtschaftspolitik, rahmen, inflation, wirtschaftliche
5	unternehmen, monaten, vergangenen, gegenüber, dreimonatsgeld, banken, unbesichertes, geschäftsbanken, hintergrund, finanzierungsbedingungen
6	euroraum, zins, realzins, inflationsrate, inflationsziel, wandel, produktionslücke, policy, monetary, taylorregel
7	banken, finanzkrise, laufzeit, prozentpunkte, geldmarkt, rückgang, interbankenmarkt, bonität, kreditklemme, durchschnittlichen
8	geldmenge, referenzwert, prozentpunkt, leitzinsen, zinssenkung, anstieg, hvpi, preisstabilität, kurzfristigen, entwicklung
9	krisenländern, banken, spanien, ländern, italien, sicherheiten, staatsanleihen, griechenland, vergangenen, krisenländer
10	rahmenbedingungen, kapitalmarktzinsen, monetären, gegenüber, euroraum, niveau, prozentpunkt, leitzinsen, konjunktur, usdollar
11	raum, geldpolitik, geldmenge, inflation, konjunktur, preise, preisniveaustabilität, geldpolitischen, ziel, strategie

4. Conclusions

We present a novelty in text mining, suited to study sequential corpora and outperforming other topic models in terms of interpretation and parametrization. *Topic Scaling* could be seen as a dual algorithm: a supervised scaling method where topics are scaled on the same ideological dimension of documents, and a robust alternative to other sequential topic models in which the estimated document scores serve as an ordered variable to retrieve topics rather than a learning process requiring a time frame. Under regularization schemes and entropy-related metrics, increasing the number of topics helps maximize the information gain and uncovering nested structures that render information about potential embedded subtopics, thus unveiling topics that signal important changes to the evolution of the corpus. Applied to study the party duality (Democrats vs Republicans) in the State Of The Union addresses, this method confirms the existence of two distinct periods correlated with the prevailing conditions throughout the modern history of the United States, with a clear dominance of foreign affairs and business discourse in post-war addresses, while recent addresses seem to prioritize security and the economic issues. For the monetary policy forecasting reports in Germany, *Topic Scaling* identified two groups of topics that translate the shift in narrative style that occurred after the financial turbulence of 2008, correlated

with *crisis* jargon used to describe the successive perturbation in financial markets and the banking sector.

Author Contributions: Conceptualization, S.D. and U.F.; methodology, S.D.; software, S.D.; validation, S.D., U.F.; formal analysis, S.D.; investigation, S.D.; resources, U.F.; data curation, S.D.; writing—original draft preparation, S.D.; writing—review and editing, S.D. and U.F.; visualization, S.D..; supervision, U.F.; project administration, U.F.; funding acquisition, U.F.. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 275693836.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Replication code is made available at https://github.com/FritscheU/Topic-Scaling (accessed 12 September 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McCallum, A.; Nigam, K. A comparison of event models for Naive Bayes text classification. In *Proceedings of the IN AAAI-98 Workshop on Learning for Text Categorization*; AAAI Press: Palo Alto, CA, USA, 1998; pp. 41–48.
2. Laver, M.; Benoit, K.; Garry, J. Extracting policy positions from political texts using words as data. *Am. Political Sci. Rev.* **2003**, *97*, 311–331. [CrossRef]
3. Slapin, J.B.; Proksch, S.O. A Scaling Model for Estimating Time-Series Party Positions from Texts. *Am. J. Political Sci.* **2008**, *52*, 705–722. [CrossRef]
4. Lauderdale, B.E.; Herzog, A. Measuring Political Positions from Legislative Speech. *Political Anal.* **2016**, *24*, 374–394. [CrossRef]
5. Grimmer, J.; Stewart, B.M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Anal.* **2013**, *21*, 267–297. [CrossRef]
6. Goet, N.D. Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015. *Political Anal.* **2019**, *27*, 518–539. [CrossRef]
7. Denny, M.J.; Spirling, A. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Anal.* **2018**, *26*, 168–189. [CrossRef]
8. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
9. Boyd-Graber, J.; Mimno, D.; Newman, D. Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In *Handbook of Mixed Membership Models and Their Applications*; CRC Handbooks of Modern Statistical Methods; Chapman and Hall/CRC: Boca Raton, FL, USA, 2014.
10. Blei, D.M.; Lafferty, J.D. Dynamic Topic Models. In Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Pittsburgh, PA, USA, 25–29 June 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 113–120.
11. McAuliffe, J.D.; Blei, D.M. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20*; Platt, J.C., Koller, D., Singer, Y., Roweis, S.T., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2008; pp. 121–128.
12. Boyd-Graber, J.; Resnik, P. Holistic Sentiment Analysis across Languages: Multilingual Supervised Latent Dirichlet Allocation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, Cambridge, MA, USA, 9–11 October 2010; Association for Computational Linguistics: Chicago, IL, USA, 2010; pp. 45–55.
13. Blei, D.M.; Jordan, M.I.; Griffiths, T.L.; Tenenbaum, J.B. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In Proceedings of the 16th International Conference on Neural Information Processing Systems, NIPS'03, Online, 16–20 November 2003; MIT Press: Cambridge, MA, USA, 2003; pp. 17–24.
14. Li, W.; McCallum, A. Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Pittsburgh, PA, USA, 25–29 June 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 577–584.
15. Lu, B.; Ott, M.; Cardie, C.; Tsou, B.K. Multi-Aspect Sentiment Analysis with Topic Models. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDMW '11, Vancouver, BC, Canada, 11 December 2011; IEEE Computer Society: Washington, DC, USA, 2011; pp. 81–88.
16. Wang, C.; Paisley, J.; Blei, D. Online Variational Inference for the Hierarchical Dirichlet Process. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2021; Gordon, G., Dunson, D., Dudík, M., Eds.; Proceedings of Machine Learning Research; JMLR Workshop and Conference Proceedings: Fort Lauderdale, FL, USA, 2011; Volume 15, pp. 752–760.
17. Greene, D.; Cross, J.P. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Anal.* **2017**, *25*, 77–94. [CrossRef]

18. Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S.K.; Albertson, B.; Rand, D.G. Structural Topic Models for Open-Ended Survey Responses. *Am. J. Political Sci.* **2014**, *58*, 1064–1082. [CrossRef]
19. Roberts, M.E.; Stewart, B.M.; Airoldi, E.M. A Model of Text for Experimentation in the Social Sciences. *J. Am. Stat. Assoc.* **2016**, *111*, 988–1003. [CrossRef]
20. Diaf, S.; Döpke, J.; Fritsche, U.; Rockenbach, I. Sharks and minnows in a shoal of words: Measuring latent ideological positions based on text mining techniques. *Eur. J. Political Econ.* **2022**, 102179. [CrossRef]
21. Shogan, C. The President's State of the Union Address: Tradition, Function, and Policy Implications. *Congr. Res. Serv. Rep.* **2016**, R40132. Available online: https://crsreports.congress.gov/product/pdf/R/R40132 (accessed on 12 September 2022).
22. Savoy, J. Text clustering: An application with the State of the Union addresses. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 1645–1654. [CrossRef]
23. Petrocik, J.R.; Benoit, W.L.; Hansen, G.J. Issue Ownership and Presidential Campaigning, 1952–2000. *Political Sci. Q.* **2003**, *118*, 599–626. [CrossRef]
24. Savoy, J. Vocabulary Growth Study: An Example with the State of the Union Addresses. *J. Quant. Linguist.* **2015**, *22*, 289–310. [CrossRef]
25. Savoy, J. Text representation strategies: An example with the State of the union addresses. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 1858–1870. [CrossRef]
26. Lei, L.; Wen, J. Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses. *Lingua* **2020**, *239*, 102762. [CrossRef]
27. Teten, R.L. Evolution of the Modern Rhetorical Presidency: Presidential Presentation and Development of the State of the Union Address. *Pres. Stud. Q.* **2003**, *33*, 333–346. [CrossRef]
28. Cummins, J. State of the Union addresses and presidential position taking: Do presidents back their rhetoric in the legislative arena? *Soc. Sci. J.* **2008**, *45*, 365–381. [CrossRef]
29. Lo, J.; Proksch, S.O.; Slapin, J.B. Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos. *Br. J. Political Sci.* **2016**, *46*, 591–610. [CrossRef]
30. Koltcov, S.; Ignatenko, V.; Boukhers, Z.; Staab, S. Analyzing the Influence of Hyper-parameters and Regularizers of Topic Modeling in Terms of Renyi Entropy. *Entropy* **2020**, *22*, 394. [CrossRef] [PubMed]
31. Benoit, K.; Watanabe, K.; Wang, H.; Nulty, P.; Obeng, A.; Müller, S.; Matsuo, A. quanteda: An R package for the quantitative analysis of textual data. *J. Open Source Softw.* **2018**, *3*, 774. [CrossRef]
32. Honnibal, M.; Montani, I.; Landeghem, S.V.; Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. **2020** Available online: https://doi.org/10.5281/zenodo.3358113 (accessed on 12 September 2022).
33. Cleveland, W.S.; Loader, C. Smoothing by Local Regression: Principles and Methods. In *Proceedings of the Statistical Theory and Computational Aspects of Smoothing*; Härdle, W., Schimek, M.G., Eds.; Physica-Verlag HD: Heidelberg, Germany, 1996; pp. 10–49.