CommunityFish:
A Poisson-based Document Scaling With Hierarchical Clustering

Sami Diaf
Universität Hamburg
Department of Socioeconomics
sami.diaf@uni-hamburg.de

Abstract

Document scaling has been a key component of modern text-as-data applications in social sciences, particularly for political scientists, who aim at uncovering differences between speakers or parties with the help of probabilistic and non-probabilistic approaches. Yet, most of these techniques employ the bag-of-word hypothesis and disregard semantic features or use prior information borrowed from external sources that may bias the results. This paper presents *CommunityFish* as an augmented version of *Wordfish* based on a prior hierarchical clustering of the word space to retrieve semantic n-grams, or *communities*, as signals emerging from the corpus to be used as an input to *Wordfish*. Instead of considering all words in the corpus as independent features, we emphasize the interpretability of the results, since communities have the ability to better scale parties or speakers, and ensure a faster convergence when considering a Poisson-based ranking model. Aside from yielding communities assumed to be subtopics summarizing the corpus' narrative signals, the application of this technique outperforms the classic *Wordfish* model by emphasizing key historical developments in the U.S. State of the Union addresses and was found to replicate the prevailing political stance in Germany when using the corpus of parties' manifestos.

1 Introduction

Comparative politics has been a prominent domain of application of what is currently known as text-as-data field, featuring the use of text mining techniques and machine learning algorithms to identify patterns that differentiate documents or track disparities at the meta-data level. Scaling techniques typically comprise an array of unsupervised methods, both probabilistic and non-probabilistic, which aim to extract one or multiple dimensions to enable metadata comparisons, based on a set of assumptions conducted at the word-level.

Earlier scaling techniques used statistical learning approaches as for matrix factorization schemes (Deerwester et al., 1990) and a probabilistic model based on the Poisson distribution as for *Wordfish* (Slapin and Proksch, 2008; Lowe and Benoit, 2013) which ranks documents on a unidimensional scale using word occurrences in the corpus. Further extensions of Poisson scaling models considered a debate structure (Lauderdale and Herzog, 2016), pre-trained embedding models (Nanni et al., 2019), word variations (Vafa et al., 2020) and semantic search strategies (Diaf and Fritsche, 2022b), providing an improved scaling of documents depending on several assumptions and use cases at the word or document levels.

Regarding *Wordfish*, the Poisson scaling model uses word counts to learn a hidden and normally-distributed dimension, assumed to be a proxy of partisanship among political parties when scaling manifestos (Slapin and Proksch, 2008). However, the Poisson distribution does not always pertain (Lowe and Benoit, 2013), as frequent words are likely to be normally distributed, while very rare words tend to substantially deviate from the Poisson paradigm (Lo et al., 2016). Another disadvantage is the dynamic word usage which needs time-varying parameters for the Poisson ranking model and further constraints on parameters to ensure its stability (Jentsch et al., 2020), or to consider the structure *document-topic-word* to get polarization at the topic level using a hybrid supervised topic model (Diaf and Fritsche, 2022a).

Although the choice of scaling techniques is abundant, it may not always meet the expectation of practitioners, as the inference is done at the word-level, while the analysis often targets documents' content in terms of groups of words that convey the interest of researchers. The word contribution to the built scale in *Wordfish* is static and cannot be fully interpretable if the corpus has undergone significant changes over time, in terms of word usage, between parties/speakers (Jentsch et al., 2020). Furthermore, the polarity of specific words could be different from the position of documents they are mostly related to, thus not in-line with experts' assessments (Hjorth et al., 2015). This issue arises from the bag-of-word assumption and the underlying agnostic hypothesis of word independence, which prevents an accurate scaling of documents based on semantic features (Nanni et al., 2019).

Advances in social network analysis indicated that hierarchical clustering can reveal homogeneous and distinct groups of users, commonly referred to as *communities*, based on their interactions, which could also be used in text mining to identify independent, semantic groups of words, in form of n-grams, that differentiate documents by their occurrences while delivering informative signals that outperform analyses based on single-word usage. One popular algorithm for studying social networks is the *Louvain* algorithm (Blondel et al., 2008) which was applied to get word groups that better represent the rhetoric used in a given corpus (Bail, 2016) or to study the lexical shift in the State Of The Union addresses (Rule et al., 2015). Other hierarchical clustering schemes were proposed as for *Infomap* (Rosvall and Bergstrom, 2008) which uses random walk map-equation instead of optimizing the modularity as for *Louvain* (Lancichinetti and Fortunato, 2009), and *Leiden* (Traag et al., 2019) which was found to outperform *Louvain* when applied to big networks, however, similar performances with *Louvain* are expected on smaller networks.

This paper extends the idea of *lexical shift* (Rule et al., 2015) by identifying communities as representative groups of words, able to achieve a fast and interpretable scaling of documents upon which a Poisson ranking model could be built, instead of considering a plain word-count model related to the bag-of-word hypothesis. I argue that communities offer a better polarization level when differentiating documents and metadata than standard bag-of-word techniques, in addition to efficiently speeding up the learning process by reducing the size of the document-term-matrix whose sparsity may hinder the convergence of Poisson models. Commonly used words are likely to form communities with a high frequency of words but are less likely to be polarized compared to communities with exclusive word usage, denoting the focus of a given speaker/party on a specific subject of item that could be identified without the need to run topic models.

Two historical corpora, in English and German, were selected to evaluate this novel approach. The application on the U.S. State Of The Union (SOTU) addresses (1854-2019) shows a dominance of historical developments as for economic issues, local affairs and foreign policy that ranked addresses on a two-regime scale whose transition could be identified during the great depression. From the analysis of German political parties' manifestos (2013, 2017 and 2022), *CommunityFish* identified granular themes at the center of election debates that were found to replicate the ideological spectrum of political parties with *AFD* and *Linke* parties being the ideological bounds of the learned scale, while other parties seem to share many featured themes, hence reinforcing their centrist positions.

The paper outlines the build-up of *CommunityFish* from a network analysis perspective (Section 2) and from statistical learning (Section 3), then implements the proposed algorithm on two corpora (Section 4) and compares to the standard *Wordfish* used by practitioners.

2 Methodology

2.1 Network Analysis

Analysis of social media drove the attention of scientists on the necessity to adopt advanced clustering methods able to extract information that describe relationships between users via the types of messages or ideas they produce (White, 2008), instead of simple relationship structures between individuals (Bail, 2016).

Network analysis witnessed important contributions on identifying distinct subgroups in social networks, built on several optimization schemes developed to offer intuitive clustering (Lancichinetti and Fortunato, 2009).

For such tasks, researchers should carefully select clustering methods for community detection and also take into account centrality scores (Mester et al., 2021). *Louvain* algorithm (Blondel et al., 2008) is one commonly used clustering technique ,usually preferred to *FastGreedy* algorithm (Clauset et al., 2004), due to its relative low complexity, as it achieves a local optimization of the modularity Q at the node-level, defined as :

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

with A_{ij} representing the edge weight between nodes i and j, k_i and k_j are the sum of the weights of the edges attached to nodes i and j, respectively; m is the sum of all of the edge weights in the graph; c_i and c_j are the communities of the nodes; and δ is Kronecker delta function $\delta(x, y) = 1$ if x=y, 0 otherwise.

Louvain clustering iteratively optimizes the modularity Q by starting with different node being its own community, and the concept is to place a node n_i to one of its neighboring nodes community, in a way to maximize the modularity change (Mester et al., 2021). Similar to users in social networks, *Louvain* algorithm can cluster words in a corpus, so to extract *communities*, in a form of n-grams of different lengths, having an independent, non-overlapping structure stemming from the specific word usage found in documents.

Traag et al. (2019) proposed *Leiden* clustering as a reliable alternative to *Louvain* in discerning small connected communities in large network structures. Altough *Leiden* was found to be faster than *Louvain*, in terms of execution, both do not differ when the network structure is relatively small, as for collection of documents with limited vocabulary, meaning the community structures of both algorithms can share many similarities and just slightly differ in the number of uncovered clusters.

2.2 Poisson ranking model

To apply *CommunityFish*, the corpus is broken down into bigrams and a minimum threshold π is set before running *Louvain* algorithm that yields K communities used as features for the Document-Term-Matrix (DTM), instead of considering all words in the corpus, hence communities serve as features to the Wordfish scaling algorithm. This scheme could be seen as a semantic clustering of the DTM that identifies correlated pairs of words in local contexts, thanks to a hierarchical clustering on bigrams, which differs from a simple bigram grouping of the initial DTM features.

The resulting DTM, as a matrix of communities' frequencies on each document in the corpus, is given as an input to *Worfish* (Slapin and Proksch, 2008) to learn document positions, or ideal points, that scale documents based on the occurrence of communities. As a scaling technique, *Wordfish* uncovers a latent scale θ, assumed to be a proxy of partisanship or ideological differences between parties or speakers, depending on the used context.

Although the use of Poisson distribution is justified by the occurrence of words in the corpus, assumed to be rare events, it is not always applicable to cases where the word usage concerns few documents, meaning the Poisson's expectation departs significantly from the variance (Lowe and Benoit, 2013; Lo et al., 2016) even though a quasi-Poisson scheme can relax the Poisson assumption of the mean-variance equality.

I argue that considering communities frees the DTM from potential biases raised by rare words and allows a faster convergence of *Wordfish* algorithm when applied to big corpora. *CommunityFish* could be seen as a double dimensionality reduction technique: first to uncover communities, as the primary unit of analysis, and second to learn one scale of ideal points using a Poisson ranking model.

Algorithm: CommunityFish

1.Community detection: Run a hierarchical algorithm (*Louvain*) over the bigram features of the corpus and extract K groups of words or *communities*, whose occurrence in the corpus is greater than π.
2.Poisson scaling model: The K communities are used as features for the Document-Term-Matrix, to be given as input to the Poisson scaling model (Slapin and Proksch, 2008) to uncover the scale θ_i from the specification:
$log(\lambda_{ij}) = \alpha_i + \psi_j + \theta_i\beta_j$, where:
λ_{ij}: frequency of the community j in document i
α_i: document fixed effect
ψ_j: community fixed effect
θ_i: the *position* of document i
β_j: the effect of community j to the document position

The hierarchical clustering applied to the corpus (*Louvain* algorithm) may be regarded as an implicit factorization of the traditional unigram DTM, yielding an interpretable feature matrix stemming from the learned communities. Aside from lowering the DTM dimension, it permits to intuitively concentrate the scaling on meaningful and independent groups of words (*communities*), that discriminate the ideal points based on their occurrences in the documents.

3 Application

3.1 State of the Union

State of the Union (SOTU) addresses consist of annual speeches given by U.S. presidents during the period (1854-2019), so to emphasize the duality democratic-republican in the scaling (Diaf and Fritsche, 2022a). The corpus was lemmatized using *udpipe* model (Straka et al., 2016) to reduce the size of the Document-Term-Matrix and learn robust communities, in comparison with the raw corpus. The application of the *Louvain* algorithm yielded 52 different communities (Table 1) with a clear historical context that spans over one and half century, tied to different episodes of modern American history. From Table 1, 22 communities, out of 52, are constituted of bigrams and the remaining are n-grams of different lengths comprising entities, expressions as well as plans or programs[1].

Communities, whose contributions to the scale β_j are different from zero, polarize the overall scale θ via their respective signs. From Figure 1, communities 45, 40, 11 and 8 contribute to documents whose positions in the overall scale (Figure 2) are positive, consisting of earlier addresses from the second half of the ninetieth century that targeted foreign policy and local administration. On the other hand, modern addresses have negative positions (Figure 2) and demonstrate a strong influence of foreign policy and defense interests (communities 38 and 49) as well as business/economic environment (communities 43 and 2). Figure 2 shows a two-regime scale of ideal points, whose transition occurred during the great depression (Hoover's addresses during the period 1929-1933, coinciding with the position $\hat{\theta} = 0$), suggesting a potential shift in the rhetoric, or a transition into modern addresses, used by U.S. presidents and captured via communities that could be assumed to be proxies for most discussed interests in their addresses.

In comparison to classic *Wordfish* application on the same corpus (Diaf and Fritsche, 2022a), the learned document positions are quiet similar, but cannot be differentiated in small periods, even if given by different speakers. Word contributions (Figure 5) obtained via *Wordfish* offer clustered, heavily centered densities, with tails dominated by rare words that occurred in a relatively small

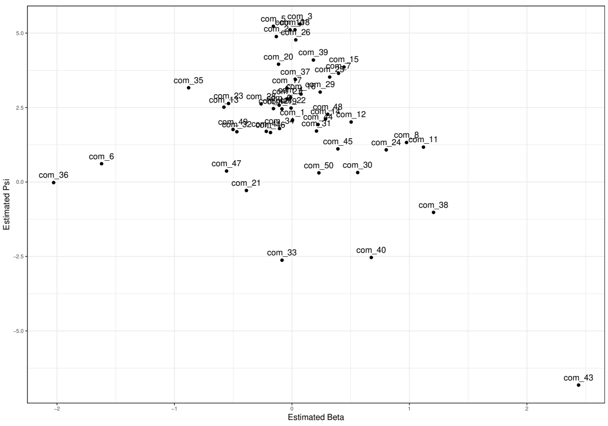

Figure 1: Communities contributions to the scale (β) vs communities' positions ψ (SOTU corpus)

number of documents.

3.2 German Manifesto

The corpus of Manifesto Project (Lehmann et al., 2022) was used to get the manifestos of six main German political parties, during the period 2013-2021 (Diaf and Fritsche, 2022b), then lemmatized using *udpipe* German language model (Straka et al., 2016) to reduce the vocabulary length of the corpus. It resulted 45 communities (Table 2) reproducing most of the debated themes in social life, politics and economic development which constitute the basis of the learned scale (Figure 4), found to replicate the prevailing political partisanship in Germany. The *AFD* and *Linke* parties represent the opposite ends of the learned scale, while the other parties hold central positions, with noticeable firm positions (small standard deviations of their ideal points) of the *Linke* and *Grüne* parties throughout the studied period. Conversely, the positions of *AFD* and *CDU* exhibit the highest variability, evidenced by wider standard errors. The blue line in Figure 4 is the local polynomial regression *Loess curve* (Jacoby, 2000) used to separate parties into two distinct classes (left-right) based on learned scale from the established communities (Table 2), resulting into a bi-partisanship *AFD-CDU-FDP* and *SPD-Grüne-Linke*.

From Figure 2, communities 40 and 45 support the position of the *Linke* party, as their contribution to the scale is strongly positive, in comparison to communities 5, 11 and 12 whose β_j are still positive but rather close to the origin. Most of the learned communities have a low contribution to the scale ($\beta_j \to 0$) and denote shared interests debated by political parties.

[1] *Leiden* clustering yielded a similar community structure to *Louvain*, with minor differences concerning two communities, out of 52. The same results were found using the German political manifesto corpus.

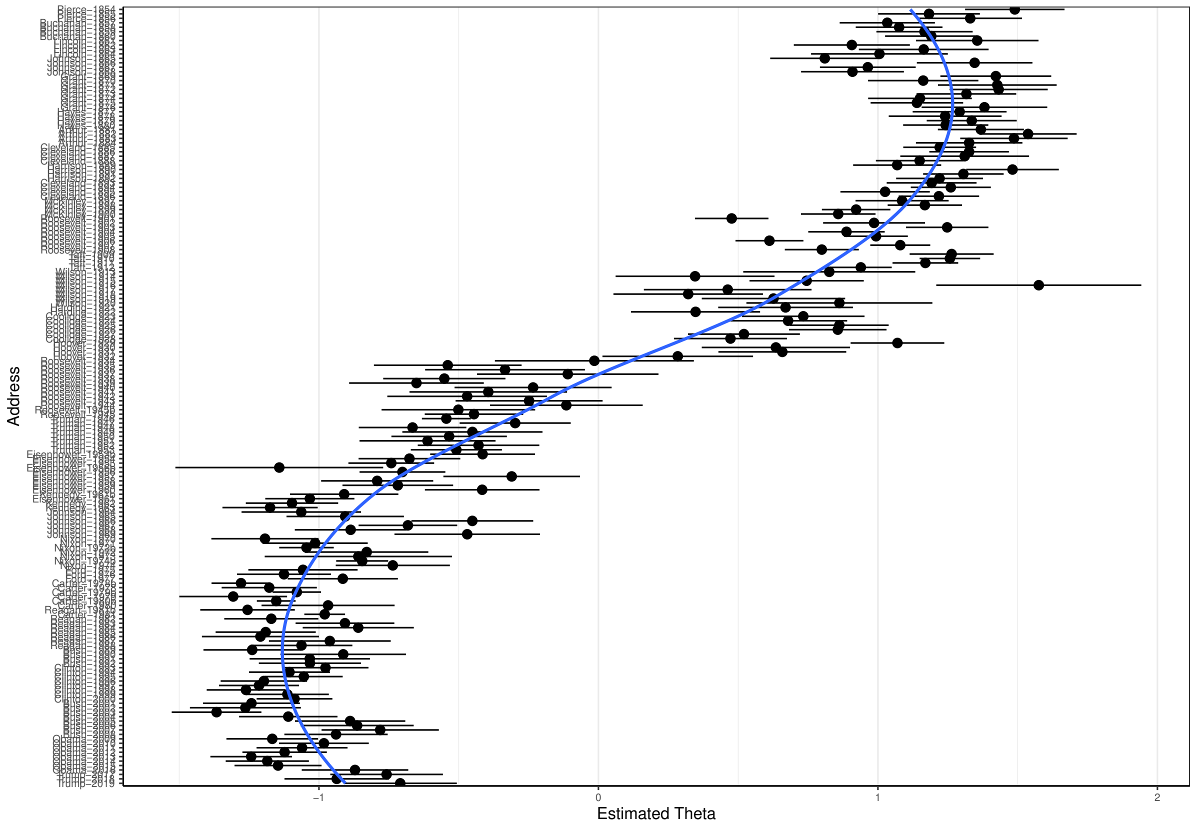

Figure 2: Learned CommunityFish ideal points with 95% confidence intervals (SOTU Corpus).

As a comparison to Wordfish (Figure 6), *CommunityFish* highlights a better polarization *AFD-Linke*, and a clear partisanship even if document positions exhibit a higher variability, in terms of standard errors, than *Wordfish*.

4 Conclusion

Scaling techniques are valuable analytical tools used by political scientists to explore partisanship among parties and to understand the ideological spectrum of speakers. Nonetheless, they are limited by the fact that they consider only words as the unit of analysis, making their application agnostic vis-à-vis semantic signals emerging from the corpus. While numerous solutions were developed to improve scaling results by incorporating external information sources as priors, the use of hierarchical clustering, as a pre-processing step, enables the identification of *communities*, as resilient clusters, with semantic effectiveness and substantial results, combined with a faster execution time. *CommunityFish* is a scaling technique that translates the unit of analysis from words to communities and an implicit factorization of the document-feature-matrix, unveiling informative sub-topic structures for an in-depth scaling of historical corpora as well as political manifestos. Optimal use of *CommunityFish* requires selecting most informative communities in an already-lemmatized corpus by mean of a clustering technique (such as *Louvain* or *Leiden* algorithms). This ensures an independent community structure when aggregating the document-feature-matrix, helping the spread of the ideological stance learned via Poisson ranking model, which was found to outperform classic *Wordfish* without calling expensive, often biased, prior information. Applied to two distinct corpora, it demonstrated a great ability in extracting communities from a language-variable corpus (SOTU) and identifying common items in debate-based documents (German manifesto) for an efficient and meaningful scaling of documents.

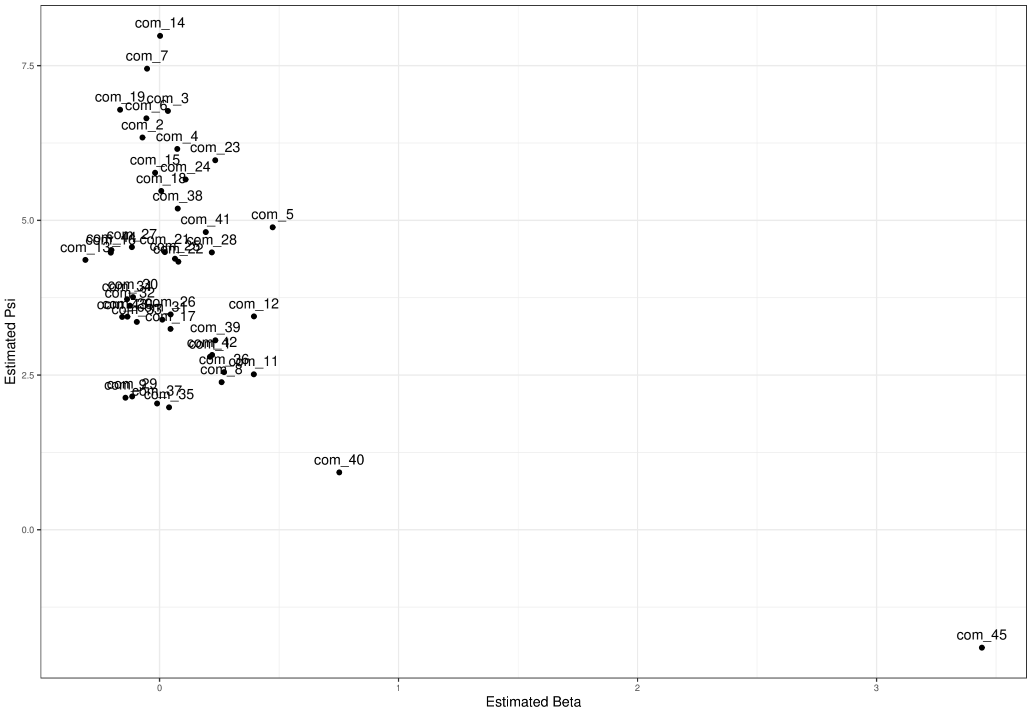

Figure 3: Communities contributions to the scale (β) vs communities' positions ψ (German Manifesto corpus)

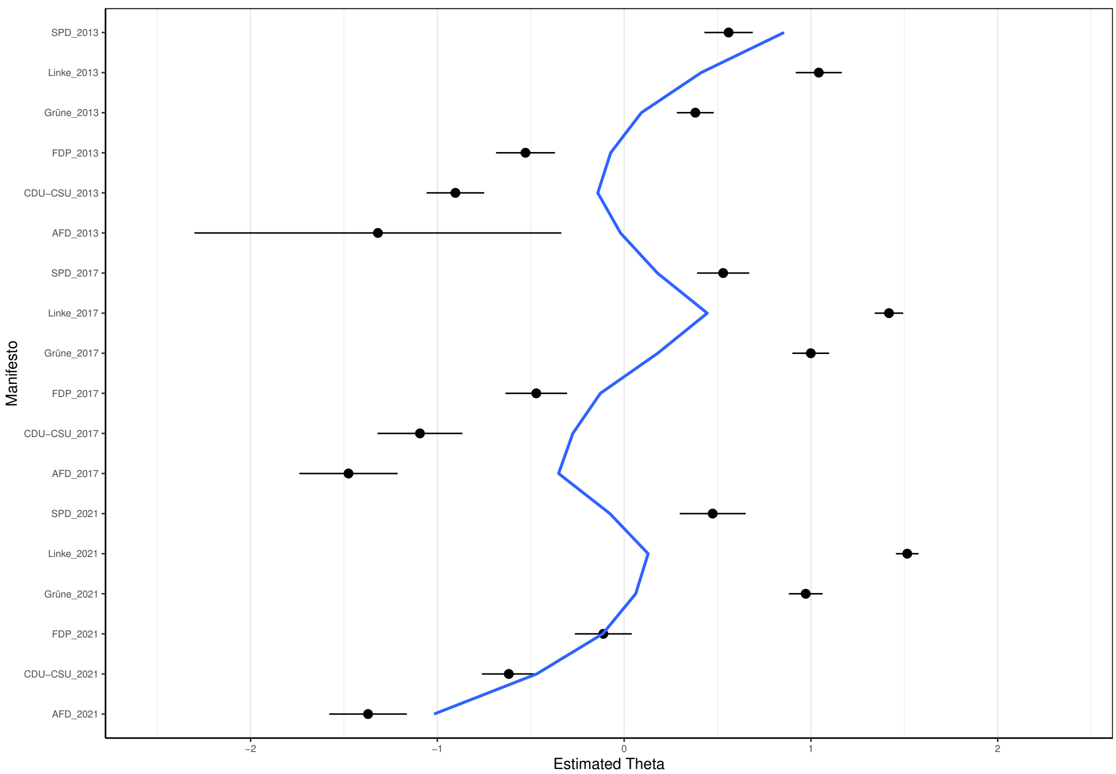

Figure 4: Learned CommunityFish ideal points with 95% confidence intervals (German Manifesto Corpus).

Table1: Communities in SOTU corpus

Community	Words
com_1	agricultural, product
com_2	american , billion , business , enlist , every , fellow , million , silver , small young , citizen , family , people , republics, dollar , man , day , americans
com_3	annual , special, message
com_4	armed , military, naval , force
com_5	ask , come , current , end , fiscal , five , four , last , many , next , past precede , previous, recent , ten , three , two , year , congress, june , session , ago , ahead
com_6	attorney , british , can , federal , general , government, local , make , must national , postmaster, self , social , spanish , supreme , help , court , sure also , continue , bank , defense , security
com_7	balanced, budget
com_8	base , call , confer , depend , enter , impose , urge , upon , attention
com_9	careful , favorable , consideration
com_10	central, latin , south , america
com_11	civil , hard , human , interest , postal , public , right , tax , work , service , rate debt , building , land , opinion , now , credit , cut , reduction, together
com_12	commerce , interstate, commission
com_13	earnestly, recommend
com_14	economic , development, growth
com_15	executive, branch , order
com_16	exist , international, law , present , tariff , enforcement , condition , system
com_17	far , thus , reach
com_18	first, time
com_19	foreign , free , great , nation , office , post , take , treasury , war , world country com_ , power , trade , britain , department, place , ii
com_20	full , employment
com_21	go , look , move , forward
com_22	god , bless
com_23	good , faith
com_24	health , medical , care , insurance
com_25	high , level , priority, school
com_26	internal, revenue
com_27	large , number, part
com_28	let, us
com_29	long, run , term
com_30	low , income
com_31	may , well
com_32	merchant, marine
com_33	middle, class , east
com_34	minimum, wage , worker
com_35	mr , speaker
com_36	natural , resource
com_37	new , job , program, york
com_38	nuclear, weapon
com_39	one , half , hundred, third
com_40	panama, canal
com_41	per , annum, cent
com_42	philippine, islands
com_43	private , enterprise, sector
com_44	progress, step , toward
com_45	puerto, rico
com_46	set , forth
com_47	several, united , states , nations
com_48	sink, fund
com_49	soviet, union
com_50	vice , president
com_51	welfare, reform
com_52	white, house

Table 2: Communities in German Manifesto corpus

Community	Words
com_1	abkomme, abkommen
com_2	afd, demokrat, deshalb , fordern , frei, linke, stehen, setzen
com_3	alt, brauchen, immer, jung, mehr , mensch million, gerechen, stark, geld, personal, transparenz, zeit
com_4	arbeit, beruflich, gut, kulturell, selbstbestimmt, arbeiten bildung, arbeitsbedingung, leben, zukunft
com_5	arbeitgeber , arbeitnehmer, patient , verbraucher , innen
com_6	arbeitsplatz , dass , deutschland , einsetzen , ganz , gestalten jed , neu , schaffen , sicherstellen, sorgen verhindern zeigen , einzeln , form , kind , technologie
com_7	beitrag , bund , dabei, etwa, gelten , gerade , gesellschaftlich, insbesondere , land , mittel, projekt , regelung sollen , sowie , teilhabe, wichtig, zugang , leisten, na, mitteln , rolle
com_8	bezahlbar, wohnraum
com_9	biologisch, vielfalt
com_10	cdu, csu
com_11	corona, krise
com_12	demokratisch, kontrolle
com_13	deutsch, bundestag, sprache
com_14	digital, it, sozial, infrastruktur, welt , sicherheit, absicherung gerechtigkeit, marktwirtschaft , netzwerk, sicherungssystem wohnungsbau, zusammenhalt
com_15	drei, euro, letzt, milliarde, mrd, pro, seit, vergangen, vier, zehn, jahr
com_16	erhalten, bleiben
com_17	erneuerbare , erneuerbaren, energie, energien
com_18	erst, schritt
com_19	eu, ebene, kommission, mitgliedstaat, staat
com_20	fair, wettbewerb
com_21	gering, hoch, mittler, einkommen , unternehmen
com_22	gesetzlich, mindestlohn, rent, rentenversicherung
com_23	gleich, recht , chance, lohn , rechte
com_24	hartz, iv
com_25	lage, versetzen
com_26	medizinisch, versorgung
com_27	nachhaltig, wirtschaftlich, entwicklung
com_28	offen, gesellschaft
com_29	qualitativ, hochwertig
com_30	rechnung, tragen
com_31	rechtlich, rundfunk
com_32	regel, regeln
com_33	schnell, internet
com_34	schon, heute
com_35	schwarz, gelb
com_36	sexuell, orientierung
com_37	start, ups
com_38	stelle, stellen
com_39	strukturschwach, region
com_40	stunde, stunden
com_41	teil, teilen
com_42	treffen, triefen
com_43	verein, vereinen
com_44	vereint, nation
com_45	vgl, kapitel

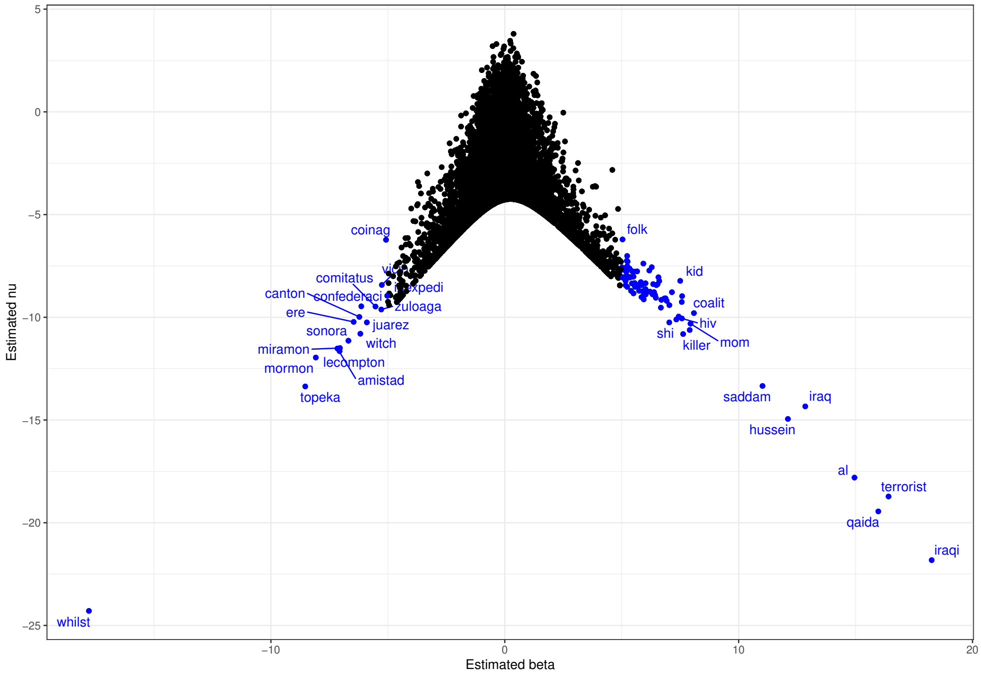

Figure 5: Word contributions from *Wordfish* (SOTU Corpus) (Diaf and Fritsche, 2022a)

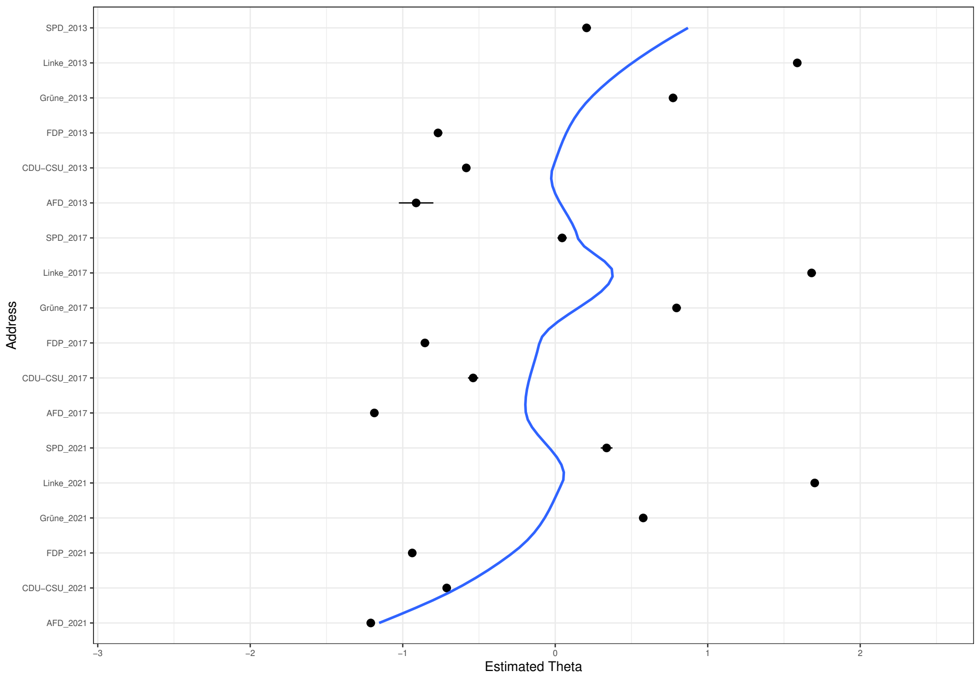

Figure 6: Learned *Wordfish* ideal points with 95% confidence intervals (German Manifesto Corpus). Blue line is the Loess curve.

References

Christopher A. Bail. 2016. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, 113(42):11823–11828.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Sami Diaf and Ulrich Fritsche. 2022a. Topic scaling: A joint document scaling-topic model approach to learn time-specific topics. *Algorithms*, 15(11).

Sami Diaf and Ulrich Fritsche. 2022b. TopicShoal: Scaling partisanship using semantic search. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 167–174, Potsdam, Germany.

Frederik Hjorth, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. 2015. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics*, 2(2):2053168015580476.

William G. Jacoby. 2000. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613.

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2020. Time-dependent poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, 142:106813.

Andrea Lancichinetti and Santo Fortunato. 2009. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117.

Benjamin E. Lauderdale and Alexander Herzog. 2016. Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394.

Pola Lehmann, Tobias Burst, Theres Matthieß, Sven Regel, Andrea Volkens, Bernhard Weßels, and Lisa Zehnter. 2022. The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2022a.

James Lo, Sven-Oliver Proksch, and Jonathan B. Slapin. 2016. Ideological clarity in multiparty competition: A new measure and test using election manifestos. *British Journal of Political Science*, 46(3):591–610.

Will Lowe and Kenneth Benoit. 2013. Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, 21(3):298–313.

Attila Mester, Andrei Pop, Bogdan-Eduard-Mădălin Mursa, Horea Greblă, Laura Dioşan, and Camelia Chira. 2021. Network analysis based on important node selection and community detection. *Mathematics*, 9(18).

Federico Nanni, Goran Glavas, Ines Rehbein, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2019. Political text scaling meets computational semantics. *arXiv preprint arXiv:1904.06217*.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Milan Straka, Jan Hajič, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1).

Keyon Vafa, Suresh Naidu, and David Blei. 2020. Text-based ideal points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5345–5357, Online. Association for Computational Linguistics.

Harrison C. White. 2008. *Identity and Control. How Social Formations Emerge (Second Edition)*. Princeton University Press, Princeton.