Grouping Democracy Concepts through Consensus Clustering of Lexical and Semantic Representations

Matjaž Kumin
matjaz.kumin@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Žiga Kolar
ziga.kolar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Student paper. Democratic theory contains labels for regimes, institutions, participation modes, ideological traditions, and technological modalities, but their conceptual relations are rarely analysed at scale. We present a standardized corpus of 100 democracy concepts. Each concept is represented by a structured English description and mapped in two complementary spaces: word n-gram TF–IDF and contextual sentence embeddings. K-means clustering is repeated across 10 random seeds and five candidate values of k; representation-specific co-association matrices are then combined into an equal-weight consensus grouping. Agreement and stability are assessed with ARI, AMI, NMI, and silhouette scores. The selected eight-cluster solution suggests reasonably balanced macrofamilies spanning constitutional-representative, electoral, digital-participatory, participatory-governance, local and community-based, direct-deliberative, representative, and group-based forms. The two spaces agree only partially, indicating that lexical and contextual evidence should be treated as complementary rather than interchangeable.

Keywords

democracy concepts, consensus clustering, TF–IDF, sentence embeddings, K-means, UMAP, normalized mutual information, political science, taxonomy

1 Introduction

Democracy is one of the most widely studied concepts in political science, but its many variants are usually compared through qualitative typologies and expert interpretation [5]. Such approaches are theoretically rich, but difficult to scale or replicate across many democracy concepts. This student paper develops a computational approach for grouping democracy concepts directly from standardized textual descriptions.

The study uses a corpus of 100 democracy concepts represented in two complementary textual spaces: lexical TF–IDF and contextual sentence embeddings. Repeated K-means clustering is used to construct co-association matrices for both representations, which are then combined into a consensus grouping. We ask which macro-families emerge from the corpus and how strongly lexical and semantic representations agree. The contribution is threefold: a reproducible concept-level pipeline, an interpretable grouping of democracy concepts, and a comparison of lexical and semantic evidence in mapping democratic theory.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.12

2 Related Work

Existing computational studies of democracy usually estimate country-level democracy scores, classify regime properties, or analyse political discourse with supervised machine-learning and NLP methods [4, 8, 10, 15]. A related literature on e-democracy and digital platforms examines how online tools (including large language models) reshape participation, legitimacy, and governance practices [1, 3, 6, 7, 9, 11, 12, 16, 17]. Foundational democratic theory, including polyarchy, deliberative democracy, liberal democracy, and embedded democracy, provides rich conceptual typologies [5, 14], but does not offer a scalable procedure for comparing many democracy concepts on shared textual ground. Our contribution is therefore concept-level rather than country-level: we compare standardized definitions of democracy concepts and derive interpretable macro-families from their lexical and semantic similarities.

3 Data and Methods

3.1 Standardized Concept Corpus

The corpus contains descriptions of 100 democracy concepts, which were identified through a review of political science sources and scholarly literature, including two political science encyclopaedias [13, 2], academic monographs and journal articles. To ensure coverage of emerging and peripheral conceptual variants, the corpus also includes concepts described primarily in non-academic sources such as Wikipedia and online educational resources; these were included because the study aims to map the full conceptual space of democracy as it appears in contemporary discourse, rather than to restrict analysis to canonically established concepts. Standardized descriptions were synthesized with a large language model (GPT-5.5) using the identified sources. The model reorganized source material into a common structure while preserving the substantive definitions found in the literature. Each description was structured according to a fixed nine-field template: title, definition, core mechanism, institutional setting, participation mode, decision rule, risks or limits, keywords, and dimensions. Concatenating these fields produces descriptions of approximately 152–176 words. The generated descriptions were subsequently checked against the original source documents using a retrieval-augmented generation system (NotebookLM), which retrieved supporting passages from the same source collection to assist consistency verification. In addition, all 100 standardized descriptions were manually compared with the underlying source material by the authors, focusing on factual fidelity, preservation of defining mechanisms, omission of important qualifications, and unsupported additions introduced during LLM-based standardization. Using a common template reduces variation caused by differences in length, writing style, and document structure, allowing similarities to reflect conceptual content rather than formatting.

3.2 Textual Representation

Two complementary textual representations were constructed.

The lexical representation was obtained using TF–IDF vectorization with word n-grams ranging from one to three tokens. Because TF–IDF vectors are high-dimensional and sparse, they primarily capture controlled terminology and explicit lexical overlap between democracy concepts.

The semantic representation was obtained using Sentence-Transformers models. Sentence embeddings were L2-normalized before similarity computation, enabling cosine similarity to be computed directly through dot products. Whereas TF–IDF emphasizes shared vocabulary, sentence embeddings capture broader semantic relationships between concept descriptions.

3.3 Dimensionality Reduction and Clustering

Dimensionality reduction was used both for visual comparison and as a standardized low-dimensional space for the consensus clustering experiments. Uniform Manifold Approximation and Projection (UMAP) was computed using cosine distance for sentence embeddings and cosine-distance matrices derived from TF–IDF representations. Following exploratory experiments, the parameters were fixed at $n_neighbors = 10$ and $min_dist = 0.05$, providing stable neighborhood structure without excessive fragmentation.

To compare lexical and semantic representations under identical conditions, UMAP projections were clustered using K-means. Multiple random initializations ($n_init = 50$) were employed together with three random seeds (42, 123, and 999). Candidate solutions were generated for $k = 4$ to 15 clusters.

The final concept grouping was derived from repeated K-means clustering over three-dimensional UMAP representations of both textual spaces. For each seed, UMAP was computed separately for the embedding and TF–IDF representations, and K-means was then repeated across ten random seeds (2, 7, 17, 29, 42, 53, 71, 123, 432, 999), and five candidate values of k ($k \in \{6, 7, 8, 9, 10\}$).

3.4 Evaluation and Consensus Analysis

Agreement between lexical and semantic clustering solutions was evaluated using the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), with Normalized Mutual Information (NMI) as a supplementary measure. These indices compare cluster assignments, while ARI and AMI correct for agreement expected by chance.

Internal cluster quality was assessed using the silhouette coefficient computed in the corresponding representation space. Rather than selecting a single clustering solution, repeated K-means runs were used to evaluate stability both within and across representations.

For each representation, co-association matrices were constructed by recording how frequently pairs of democracy concepts appeared in the same cluster across repeated runs. The embedding and TF–IDF based co-association matrices were then averaged to obtain a consensus similarity matrix. A consensus distance matrix ($D = 1 - C$) was subsequently clustered using average-linkage agglomerative clustering, producing the final set of macro-families. Membership strength was calculated as the average co-association of each democracy concept with the other members of its assigned cluster, providing an indicator of cluster stability and conceptual coherence.

4 Results

4.1 Clustering Stability and Metric Agreement

K-means was applied independently in the sentence embedding space and in the TF–IDF UMAP space across ten random seeds and five candidate values of k ($k \in \{6, 7, 8, 9, 10\}$). Agreement between the two representations was quantified for each (k, seed) pair using ARI, AMI, and NMI.

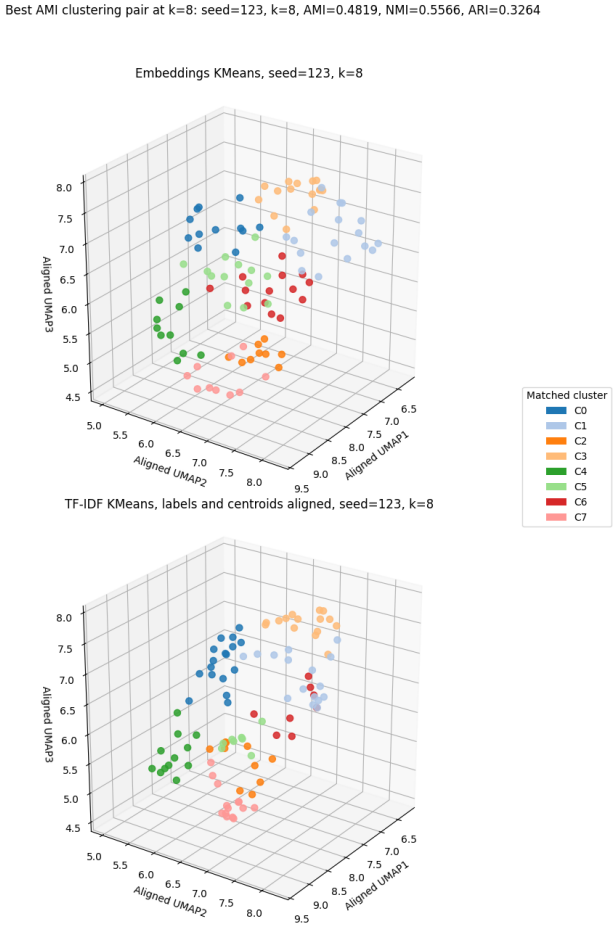


Figure 1: Aligned three-dimensional UMAP projections of embedding space and TF–IDF space K-means clustering (seed = 123, $k = 8$).

To allow visual comparison, TF–IDF cluster labels were matched to embedding cluster labels using Hungarian assignment on the contingency table, and the TF–IDF UMAP coordinates were aligned to the embedding UMAP coordinates via Procrustes rotation and scaling of cluster centroids.

Several clusters occupy comparable regions of the aligned space in both representations, while other items shift between neighbouring clusters depending on representation. The two representations show overlapping but not a one-to-one correspondence between all clusters.

The choice of $k = 8$ was not based on a single best metric. The highest consensus silhouette was obtained for $k = 6$ (0.5165), but this solution produced a large maximum family of 30 concepts and therefore merged several theoretically distinct regions of the corpus. $k = 10$ produced smaller and more balanced families, but at the cost of a more fine-grained partition that was less suitable

for a macro-level concept grouping. The $k = 8$ solution was therefore retained as an interpretability-oriented compromise: it preserved a reasonably high consensus silhouette (0.4824), avoided a dominant cluster (family sizes ranged from 5 to 16), and yielded eight substantively interpretable macro-families. Agreement generally increased from lower to higher values of k, although the improvement was modest and not uniform across all metrics. At $k = 8$, cross-representation ARI averaged approximately 0.28 and AMI approximately 0.44. The ARI indicates only partial agreement beyond chance and therefore limited robustness to the choice of textual representation. The $k = 8$ partition should consequently not be interpreted as a representation-invariant taxonomy: individual concepts may change family when lexical rather than contextual evidence is emphasized. Instead, the solution represents a consensus compromise between two complementary but only partly concordant views of the corpus. Its selection is supported by the consensus silhouette, cluster-size balance, and substantive interpretability, while the relatively low cross-representation ARI remains an important limitation.

4.2 Consensus Concept Grouping

To integrate lexical and semantic evidence, we averaged the representation-specific co-association matrices with equal weight to obtain a consensus similarity matrix C. The consensus distance matrix $D = 1 - C$ was then partitioned using average-linkage agglomerative clustering.

The $k = 8$ solution was selected as the primary concept grouping. Cluster sizes ranged from 5 to 16 items, indicating reasonably balanced families without any severely dominant or singleton groups.

4.3 Eight Macro-Families

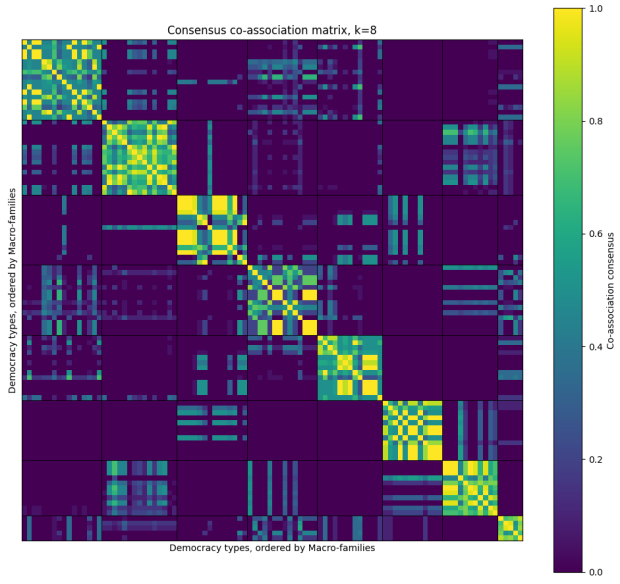


Figure 2: Macro-families of democracy concepts

Table 1 summarizes the eight macro-families derived from the consensus co-association concept grouping for $k = 8$. The resulting families vary in size from 5 to 16 democracy concepts, indicating that the corpus is distributed across several medium-sized conceptual regions rather than being dominated by one large cluster.

Mean internal consensus ranged from 0.467 to 0.750 across the eight families. The highest internal cohesion was observed in the Group-Based Democracy family (0.750), the Representative Democracy family (0.749), and the Direct and Deliberative Democracy family (0.720). These values suggest that the concepts grouped in these families repeatedly co-occurred across the consensus clustering runs and therefore form relatively compact regions in the combined TF–IDF and embedding-based representations.

The Electoral Democracy family (0.665) and the Local and Community-Based Democracy family (0.622) showed intermediate cohesion. These families appear to be internally coherent, but less compact than the strongest clusters. This is expected because electoral democracy includes both procedural and contested forms of democratic competition, while local and community-based democracy overlaps with federal, municipal, multilevel, and participatory arrangements.

The Digital Participatory Governance family (0.580) and the Constitutional Representative Democracy family (0.561) showed lower but still interpretable consensus. The digital family is connected by technological and participatory vocabulary, but it includes several distinct mechanisms, such as AI augmentation and algorithmic mediation. The constitutional-representative family is similarly broad, combining constitutional and representative elements.

The lowest internal consensus was found in the Participatory Governance family (0.467). This indicates that these democracies form a recognizable macro-family, but one that is internally heterogeneous. The lower value suggests that this family may require further subdivision in later versions of the grouping.

The representative concept in each row was selected computationally as the within-family medoid, that is, the concept with the highest average co-association with other members of the same family. It should not be interpreted as the theoretically most canonical example of the family. For instance, *Illiberal Democracy* and *Adversarial Democracy* appear as representative concepts because of their position in the consensus matrix, not because they are theoretically central examples of electoral or representative democracy.

Figure 2 provides a complementary visual diagnostic. The diagonal block structure indicates that the consensus solution recovers identifiable macro-families. At the same time, the off-diagonal associations show that several democracy concepts remain linked across family boundaries. This is expected in a corpus of democracy concepts, where many concepts combine multiple dimensions.

Figure 3 provides a hierarchical view of the consensus structure. Some families merge only at relatively high consensus distances, indicating stronger internal distinctiveness, whereas others join earlier, suggesting looser boundaries and greater conceptual overlap.

5 Conclusion

In this student paper, we presented a computational grouping of 100 democracy concepts using lexical TF–IDF and contextual sentence embeddings. By combining repeated K-means solutions in a consensus co-association matrix, we identified eight interpretable macro-families of democracy.

The main limitations concern heterogeneous source material, reliance on LLM-generated standardized descriptions, and restriction to English. LLM standardization and pretrained embedding

ID	Macro-family	Size	Consensus	Medoid concept
1	Constitutional Representative Democracy	16	0.561	Absolute Democracy
2	Electoral Democracy	15	0.665	Illiberal Democracy
3	Digital Participatory Governance	14	0.580	AI-Augmented Democracy
4	Participatory Governance	14	0.467	Cooperative Democracy
5	Local and Community-Based Democracy	13	0.622	Federal Democracy
6	Direct and Deliberative Democracy	12	0.720	Direct Democracy
7	Representative Democracy	11	0.749	Adversarial Democracy
8	Group-Based Democracy	5	0.750	Multicultural Democracy

Table 1: Eight macro-families of democracy concepts derived from the consensus co-association grouping ($k = 8$, $n = 100$).

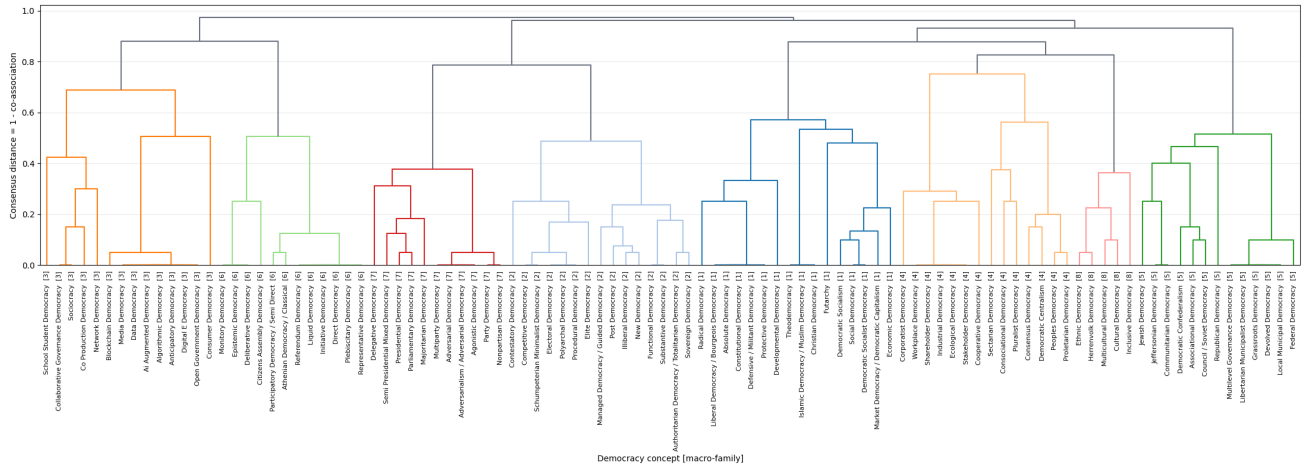


Figure 3: Dendrogram from the consensus matrix

models may introduce model-specific framing, semantic smoothing, unsupported details, and homogenized terminology. Because both textual representations are derived from these descriptions, such biases may propagate into the resulting cluster structure. The English-only corpus may also privilege Anglophone political-science terminology and underrepresent concepts rooted in other linguistic, legal, or historical traditions.

Future work should test robustness with additional embedding models, compare LLM-generated descriptions with expert-written definitions, extend the analysis to multilingual corpora, and externally validate the eight macro-families against independent expert-coded political-science taxonomies using measures such as ARI, AMI, and NMI.

Acknowledgements

The authors acknowledge the funding from the Slovenian Research and Innovation Agency (ARIS), Grant (PR-10495) and Basic core funding P2-0209.

References

[1] Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2, 3, 1–22.

[2] Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino, eds. 2011. *International Encyclopedia of Political Science*. SAGE, Thousand Oaks. ISBN: 978-1-4129-5963-6.

[3] Kenneth Benoit, Scott De Marchi, Conor Laver, Michael Laver, and Jinshuai Ma. 2025. Using large language models to analyze political texts through natural language understanding. *American Journal of Political Science*.

[4] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetrò, and Juan Carlos De Martin. 2019. The invisible power of fairness. how machine learning shapes democracy. In *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32*. Springer, 238–250.

[5] Robert A Dahl. 1998. *On democracy*. Yale university press.

[6] Shih-chan Dai. 2025. Conceptualizing digital democracy: a policy-oriented approach through computational text analysis. *Policy & Internet*, 17, 4, e70022.

[7] Ana Raquel de Melo Pinto. 2021. *Using Machine Learning to Measure Democracy and Economic Development, and the Causal Relationship Between the Two*. English. PhD thesis. Universidade do Porto, Porto, 114. ISBN: 9798383334102. https://www.proquest.com/dissertations-theses/using-machine-learning-measure-democracy-economic/docview/3085970586/se-2.

[8] Eleonore Fournier-Tombs and Michael K MacKenzie. 2021. Big data and democratic speech: predicting deliberative quality using machine learning techniques. *Methodological Innovations*, 14, 2, 20597991211010416.

[9] Tea Golob, Maruša Gorišek, and Matej Makarovič. 2023. Authoritarian and populist challenges to democracy correspond to a lack of economic, social, and cultural capitals. *Societies*, 13, 8, 181.

[10] Klaus Gründler and Tommy Krieger. 2016. Democracy and growth: evidence from a machine learning indicator. *European Journal of Political Economy*, 45, 85–107. On Institutions and Well Being. doi:https://doi.org/10.1016/j.ejpoleco.2016.05.005.

[11] Natali Helberger. 2019. On the democratic role of news recommenders. *Digital Journalism*, 7, 8, 993–1012.

[12] Ajay Kulkarni, Deri Chong, and Feras A Batarseh. 2020. Foundations of data imbalance and solutions for a data democracy. In *data democracy*. Elsevier, 83–106.

[13] George Thomas Kurian, ed. 2010. *The Encyclopedia of Political Science*. CQ Press, Washington, D.C. ISBN: 978-1-60871-243-4.

[14] Wolfgang Merkel. 2004. Embedded and defective democracies. *Democratization*, 11, 5, 33–58.

[15] Mustafa Sagir and Said Varlioglu. 2020. Explaining the relationship between internet and democracy in partly free countries using machine learning models. *CoRR*, abs/2004.05285. https://arxiv.org/abs/2004.05285 arXiv: 2004.05285.

[16] Gitte Stald. 2023. Mobile democracy: changing conditions for young danes' democratic information and participation. *Journalism and Media*, 4, 1, 272–288.

[17] Angel Torres-Toukoumidis, Isidro Marín Gutiérrez, Mónica Hinojosa Becerra, Tatiana León-Alberca, and Concha Pérez Curiel. 2023. Let's play democracy, exploratory analysis of political video games. *Societies*, 13, 2, 28.