Semantic Component Analysis: Introducing Multi-Topic Distributions to Clustering-Based Topic Modeling

Florian Eichin[1], **Carolin M. Schuster**[2], **Georg Groh**[2], **and Michael A. Hedderich**[1]
[1]LMU Munich and Munich Center for Machine Learning (MCML)
[2]Technical University of Munich
{feichin,hedderich}@cis.lmu.de, {carolin.schuster,groh}@in.tum.de

Abstract

Topic modeling is a key method in text analysis, but existing approaches fail to efficiently scale to large datasets or are limited by assuming one topic per document. Overcoming these limitations, we introduce Semantic Component Analysis (SCA), a topic modeling technique that discovers multiple topics per sample by introducing a decomposition step to the clustering-based topic modeling framework. We evaluate SCA on Twitter datasets in English, Hausa and Chinese. There, it achieves competetive coherence and diversity compared to BERTopic, while uncovering at least double the topics and maintaining a noise rate close to zero. We also find that SCA outperforms the LLM-based TopicGPT in scenarios with similar compute budgets. SCA thus provides an effective and efficient approach for topic modeling of large datasets.

1 Introduction

Topic modeling is an essential tool for automated text analysis in digital humanities, computational social science, and related disciplines (Grimmer et al., 2022; Rauchfleisch et al., 2023; Blei, 2012; Maurer and Nuernbergk, 2024). It aims to find high-level semantic patterns, the *topics*, in a corpus and then assign them to documents. The state-of-the-art (SOTA) for large-scale datasets, BERTopic (Grootendorst, 2022), assumes a single topic per document, which often does not hold even for short texts (Paul and Girju, 2010), as can be seen in Figure 1. Additionally, a high rate of samples is usually assigned to no topic, i.e. they have a high *noise rate* (Egger and Yu, 2022). LLM-based approaches like TopicGPT (Pham et al., 2024) perform very well, but due to their high computational cost, it is questionable how well they scale to large datasets under resource constraints.

To overcome these limitations, we propose **Semantic Component Analysis** (SCA) as a novel method to discover multiple topics per document. SCA is a topic modeling technique, that formalizes the discovery of *semantic components*, which intuitively are mathematical directions in the embedding space, each corresponding to a topic. It scales to large datasets while maintaining a noise rate close to zero and matches BERTopic in topic coherence and diversity. Based on clustering sequence embeddings as proposed by Angelov (2020), for each cluster, we introduce a decomposition step and iterate over the residual embeddings (see Figure 2). This decomposition enables us to represent a sample's content independently of the topics found in previous iterations and to find additional topics in the residuals.

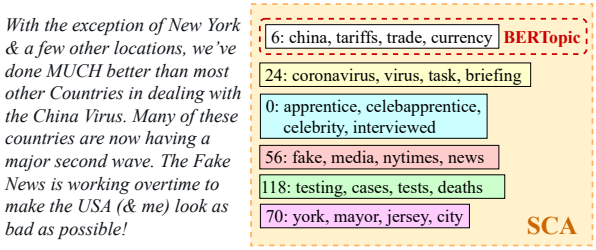

Figure 1: A Tweet from the Trump dataset (Brown, 2021) decomposed into its semantic components as defined by the TOP3 component activation scores of iteration 1 (components 6, 24, and 0) and subsequent iterations (components 56, 118, 70).

We evaluate our method on four datasets, one of Tweets by Donald Trump (Brown, 2021), another one of Tweets in Hausa (Muhammad et al., 2022), one of Tweets by Chinese-language news outlet accounts that we publish alongside this paper, and a collection of bills by the US congress (Adler and Wilkerson, 2012). We show that our method is able to match BERTopic in terms of topic coherence and diversity, while additionally providing a large number of topics missed by the SOTA approach. Furthermore, we show that given a comparable compute budget, SCA outperforms the LLM-based

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17748–17771
November 4-9, 2025

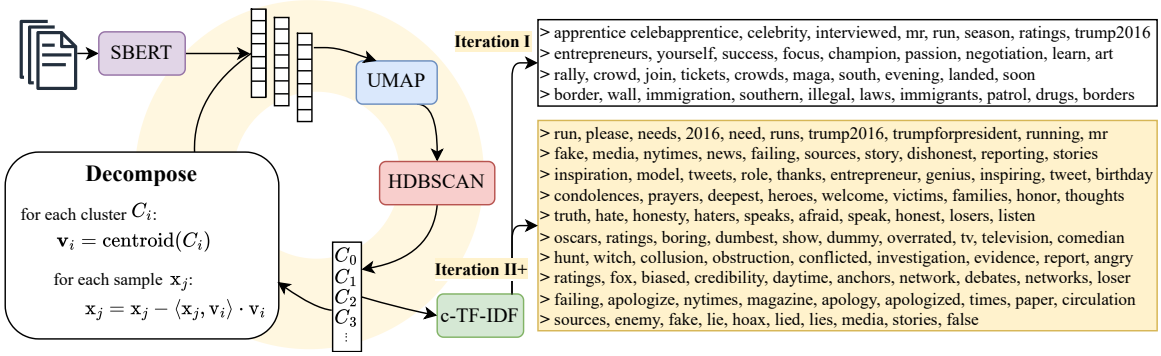

Figure 2: Illustration of our proposed method, that clusters residual embeddings obtained from linear decompositions of the embedding space. Top-Right: TOP4 components (by cluster size) of the first iteration equivalent to a BERTopic run. Bottom-right: components uncovered in subsequent iterations (see Appendix C.2 for more).

TopicGPT baseline on the Bills dataset, a benchmark proposed by Pham et al. (2024). Lastly, we also show that our method provides an interpretable transformation matching the base embedding in a variety of downstream tasks.

2 Related Work

Topic modeling's goal is to find semantic patterns in a corpus and to assign them to the documents. Latent Dirichlet Allocation (LDA, Blei et al. 2003; Pritchard et al. 2000) models multi-topic documents, works well on small datasets of long texts, and has been generalized to larger datasets (Hoffman et al., 2012). However, it performs poorly on short texts (Ushio et al., 2021; de Groot et al., 2022), is generally vulnerable to noisy data, and lacks domain adaption (Egger and Yu, 2022). Recent advancements have integrated neural networks (Srivastava and Sutton, 2017; Liu et al., 2015; Dieng et al., 2019; Qiang et al., 2016; Zhao et al., 2021), but they tend to struggle with stability and alignment with human-determined categories (Hoyle et al., 2022). Topic modeling based on LLMs (Pham et al., 2024; Doi et al., 2024) is promising, but their cost is a barrier for large-scale applications (see discussion in Appendix B.2) and their generated topics tend to be too generic or general (Li et al., 2025). Methods based on clustering embeddings stand out because of their efficiency, domain-adaption, and topic coherence on large corpora of short texts. Top2Vec (Angelov, 2020) clusters document embeddings generated by Doc2Vec (Le and Mikolov, 2014). There, UMAP (McInnes et al., 2020) reduces embedding dimensionality mitigating the "curse of dimensionality" (Reimers and Gurevych, 2021), and clustering is performed using HDBSCAN (Campello et al., 2015). BERTopic (Grootendorst, 2022) improves the Top2Vec framework by using SBERT embeddings (Reimers and Gurevych, 2019) and c-TF-IDF weighting for representing the topics. It is considered SOTA for large datasets of short texts (Egger and Yu, 2022). However, BERTopic does not provide a way to generalize to multi-topic documents. The cluster-hierarchy of HDBSCAN could be used to assign multiple topics per sample, but this approach is insufficient for finding patterns that are not subtopics of each other. Additionally, BERTopic's results usually suffer from a high noise rate (Egger and Yu, 2022), which can reach up to 74% (de Groot et al., 2022). In summary, the topic modeling field does not provide a solution for scenarios with large, noisy datasets of short texts containing multiple topics per sample, that matches the SOTA topic quality.

3 Methodology

SCA is a clustering pipeline that introduces a decomposition step to find unit vectors in the embedding space—which we call *semantic components*—that each correspond to a topic. For clustering, we employ the cosine distance in UMAP to map the embeddings to a 5-dimensional, Euclidean space on which HDBSCAN operates, mirroring the setup in BERTopic. Visualized in Figure 2, it works as follows:

0. **Embed** input sequences
1. **Reduce and Cluster** the embeddings $\mathbf{x}_j$
2. **Represent** component defined by cluster C_i through normalized centroid $\mathbf{v}_i \in \mathbb{R}^D$

$$\mathbf{v}_i = \frac{\mathbf{v}'_i}{||\mathbf{v}'_i||}, \quad \mathbf{v}'_i = \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j \qquad (1)$$

3. **Decompose** into linear components along the $\mathbf{v}_i$. For $\alpha_{i,j} = \frac{\langle \mathbf{x}_j, \mathbf{v}_i \rangle}{||\mathbf{x}_j||}$:

$$\mathbf{x}'_i = \mathbf{x}_i - \mu \mathbb{1}_{\alpha_{i,j} > \alpha} \langle \mathbf{x}_j, \mathbf{v}_i \rangle \mathbf{v}_i \quad (2)$$

4. **Repeat** steps 1-3 on residual embeddings
5. **Represent** components by c-TF-IDF
6. **Merge** topics whose token representations overlap by more than a threshold θ.

After each decomposition step in 3, we set $\mathbf{x}_j = \mathbf{x}'_j$ for all j and repeat it for the next cluster. The algorithm converges when the residuals are zero. We introduce other practical stopping criteria listed in Appendix A.3. Note that SCA has two hyperparameters: $\mu \in [0, 1]$, controlling the amount of the decomposition, and a threshold α, which is compared against $\alpha_{i,j}$, the *cosine similarity* between $\mathbf{x}_j$ and $\mathbf{v}_i$. See Appendix A.1 for a detailed discussion of α and μ. We assume that each cluster found represents a topic. To obtain an interpretable representation, we follow BERTopic and use c-TF-IDF to rank the tokens according to their importance for a cluster storing the TOP10 set of tokens for each cluster.

Merging step. To avoid duplicates arising in larger clusters spread across multiple dimensions, we introduce a merging step. First, for each pair of topics, we compute the token overlap $O(R_1, R_2) := \frac{1}{10}|R_1 \cap R_2|$ between their token representations R_1, R_2, where the division by 10 serves as a normalization factor for the number of tokens per representation. If $O(R_1, R_2) > \theta$ for some threshold $\theta \in [0, 1]$, we merge the two corresponding topics by assigning all the same topic ID and keeping the token and component representations of the topic that appeared earlier.

Topic assignment. SCA can assign topics in two ways: (i) Assign topics based on the clusterings in each iteration or (ii) obtain a full topic distribution by mapping the residual embeddings $\mathbf{x}'_j$ to the component space, where each dimension $\langle \mathbf{x}'_j, \mathbf{v}_i \rangle$ is interpreted as relevance of the topic represented by component $\mathbf{v}_i$ (see Appendix A.2).

SCA has the same modular approach as BERTopic, in which the embedding, dimensionality reduction, clustering, and representation method can be replaced.[1] Our specific choices for these modules, namely SBERT, UMAP, HDBSCAN, and c-Tf-IDF, make the first iteration of SCA equivalent to a BERTopic run, which allows for a direct comparison of the methods.

4 Data & Evaluation

Datasets. We evaluate our method on four datasets of different sizes and languages.[2] For an English-language use-case, we include all Tweets by Donald Trump until December 2024 (60K samples, Brown 2021). To cover a low-resource language, we include a set of 512K Tweets in Hausa collected by Muhammad et al. (2022). To evaluate on a high-resource non-English language, we collect and publish 1.5M Chinese-language Tweets from a list of news outlets (details in Appendix B.1). To evaluate in another domain against a ground truth, we use the Bills dataset (32K samples, Adler and Wilkerson 2012) used by Pham et al. (2024).

Evaluation metrics. We evaluate our method in terms of number of topics, noise rate, topic coherence, and diversity (see Appendix B.3 for definitions) and compare it to LDA and BERTopic on the datasets without ground truth. Furthermore, we perform a grid run over the hyperparameters to assess their influence. To check if the topics found are indeed beyond what is found by BERTopic, we compute the token overlap between the topics found by SCA and the topics found by BERTopic, as well as the sample overlap with the cluster hierarchy in HDBSCAN of BERTopic and the topics found by SCA (see Appendix B.3 for details). Evaluating on the Bills dataset, we compare SCA to TopicGPT (Pham et al., 2024) in terms of Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Purity (P_1) as proposed by them. Note that running TopicGPT requires a high budget of several thousand dollars for large datasets (see Discussion in Appendix B.2). We thus also compare to a version of TopicGPT with similar compute budget as SCA.

MTEB evaluation. Finally, we evaluate the SCA's transformation into the component space (see Appendix A.2 for details) on the MTEB benchmark (Muennighoff et al., 2023), which consists of various downstream tasks, including sequence classification, summarization and retrieval. We do this to quantify the faithfulness of the representation provided by SCA-components and compare against the base embedding model and a PCA.

[1] see https://maartengr.github.io/BERTopic/

[2] https://github.com/mainlp/semantic_components

	Trump			Hausa			Chinese News		
	BT	SCA	LDA	BT	SCA	LDA	BT	SCA	LDA
#Topics ↑	55	182	182	102	331	331	212	594	594
Noise R. ↓	0.275	0.000	0.003	0.574	0.001	0.05	0.429	0.000	0.802
NPMI ↑	-0.135	-0.168	-0.139	-0.028	-0.061	-0.156	0.186	0.120	-0.149
CV ↑	0.370	0.360	0.334	0.431	0.400	0.393	0.683	0.595	0.198
Topic D. ↑	0.947	0.703	0.698	0.909	0.804	0.444	0.929	0.803	0.003

	Bills					
	BT	SCA	LDA*	TGPT	TGPT	TGPT*
#Topics	79	79	79	23	15	79
Param.	0.3B	0.3B	n/a	1B	8B	unk.
P_1 ↑	0.42	0.42	0.39	0.13	0.16	0.57
ARI ↑	0.16	0.18	0.21	0.00	0.00	0.42
NMI ↑	0.38	0.38	0.47	0.01	0.09	0.52

Table 1: Topic modeling statistics. BERTopic and SCA use the same hyperparameters. Arrows (↑↓) indicate optimal metric directions (see Appendix B.3). Results marked with * are copied from Pham et al. (2024), who report using `GPT-3.5` through the OpenAI API. Additional runs of TopicGPT use `Llama-3.2-1B-Instruct` and `Llama-3.2-8B-Instruct` (Grattafiori, 2024). We follow their evaluation to compare SCA to TopicGPT, which is restricted to the TOP79 topics to ensure comparability with the LLM based approach. Since the two TopicGPT baselines do not find this many topics, we report the results for the respective smaller numbers of topics.

5 Results

Statistical results. We provide statistical results in Table 1. For all datasets, SCA finds a large number of topics, with the Trump dataset yielding 182 topics in total after merging, the Hausa dataset 331 and the Chinese News dataset 594, surpassing the 55, 102, and 212 respective topics found by BERTopic. These topics relate to patterns distinct from the topics found by BERTopic, which is evident from the low average maximum token overlap scores of 1.85, 1.32 and 1.39 on the three datasets respectively. On the Trump dataset, only 8 topics with a token overlap of higher than 0.5 are merged, on the Hausa dataset it is 9, and for the Chinese News dataset 19. The average maximum sample overlaps with any of BERTopic's hierarchical subtopics are small and between 0.077 and 0.093 on all datasets. This indicates that the additional topics are not sub- or supertopics of the topics found by BERTopic.

The noise rate is close to zero for all datasets, i.e. almost all samples are assigned at least one topic. The topic coherence scores are high for all datasets, with the Chinese dataset showing the highest NPMI coherence of 0.120 and the Trump dataset showing the lowest of -0.168, all matching the baseline. The topic diversity is high for all datasets, between 0.7 and 0.8, but about 10-15% lower than in the baseline.

In sum, this shows that the quality of the topics found matches that of BERTopic while providing a more nuanced view of the data through the large number of extra topics found and the low noise rate. Ablating the hyperparameters on the Trump dataset, we find that for $\alpha > 0.7$ and $\mu < 0.4$, SCA reliably finds at least double the number of topics found by BERTopic, usually ranging up to four times.

Exploration of additional topics. The additional topics found intuitively match the expected distribution of the datasets (full TOP10 topic tables in Appendix C). For example, in the Trump dataset, beyond the topics found by BERTopic (about Trump himself, entrepreneurship, election campaigning), we uncover topics about "fake news", inspiration and role models, paying condolences, and the "witch hunt" narrative among others (see Appendix C.2) underlining SCA's ability to find important topics not captured by the baseline.

This view is reinforced by the topic distributions of single samples, as shown in Figure 1. Similarly, in the Chinese dataset, the additional topics relate to important motives in Chinese news reporting, such as protests, family, the trade war with the US, and currency (see Appendix C.4). In the Hausa dataset, the additional insights include topics about poverty, courts and law, and work (see Appendix C.5). In this low-resource scenario, some topics relate to semantically more narrow concepts, which might be caused by lack of expressive power in the embeddings.

Tweet 1: *ALWAYS BORROW MONEY FROM A PESSIMIST BECAUSE HE WILL NEVER EXPECT IT TO BE PAID BACK!*	
TOP1: entrepreneurs, yourself, success, think, focus, champion, learn, art, passion, touch	
TOP2: (2+ It.) charity, money, ads, million, bankrupt, bankruptcy, website, ad, raised, donate	
TOP3: impeachment, collusion, democrats, investigation, hoax, whistleblower, evidence, hunt, dems, witch	
Tweet 2: *"Leverage: don't make deals without it." – The Art of the Deal*	
TOP1: (1st It.) entrepreneurs, yourself, success, think, focus, champion, learn, art, passion, touch	
TOP2: (2+ It.) art, deal, negotiation, desperate, seem, possibly, deals, tip, worst, theartofthedeal	
TOP3: (1st It.) impeachment, collusion, democrats, investigation, hoax, whistleblower, evidence, hunt, dems, witch	
Tweet 3: *95% Approval Rating in the Republican Party. Thank you!*	
TOP1: (2+ It.) approval, rating, overall, party, republican, 52, 51, 53, 93, 50	
TOP2: (1st It.) impeachment, collusion, democrats, investigation, hoax, whistleblower, evidence, hunt, dems, witch	
TOP3: (1st It.) poll, leads, surges, lead, bush, surging, postdebate, tops, 28, favorability	
Tweet 4: *Our govt is so pathetic that some of the billions being wasted in Afghanistan are ending up with terrorists*	
TOP1: (1st It.) impeachment, collusion, democrats, investigation, hoax, whistleblower, evidence, hunt, dems, witch	
TOP2: (1st It.) isis, oil, fighters, caliphate, attack, chemical, rebels, terrorist, troops, terrorists	
TOP3: (2+ It.) leadership, leader, worst, incompetent, need, politicians, needs, stamina, leaders, lead	

Table 2: Samples from the Trump dataset for qualitative error analysis. Note that the topics are taken from a different run than the one described in the Appendix, but with the same hyperparameters. **Top two samples** were selected at random from the **entrepreneurship topic** and the **bottom two** were selected from the noise cluster of iteration 1, i.e. would have been classified as noise by BERTopic. We also highlight the largest topic colored in **red**. Topics shown are the TOP3 topics assigned by the scores of the transformation into the component space and annotated on whether they were discovered in the **first iteration** or **subsequent iterations**.

Error analysis. While the discovered topics are generally high quality, we observe several recurring error patterns. In all datasets, SCA produces a power-law distribution of topic sizes, with a few large and many smaller topics. The largest topics tend to be generic and over-assigned, as seen in Table 2. This is likely due to HDBSCAN, which tends to connect clusters in dense areas of the embedding space. However, in later iterations SCA refines the broad entrepreneurship topic into more specific topics such as money and negotiation assigned to Tweets 1 and 2 respectively. We also sample two Tweets from the iteration-1 noise cluster—i.e., Tweets BERTopic failed to assign—and find that SCA correctly identifies topics like approval ratings, polling, terrorism, and leadership criticism. We also note that the largest topic, which seems to encompass patterns such as impeachment and the Democratic party, is wrongly predicted across all samples. Lastly, while most TOP10 token lists support interpretation, some topics, especially the generic ones, are hard to interpret.

Comparison with TopicGPT. Comparing against TopicGPT, we find that both SCA and BERTopic achieve similar performance in terms of purity, NMI and ARI outperforming the TopicGPT baselines based on the Llama models (i.e. with similar compute budgets). The good results reported by Pham et al. (2024) are based on `GPT-3.5`, suggesting that compute-intensive and expensive models are necessary for TopicGPT to perform well; which might not be available in application settings like social science.

MTEB evaluation of transformed embeddings. Finally, the results of the topic distribution provided by SCA can match the performance of PCA and the base embedding model in the MTEB benchmark (see Appendix C.6), indicating the method's faithfulness in representing the information of the samples in the component space. This suggests that the transformed embeddings provided by SCA can be seen as an interpretable alternative to the base embedding model at no cost in performance.

6 Conclusion

We introduce Semantic Component Analysis as a novel topic modeling method that can discover multiple topics and scale to large datasets. We show that the topics found by SCA indeed provide new, more nuanced insights into the data while maintaining a noise rate close to zero and matching the large-scale SOTA in terms of topic quality. Furthermore, we show that the method is able to generalize to multiple languages. However, future work is required to address the problem of generic, overpredicted clusters and the interpretability of the topic representations, perhaps by downweighing larger clusters and exploring other forms of cluster representations respectively. Taken together, we believe that SCA provides a substantial improvement in large-scale short text analysis.

Limitations

While SCA provides a novel perspective on topic modeling by providing an efficient way to discover multiple topics per sample and thus a more detailed explanation of the pattern distribution in a dataset, it has its limitations.

As is common with pipelines that rely on clustering embeddings, SCA inherits several potential limitations inherent to this class of algorithms. Firstly, using a sequence of modular steps might introduce errors that could be propagated and thus amplified in subsequent steps. For example, if the clustering step fails to find meaningful clusters, the topics found will likely be of low quality. To an extent, this problem can be mitigated by carefully chosing the modules and hyperparameters in accordance with the data at hand and performing multiple runs to test robustness. However, the risk of error propagation in clustering-based topic modelling remains and should be investigated further in future work.

Furthermore, the similarity modeled by the input embeddings is crucial for the quality of the topics found. If the embedding model does not capture the semantic similarity in a way that is suitable for the data at hand, the topics found will likely be of low quality. This is especially relevant in low-resource languages or domains, where high-quality embedding models are often not available. While SCA can be used with any embedding model, it is important to carefully select and validate the base embedding model to ensure as fair and unbiased representations as possible.

Besides, SCA is not able to capture non-linear, multi-dimensional components in the data, as it is based on a linear decomposition of the embedding space. In other words, SCA is based in the linear representation hypothesis (Park et al., 2024). To some extent, the iterative procedure of SCA is able to mitigate this problem: Because each cluster is represented by the centroid, the decomposition might retain a part of the cluster-information in the residual embeddings in cases, when one linear unit vector is not able to capture the full complexity in terms of dimension or linearity. In such cases, similar topics might be found in subsequent iterations, which can be merged by the overlap threshold θ to avoid duplicate topics.

Furthermore, while we show SCA's applicability to a range of datasets and languages, we restrict our investigation to short texts, more precisely Tweets. We argue that problems with longer documents can partially be mitigated by splitting them into smaller parts, such as paragraph level, and apply SCA to the resulting dataset. Since SCA scales to large numbers of samples even on small systems, this should be feasible for most applications. Adding to this, our evaluation does not include a user study to assess the interpretability of the topics found. While we provide qualitative examples and find that the topics discovered match the expected distribution in the datasets, a user study could provide additional insights into the practical applicability of SCA.

As is common with topic modeling, the method's token-based representation of topics has its limitations, too, as it might not able to capture the full complexity of the topics found. While additional methods like LLM summarizations and Medoid representations can provide additional insights, a core problem with more expressive approaches are uncertain faithfulness, a lack of good evaluation metrics as well as increased computational cost.

Ethics Statement

There are several risks and ethical considerations associated with the use of SCA.

Firstly, there is a risk of misinterpretation or misrepresentation of the discovered topics. The interpretation of these topics relies on the assumption that they represent meaningful and coherent topics or themes in the data. However, there is always a possibility of misinterpreting or misrepresenting the true meaning of these topics, leading to incorrect conclusions or biased interpretations. Even earlier in the pipeline, there is a risk of algorithmic bias in the clustering and representation process. Especially the base embedding model can introduce biases based on its training data and model architecture that SCA can inherit. It is important to carefully select and validate the base embedding model to ensure as fair and unbiased representations as possible.

Secondly, there is a risk of privacy infringement when working with sensitive or personal data. SCA relies on analyzing and clustering text data, which may contain sensitive information about individuals or groups. It is important to handle and protect this data in accordance with privacy regulations and ethical guidelines and to ensure that the topics extracted from the data do not reveal sensitive or personally identifiable information to unauthorized parties. The usage of the data in our evaluation

does not expose any individual to such risks, as the data was either publicly available to begin with (the Hausa Dataset and Trump Archive) or does not contain any data produced by private Twitter accounts, as the Chinese News Data we scrape only consists of Tweets by the official accounts of the news outlets we include.

Thirdly, the nature of online social media data, such as Tweets, can introduce additional risks and challenges. Social media data can be subject to manipulation or misinformation, and contain offensive or harmful content. SCA may inadvertently amplify or propagate such content if not properly filtered or evaluated. Additionally, using the insights from SCA might inadvertently lead to amplifying or reinforcing existing biases or stereotypes. It is important to consider the potential impact of the data and the extracted topics on individuals, communities, and society as a whole.

Overall, it is crucial to approach the use of SCA with caution, considering the potential risks and ethical implications associated with its application.

Use of AI assistants. The authors acknowledge the use of ChatGPT for correcting grammatical errors and improving the overall readability of the paper. Furthermore, GitHub Copilot was used to assist in writing the code. The authors have reviewed and edited the output of these AI assistants to ensure accuracy and clarity.

References

E. Scott Adler and John D. Wilkerson. 2012. *Congress and the Politics of Problem Solving*. Cambridge University Press, Cambridge [England] ; New York.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics. *Preprint*, arXiv:2008.09470.

David M. Blei. 2012. Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, Vol. 2,(No. 1 Winter 2012).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40.

Brendan Brown. 2021. The Trump Twitter Archive. www.thetrumparchive.com.

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):5:1–5:51.

Muriël de Groot, Mohammad Aliannejadi, and Marcel R. Haas. 2022. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. https://arxiv.org/abs/2212.08459v1.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. Topic Modeling in Embedding Spaces. *Preprint*, arXiv:1907.04907.

Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. 2024. Topic Modeling for Short Texts with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 21–33, Bangkok, Thailand. Association for Computational Linguistics.

Roman Egger and Joanne Yu. 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. *Preprint*, arXiv:2007.01852.

Aaron et al. Grattafiori. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *Preprint*, arXiv:2203.05794.

Matt Hoffman, David M. Blei, Chong Wang, and John Paisley. 2012. Stochastic Variational Inference. https://arxiv.org/abs/1206.7051v3.

Alexander Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are Neural Topic Models Broken? *Preprint*, arXiv:2210.16162.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196. PMLR.

Zongxia Li, Lorena Calvo-Bartolomé, Alexander Hoyle, Paiheng Xu, Alden Dima, Juan Francisco Fung, and Jordan Boyd-Graber. 2025. Large Language Models Struggle to Describe the Haystack without Human Help: Human-in-the-loop Evaluation of Topic Models. *Preprint*, arXiv:2502.14748.

Yang Liu, Zhizuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical Word Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume Vol. 29 No. 1 (2015).

Peter Maurer and Christian Nuernbergk. 2024. No watchdogs on Twitter: Topics and frames in political journalists' tweets about the coronavirus pandemic. *Journalism*, page 14648849241266722.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Preprint*, arXiv:1802.03426.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. *Preprint*, arXiv:2210.07316.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. *Preprint*, arXiv:2201.08277.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. *Preprint*, arXiv:2311.03658.

Michael Paul and Roxana Girju. 2010. A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):545–550.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A Prompt-based Topic Modeling Framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

J. K. Pritchard, M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. 2016. Topic Modeling over Short Texts by Incorporating Word Embeddings. *Preprint*, arXiv:1609.08496.

Adrian Rauchfleisch, Tzu-Hsuan Tseng, Jo-Ju Kao, and Yi-Ting Liu. 2023. Taiwan's Public Discourse About Disinformation: The Role of Journalism, Academia, and Politics. *Journalism Practice*, 17(10):2197–2217.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Preprint*, arXiv:1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Preprint*, arXiv:2004.09813.

Nils Reimers and Iryna Gurevych. 2021. The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes. *Preprint*, arXiv:2012.14210.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. WSDM '15, pages 399–408, New York, NY, USA. Association for Computing Machinery.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. *Preprint*, arXiv:1703.01488.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021. Distilling Relation Embeddings from Pre-trained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic Modelling Meets Deep Neural Networks: A Survey. *Preprint*, arXiv:2103.00498.

Appendices

A Methodology Appendix

A.1 Hyperparameter Discussion

Our method involves two hyperparameters that influence the decomposition step, α and μ. Intuitively, both can be used to influence the level of superposition allowed in the decomposition, i.e. the number of semantic components that can occupy linearly dependent parts of the embedding space and thus the the level of detail and the number of discovered topics. This can be necessary in cases of large data where the number of topics assumed is higher than the dimension of the embedding space.

In the simplest case, when no superposition is assumed, we set $\mu = 1$ and $\alpha = 0$. We illustrate this version of the algorithm in Figure 2. Setting $\mu < 1$ is the first strategy to allow for superposition. This way, each decomposition step will only remove a fraction of the feature from the activation, allowing subsequent clustering passes to discover more, linearly dependent features. The second strategy is to set $\alpha > 0$, which introduces conditional decomposition. Here, a feature is only decomposed from an activation if the cosine similarity between the activation and the feature is above the threshold α.

Furthermore, we introduce the merging threshold θ to merge components that are similar in terms of their token representation. This is necessary to avoid duplicate components that arise for large clusters spread across a non-linear or multi-dimensional part of the embedding space. We note that the choice of θ is somewhat arbitrary and should be chosen based on the specific use case. In our evaluation, we set $\theta = 0.6$ for the Trump dataset, which resulted in 11 components being merged. For the Chinese and Hausa datasets, we did not employ component merging.

A.2 SCA as a Transformation

The decomposition step in SCA is a linear transformation of the embeddings into a space where each dimension corresponds to the score along one semantic component. The decomposition is based on the cosine similarity between the embedding and the centroid of the cluster, which is the semantic component representing the connected topic. For an embedding $\mathbf{x}$ of a sample s, e define the i-th residual in an inductive way:

$$\mathbf{x}'_0 = \mathbf{x}, \mathbf{x}'_{i+1} = \mathbf{x}'_i - \mu \mathbb{1}_{\alpha_{i,i} > \alpha} \langle \mathbf{x}_i, \mathbf{v}_i \rangle \mathbf{v}_i \tag{3}$$

where $\mathbf{v}_i$ is the i-th component and $\alpha_{i,i} = \frac{\langle \mathbf{x}_i, \mathbf{v}_i \rangle}{||\mathbf{x}_i||}$. The activation of the sample $\mathbf{x}$ along the j-th is thus defined as the dot product of the residual and the component:

$$a_j = \mu \mathbb{1}_{\alpha_{i,i} > \alpha} \langle \mathbf{x}'_j, \mathbf{v}_j \rangle \tag{4}$$

From this, we can assign additional components to the samples by computing the activation along the components and applying a threshold per iteration to choose the relevant ones. This ensures, that the decomposition is able to capture multiple topics per sample even beyond the cluster assignments.

A.3 Stopping Criteria

Without additional stopping criteria, the algorithm would spend a large number of iterations decomposing residual embeddings that are already close to zero and lack any meaningful information. To avoid this, we introduce a number of stopping criteria that can we use to terminate the algorithm. These criteria are as follows:

- **F**: Terminate after a fixed number of iterations I
- **NC-S**: Number of new clusters found in the last S iterations is lower than a threshold T
- **RN**: Matrix 2-norm of residual embeddings lower than a threshold M

For our evaluation, we set $I = 10$, $S = 2$, $T = 5$, and $M = 0.01$ for all use cases.

B Data & Evaluation Appendix

B.1 Implementation and Preprocessing

The datasets we use for evaluating our method have been obtained from different sources. The Tweets made by Donald Trump are available in the *Trump Twitter Archive* (Brown, 2021), from where we download the dataset directly as `.csv` which resulted in 60,053 samples as of 10th of December 2023. The Chinese News Tweets have been scraped from a list of Chinese news outlets, which we compiled by two criteria: (1) The majority of the Tweets uses simplified or traditional Chinese characters and (2) the news outlet is based outside of the borders of the People's Republic of China. The list is far from complete, but we stopped adding new accounts when a threshold of 1M Tweets was reached. The accounts included are: "voachinese", "chinesewsj", "abcchinese", "reuterschina", "dw_chinese", "rfi_tradcn", "asiafinance", "initiumnews", "cdtchinese", "kbschinese", "rijingzhongwen", "weiquanwang", "zaobaosg", "rfa_chinese", "rfi_cn", "rcizhongwen", "kbschines", and "chosunchinese".

From this list, we scraped all the Tweets available through the Twitter API which resulted in 1,463,150 samples until the closure of our academic API access in April 2023. The Hausa Tweets have been collected for sentiment classification by Muhammad et al. (2022) and, at the time of acquiring the dataset, amounted to 511,523 samples in Hausa. The current version, of which our sample is a subset of, with slightly more samples (516,523) is available on GitHub.[3]. The hyperparameters chosen for each run of the are listed in Figure 3.

For computing the embeddings of the datasets, we used one NVidia A100 GPU, which processed all of the data in less than an hour. The computations for SCA were entirely CPU-based (AMD EPYC 7662) with an allocation of 8 CPU cores and 64 GB of RAM for the Hausa and Chinese News datasets and 16GB for the Trump dataset. On that configuration, the total runtime for SCA (excluding the time to compute the embeddings) on the Trump dataset is 9 minutes, on the Chinese News dataset 345 minutes and on the Hausa dataset 91 minutes. The main bottleneck for the computations is UMAP, of which there exists a GPU implementation through cuML[4]. We tested this version and can indeed confirm that the runtimes are much faster using this implementation. However, we found the results to be inconsistent with the CPU version, which is why we decided to stick to the CPU version for the evaluation.

Before embedding the documents, we preprocess the data by removing URLs, mentions, hashtags, emojis, and special characters. The sentence transformer models we use, namely `paraphrase-multilingual-mpnet-base-v2` (Reimers and Gurevych, 2020) for the Trump and Chinese News Tweets and `LaBSE` (Feng et al., 2022), are trained on a large variety of domains and languages. We use the default settings and tokenizers for the models, which are able to handle sequences of up to 512 tokens. We use the `umap-learn` implementation of UMAP (McInnes et al., 2020) with the default settings for dimensionality reduction. For clustering, we use the `hdbscan` implementation of HDBSCAN (Campello et al., 2015) including `min_cluster_size` and `min_samples` as hyperparameters of SCA. We use the `OCTIS` implementation of the coherence measures (Terragni et al., 2021).

B.2 TopicGPT Comparison

We compare SCA against the TopicGPT method (Pham et al., 2024), which uses a large language model to generate topics from the embeddings. To ensure a fair comparison in the resource-restricted scenario SCA is targeted, we aim to evaluate TopicGPT with backbone models of similar size as the embedding model we use for our experiments, i.e. 278M paramaters for `paraphrase-multilingual-mpnet-base-v2`. We use the instruction tuned `Llama-3.2` model (Grattafiori, 2024) with 1B and 8B parameters, namely `Llama-3.2-1B-Instruct` and `Llama-3.2-8B-Instruct` and run TopicGPT in the standard configuration without changing any hyperparameters or steps of the procedure as defined in their GitHub repository.[5]

We note that the authors of TopicGPT report their results using the `GPT-3.5` model through the OpenAI API, which is not available for local installations. We include these results in Table 1 for comparison.

[3] https://github.com/hausanlp/NaijaSenti/blob/main/sections/unlabeled_twitter_corpus.md
[4] https://docs.rapids.ai/api/cuml/stable/execution_device_interoperability/
[5] See https://github.com/chtmp223/topicGPT

Furthermore, note the extensive compute budget necessary to run TopicGPT on the Bills dataset, which has only 32,661 samples. There, the authors report a total cost of $88 US dollars, of which $30 is for the topic generation, $10 for topic refinement and $48 for the for the assignment and correction step using the OpenAI API. Assuming the former two do not scale with dataset size, the cost for the assignment and correction step would be $1.47 per 1000 samples. Extrapolating this cost to the Trump dataset, which has 60,053 samples, would result in a total cost of $88.57 for the assignment and correction step alone. For the Hausa dataset, which has 511,523 samples, the cost would be $751.75 and for the Chinese News dataset, which has 1,463,150 samples, the cost would be $2150.30. We argue that this kind of cost is not feasible for most applications, especially in the low-resource setting SCA is targeted at.

B.3 Evaluation Metrics

For each run, we calculate a number of statistics to evaluate the performance of the method. We provide the results for the three datasets in the following tables. The statistics are as follows:

- **Noise rate**: The fraction of documents that are not assigned to any cluster in the first iteration (*1st*) and over all iterations.

$$\mathcal{L}_{noise} = \frac{|\{c \in C | c = -1\}|}{|C|} \tag{5}$$

- **Number of clusters**: The number of unique clusters in the clustering result.

$$\mathcal{L}_{num} = |C| \tag{6}$$

- **TOP10 topic diversity**: As defined in (Terragni et al., 2021), it is the number of unique TOP10 representative tokens (as computed by c-TF-IDF weighting) divided by the total number of representative tokens.

Additionaly, we define the sample overlap between sets $C_1 = \{s_1, s_2, ..., s_n\}$ and $C_2 = \{s'_1, s'_2, ..., s'_m\}$ as the fraction of elements that are in both clusters.

$$o(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \tag{7}$$

from this, we define additional metrics to quantify the novelty of the topics uncovered by SCA compared to those found by BERTopic:

- the **average maximum sample overlap** of the clusters found after the first iteration with any of the clusters in the complete HDBSCAN-hierarchy of the first iteration. For $C_I = \{C_1, C_2, ..., C_M\}$ the set of possible clusters in the first iteration as defined by the HDBSCAN-hierarchy and $C_+ = \{C_1^+, C_2^+, ..., C_N^+\}$ the set of clusters found by SCA after the first iteratiom, we define the average maximum sample overlap as:

$$\mathcal{L}_{ASO} = \frac{1}{N} \sum_{i=1}^{N} \max_{j=1,...,M} o(C_i^+, C_j) \tag{8}$$

  In other words, this metric tries to understand to what extend the additional components found by SCA where present in the topic-hierarchy of BERTopic. High values close to one would indicate, that most compnents found by SCA are sub- or supertopics of the topics found by BERTopic, low values close to zero would indicate that the components found by SCA are not present in the topic-hierarchy of BERTopic and hence provide additional insights.

- In an analog way, for the component token representations of the first iteratiom $R_I = \{R_1, R_2, ..., R_M\}$ and the representations of following iterations $R_+ = \{R_1^+, R_2^+, ..., R_N^+\}$, we define the **average maximum token overlap** as:

$$\mathcal{L}_{ATO} = \frac{1}{N}\sum_{i=1}^{N} \max_{j=1,\dots,M} o(R_i^+, R_j) \tag{9}$$

The coherence scores are calculated over the token representations of each topic using the respective `OCTIS` implementations. For some confirmation measure $m : V \times V \to \mathbb{R}$ that maps two words to a real number, the topic coherence score is defined as

$$\mathrm{TC}_m = \frac{1}{T}\sum_{i=1}^{T} \frac{1}{\binom{m}{2}} \sum_{r=2}^{K} \sum s = 1r - 1m(w_i^r, w_i^s) \tag{10}$$

- **NPMI Coherence** (Aletras and Stevenson, 2013) propose to use the NPMI measure (Bouma, 2009) due to it being closer to human topic ratings than previous versions.

$$m(w_i^r, w_i^s) = NPMI(w_i^r, w_i^s) = \frac{\log_2 \frac{p(w_i^r, w_i^s)+\epsilon}{p(w_i^r)p(w_i^s)}}{-\log_2(p(w_i^r, w_i^s)+\epsilon)} \tag{11}$$

- **CV Coherence** (Röder et al., 2015) propose to use the CV measure, which scales the coherence terms of NPMI by a power of a parameter γ.

$$m(w_i^r, w_i^s) = \mathrm{CV}(w_i^r, w_i^s) = \mathrm{NPMI}^{\gamma}(w_i^r, w_i^s) \tag{12}$$

C Extended Results

C.1 Run Statistics and Grid Run

Statistic	D.T.	Hau.	News
α	0.20	0.10	0.10
μ	0.95	1.00	1.00
`min_cluster_size`	100	300	300
`min_samples`	50	300	300
Overlap Threshold θ	0.5	0.5	0.5
No. of Components (1st)	55	102	212
No. of Components	182	331	594
No. of Clusters	190	340	613
Noise Rate (1st)	0.275	0.574	0.429
Noise Rate	0.000	0.001	0.000
Avg. Maximum Sample Overlap	0.079	0.093	0.077
Avg. Maximum Token Overlap	1.387	1.852	1.318
NPMI Coherence	-0.168	-0.061	0.120
CV Coherence	0.360	0.400	0.595
Topic Diversity	0.703	0.804	0.803
NPMI Coherence (1st)	-0.135	-0.028	0.186
CV Coherence (1st)	0.370	0.431	0.683
Topic Div. (1st)	0.947	0.909	0.929

Figure 3: Run statistics for the three datasets. (1st) denotes the statistic calculated only over the clusters/components of the first iteration of SCA, which is equivalent to a BERTopic run with the same hyperparameters. Table 1 provides the same values, but we keep this table here for reference mirroring the layout of our original submission.

No. of components

μ/α	0.0	0.05	0.1	0.2	0.3	0.4	0.5
0.5	558	416	384	218	195	577	945
0.6	582	421	410	285	208	303	680
0.7	491	444	364	305	107	107	177
0.8	363	341	358	317	281	252	108
0.9	466	365	264	319	292	246	371
0.95	336	353	347	336	231	250	302
1.0	333	358	353	336	287	289	261

Topic diversity

μ/α	0.0	0.05	0.1	0.2	0.3	0.4	0.5
0.5	0.539	0.631	0.635	0.728	0.710	0.259	0.096
0.6	0.530	0.651	0.663	0.690	0.726	0.553	0.213
0.7	0.611	0.643	0.697	0.699	0.859	0.857	0.709
0.8	0.702	0.702	0.706	0.725	0.706	0.689	0.853
0.9	0.655	0.713	0.749	0.714	0.713	0.710	0.589
0.95	0.728	0.710	0.724	0.724	0.746	0.705	0.638
1.0	0.725	0.719	0.716	0.717	0.720	0.682	0.658

Noise rate

μ/α	0.0	0.05	0.1	0.2	0.3	0.4	0.5
0.5	0.003	0.003	0.004	0.000	0.000	0.221	0.398
0.6	0.001	0.003	0.010	0.012	0.003	0.001	0.240
0.7	0.002	0.012	0.001	0.004	0.000	0.000	0.000
0.8	0.000	0.002	0.001	0.001	0.000	0.016	0.001
0.9	0.000	0.000	0.005	0.004	0.000	0.000	0.000
0.95	0.000	0.006	0.001	0.000	0.000	0.000	0.001
1.0	0.002	0.003	0.003	0.001	0.000	0.000	0.003

NPMI Coherence

μ/α	0.0	0.05	0.1	0.2	0.3	0.4	0.5
0.5	-0.192	-0.191	-0.187	-0.162	-0.160	-0.169	-0.196
0.6	-0.202	-0.198	-0.198	-0.178	-0.161	-0.168	-0.173
0.7	-0.202	-0.203	-0.195	-0.186	-0.153	-0.153	-0.150
0.8	-0.191	-0.197	-0.197	-0.190	-0.182	-0.177	-0.152
0.9	-0.206	-0.198	-0.194	-0.194	-0.187	-0.180	-0.197
0.95	-0.199	-0.199	-0.198	-0.197	-0.180	-0.179	-0.188
1.0	-0.200	-0.201	-0.200	-0.201	-0.191	-0.193	-0.181

CV Coherence

μ/α	0.0	0.05	0.1	0.2	0.3	0.4	0.5
0.5	0.356	0.362	0.362	0.368	0.368	0.364	0.353
0.6	0.355	0.361	0.357	0.363	0.366	0.372	0.359
0.7	0.361	0.363	0.358	0.365	0.374	0.374	0.374
0.8	0.359	0.365	0.362	0.365	0.361	0.366	0.375
0.9	0.359	0.358	0.367	0.363	0.361	0.363	0.359
0.95	0.365	0.364	0.362	0.359	0.362	0.367	0.360
1.0	0.369	0.361	0.362	0.365	0.360	0.365	0.365

Figure 4: Grid run results for different hyperparameters. All configurations were run with $\theta = 1.0$ to avoid occluding "bad" results by merging many components.

C.2 Trump Dataset: SCA Topics

ID	N	**c-TF-IDF representation** and *medoid*
0	14257	**apprentice, celebapprentice, celebrity, interviewed, mr, run, season, ratings, trump2016, please** *"""@djf123: @realDonaldTrump @PJTV #DonaldWillBeTheGOPNominee!"*
1	1634	**entrepreneurs, yourself, success, focus, champion, passion, negotiation, learn, art, momentum** *"A very good way to pave your own way to success is simply to work hard and to be diligent" - Think Like a Champion*
2	1217	**rally, crowd, join, tickets, crowds, maga, south, evening, landed, soon** *Leaving Michigan now, great visit, heading for Iowa. Big Rally!*
3	1215	**border, wall, immigration, southern, illegal, laws, immigrants, patrol, drugs, borders** *Our Southern Border is under siege. Congress must act now to change our weak and ineffective immigration laws. Must build a Wall. Mexico, which has a massive crime problem, is doing little to help!*
4	936	**sleepy, fracking, 47, suburbs, firefighter, pack, basement, radical, abolish, wins** *A vote for Joe Biden is a vote to extinguish and eradicate your state's auto industry. Biden is a corrupt politician who SOLD OUT Michigan to CHINA. Biden is the living embodiment of the decrepit and depraved political class that got rich bleeding America Dry!*
5	893	**hotel, chicago, tower, luxury, building, rooms, restaurant, hotels, sign, spa** *.@TrumpSoHo features a striking glass walled building w/ loft inspired interiors http://t.co/yUn5d14t0Q NYC's trendiest luxury hotel*
6	891	**china, tariffs, trade, currency, products, tariff, countries, barriers, goods, billions** *....with the U.S., and wishes it had not broken the original deal in the first place. In the meantime, we are receiving Billions of Dollars in Tariffs from China, with possibly much more to come. These Tariffs are paid for by China devaluing, pumping, not by the U.S. taxpayer!*
7	852	**investigation, report, dossier, director, spy, emails, documents, collusion, page, obstruction** *The Mueller probe should never have been started in that there was no collusion and there was no crime. It was based on fraudulent activities and a Fake Dossier paid for by Crooked Hillary and the DNC, and improperly used in FISA COURT for surveillance of my campaign. WITCH HUNT!*
8	838	**discussing, obama, records, college, birth, charity, interview, certificate, applications, offer** *What is your thought as to why Obama refused millions for charity and did not show his records and applications?*
9	702	**golf, course, links, courses, club, monster, blue, championship, international, aberdeen** *Named best golf course in the world by @RobbReport, Trump Int'l Golf Links Scotland is a 7,400 yd par 72 http://t.co/7FYb3dEQcD*

Figure 5: **Iteration 1:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids.

ID	N	**c-TF-IDF representation** and *medoid*	Overl.
55	2468	**run, please, needs, 2016, need, runs, trump2016, trumpforpresident, running, mr** *"""@tlrchrstphrbrsn: @realDonaldTrump please run for president"""*	0.4 0.75
56	2099	**fake, media, nytimes, news, failing, sources, story, dishonest, reporting, stories** *So much Fake News being put in dying magazines and newspapers. Only place worse may be @NBCNews, @CBSNews, @ABC and @CNN. Fiction writers!*	0.0 0.54
57	984	**apprentice, celebrity, season, celebrityapprentice, celebapprentice, episode, nbc, cast, wait, 13th** *Celebrity Apprentice is rebroadcasting last weeks episode at 9 P.M. WITH A GREAT NEW EPISODE FEATURING @MELANIA TRUMP AT 10 P.M. - AMAZING!*	0.4 0.70
58	767	**inspiration, model, tweets, role, thanks, entrepreneur, genius, inspiring, tweet, birthday** *"""@balango212 Donald J. Trump, my biggest inspirational in life."" Thank you."*	0.0 0.49
60	625	**condolences, prayers, deepest, heroes, welcome, victims, families, honor, thoughts, shooting** *Saddened to hear the news of civil rights hero John Lewis passing. Melania and I send our prayers to he and his family.*	0.3 0.22
61	604	**truth, hate, honesty, haters, speaks, afraid, speak, honest, losers, listen** *"""@HeyMACCC: I feel like the only people who hate @realDonaldTrump are people who can't handle the truth...he tells it how it is."""*	0.0 0.27
62	598	**oscars, ratings, boring, dumbest, show, dummy, overrated, tv, television, comedian** *@DanielRobinsMD. @Lawrence is a very dumb guy who just doesn't know it-his show is a critical and ratings disaster!*	0.1 0.40
63	539	**hunt, witch, collusion, obstruction, conflicted, investigation, evidence, report, angry, witnesses** *The Greatest Witch Hunt In American History!*	0.4 0.27
64	534	**approval, rating, poll, leads, overall, surges, party, republican, 52, polls** *96% Approval Rating in the Republican Party. 51% Approval Rating overall in the Rasmussen Poll. Thank you!*	0.1 0.20
65	501	**entrepreneurs, touch, yourself, focus, goals, opportunities, momentum, passion, vision, requires** *Entrepreneurs: Stay focused and be tenacious. Pay attention to people who know what they're talking about. Stay fixed on your goals!*	0.5 0.77

Figure 6: **Iteration 2:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids. The overlap scores refer to the maximum token overlap (top) with any component from the first iteration, i.e. a BERTopic run, and the maximum sample overlap (bottom) with any cluster in the cluster hierarchy of HDBSCAN from the first iteration (i.e. all possible topics that could be found through BERTopic).

Topic ID	Topic Description
Trade	The exchange of goods, services, and capital.
Agriculture	The production, cultivation, and harvesting of crops and livestock.
Fluopyram	A medication used to treat certain types of pain.
Ballot Scam	A type of election-related scam where voters are tricked into returning a ballot to the wrong address.
None	General Topics

Figure 8: Topics generated by TopicGPT on the Trump dataset.

ID	N	**c-TF-IDF representation** and *medoid*	Overl.
132	1502	**hotel, golf, course, tower, luxury, links, club, resort, hotels, blue** *We do have a Trade Deficit with Canada, as we do with almost all countries (some of them massive). P.M. Justin Trudeau of Canada, a very good guy, doesn't like saying that Canada has a Surplus vs. the U.S.(negotiating), but they do...they almost all do...and that's how I know!*	0.5 0.52
135	554	**hurricane, prayers, responders, condolences, storm, victims, coast, local, water, rescue** *I have instructed the United States Navy to shoot down and destroy any and all Iranian gunboats if they harass our ships at sea.*	0.4 0.27
137	470	**ratings, fox, biased, credibility, daytime, anchors, network, debates, networks, loser** *RT @DailyCaller: CNN Receives Bad TV Rating For Speaker Pelosi Town Hall https://t.co/YxzzHljJAK*	0.1 0.19
138	440	**failing, apologize, nytimes, magazine, apology, apologized, times, paper, circulation, newspaper** *The New York Times has apologized for the terrible Anti-Semitic Cartoon, but they haven't apologized to me for this or all of the Fake and Corrupt news they print on a daily basis. They have reached the lowest level of "journalism," and certainly a low point in @nytimes history!*	0.0 0.22
140	424	**sources, enemy, fake, lie, hoax, lied, lies, media, stories, false** *"Anytime you see a story about me or my campaign saying ""sources said,"" DO NOT believe it. There are no sources, they are just made up lies!"*	0.1 0.17
141	351	**economy, change, changing, stronger, quarter, market, stock, growth, booming, wand** *Our economy is perhaps BETTER than it has ever been. Companies doing really well, and moving back to America, and jobs numbers are the best in 44 years.*	0.2 0.13
142	340	**leaving, crowd, landed, heading, evening, landing, soon, south, ready, departing** *I'm leaving now for Ireland, Spain, Scotland and elsewhere–crazy life!*	0.5 0.17
143	338	**billionaire, businessman, smartest, business, rich, genius, smart, chess, investor, businessmen** *""""It's more important to be smart than tough. I know businessmen who are brutally tough, but they're not smart." - Think Like A Billionaire"*	0.0 0.05
145	300	**rather, wish, prefer, would, better, awesome, actor, achieve, greatest, best** *""""@iamapatsfan: I would rather Donald Trump be the president than Jeb Bush."" Would do a much better job!"*	0.1 0.05
146	280	**penalty, killer, prison, death, prisoners, hostage, hostages, jail, sentenced, thug** *caught, he cried like a baby and begged for forgiveness...and now he is judge, jury. He should be the one who is investigated for his acts.*	0.0 0.04

Figure 7: **Iteration 4:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids. The overlap scores refer to the maximum token overlap (top) with any component from the first iteration, i.e. a BERTopic run, and the maximum sample overlap (bottom) with any cluster in the cluster hierarchy of HDBSCAN from the first iteration (i.e. all possible topics that could be found through BERTopic).

C.3 Trump Dataset: TopicGPT Topics

The TopicGPT baseline does not perform well with a backbone model (`Llama-3.2-1B`) of comparable size to our embedding model (278M parameters). In particular, the topic generation phase seems to have troubles identifying meaningfull patterns. We plot the resulting topics in Figure 8.

C.4 Chinese News Outlet Tweets: Components

ID	N	**c-TF-IDF representation** and *medoid* with (GPT-4o translations)
0	47336	**开庭, 看守所, 寻衅滋事, 律师, 刑事拘留, 派出所, 监狱, 刑拘, 罪, 访民** court session, detention center, picking quarrels and provoking trouble, lawyer, criminal detention, police station, prison, criminal detention, crime, petitioner 关注陈卫！关注四川！维权人士陈卫案开庭在即，欧阳懿被警方带走 *http://t.co/5k7PqPZH* Pay attention to Chen Wei! Pay attention to Sichuan! The trial of rights activist Chen Wei is about to begin, Ouyang Yi has been taken away by the police.
1	29472	**普京, 乌克兰, 俄, 烏克蘭, 基辅, 连斯基, 俄军, 俄罗斯, 泽, 莫斯科** Putin, Ukraine, Russia, Ukraine, Kyiv, Liansky, Russian army, Russia, Ze, Moscow 美国与北约：俄罗斯在乌克兰附近正在增兵而不是撤军 *https://t.co/IZ816MM4IM* The United States and NATO: Russia is increasing troops near Ukraine rather than withdrawing.
2	24930	**十八, 智慧, 常委, 改革开放, 二十大, 十九, 习近平, 政治局, 中共, 习** eighteen, wisdom, standing committee, reform and opening up, 20th National Congress, nineteen, Xi Jinping, Politburo, Chinese Communist Party, Xi 【# 时事大家谈“小人物”掀大风浪：凤姐动了谁的奶酪?】何清涟：胡锦涛时代和现在有很大的不同。胡时代只要不反党，玩什么都行。习近平这两年达到的效果，让人们除了主旋律假大空的话以外，什么话都不敢说。像凤姐这样的事情，习近平肯定看不来。*https://t.co/cY7KoWd1lf* [#CurrentAffairsTalk "Little People" Stir Up Big Waves: Who Did Sister Feng Offend?] He Qinglian: There is a big difference between the Hu Jintao era and now. During Hu's era, as long as you didn't oppose the party, you could do anything. The effect achieved by Xi Jinping in the past two years has made people afraid to say anything except for the main theme of empty and false words. Xi Jinping definitely wouldn't like things like Sister Feng.
3	23631	**金正恩, 韩美, 朝鲜, 北韩, 发射, 核试验, 弹道导弹, 半岛, 平壤, 无核化** Kim Jong-un, South Korea and the United States, North Korea, North Korea, launch, nuclear test, ballistic missile, peninsula, Pyongyang, denuclearization 韩美外长称北韩应履行无核化前期措施努力改善南北韩关系：[国际]：韩国国际广播电台报道：韩国外交通商部长官金星焕和美国国务卿希拉里·克林顿 *9* 日在美国华盛顿举行韩美外长会谈，就北韩核问题深入交换了意见。 在会谈上，两... *http://t.co/S5XD8QCP* South Korean and U.S. foreign ministers say North Korea should implement preliminary denuclearization measures and strive to improve inter-Korean relations.
4	21072	**特首, 普选, 立法会, 反送, 郑月, 香港立法会, 民主派, 梁振英, 娥, hong** Chief Executive, universal suffrage, Legislative Council, anti-extradition, Cheng Yuet, Hong Kong Legislative Council, pro-democracy camp, Leung Chun-ying, Carrie Lam, Hong 眼见 # 香港民主自由倒退，包括美国、英国和加拿大等身居海外的港人将筹备香港议会普选，一人一票直选民意代表，以维系香港的自决权，让国际社会能更鼎力支持香港人。*https://t.co/rurQAunPoF* Seeing the regression of democracy and freedom in Hong Kong, Hongkongers living overseas, including in the United States, the United Kingdom, and Canada, are preparing for a universal suffrage election for the Hong Kong parliament, with one person, one vote to directly elect representatives, in order to maintain Hong Kong's right to self-determination and to gain stronger support from the international community.
5	17729	**病毒, 肺炎, 病例, 感染, 核酸, 确诊, 防疫, 例, 检测, 武汉** virus, pneumonia, case, infection, nucleic acid, confirmed, epidemic prevention, case, testing, Wuhan 【# 时事大家谈热点快评：美国疫情持续恶化公众到底该不该戴口罩?】根据美国约翰霍普金斯大学的统计，截至 *4* 月 *2* 日早上 *7* 时 *40* 分，全美新冠确诊病例总数已超过 *21* 万人，疫情已造成超过 *5100* 人死亡。美国疾病控制与预防中心主任罗伯特雷德菲尔德本周透露，*CDC* 可能会建议公众使用口罩来抵御病毒在社区的传播。*https://t.co/umA0pMrozA* [#CurrentAffairsHotCommentary: Should the Public Wear Masks as the U.S. Pandemic Worsens?] According to statistics from Johns Hopkins University, as of 7:40 a.m. on April 2, the total number of confirmed COVID-19 cases in the U.S. has exceeded 210,000, with the pandemic causing over 5,100 deaths. Robert Redfield, director of the U.S. Centers for Disease Control and Prevention, revealed this week that the CDC might recommend the public to use masks to prevent the spread of the virus in the community.
6	11737	**世界杯, 奥运会, 奥运, 选手, 比赛, 运动员, 体育, 冬奥会, 金牌, 决赛** World Cup, Olympic Games, Olympics, athlete, competition, athlete, sports, Winter Olympics, gold medal, final 美国小将力克群雄摘得里约奥运射击金牌：巴西奥运会星期六进行第一天正式比赛，比赛项目包括公路自行车、沙滩排球、拳击、射击以及几个游泳项目。*https://t.co/ztp8Om4P0j* American young athlete overcomes competitors to win Rio Olympic shooting gold: The first day of official competitions at the Brazilian Olympics on Saturday included events such as road cycling, beach volleyball, boxing, shooting, and several swimming events.
7	9314	**航母, 南中国海, 南海, 海军, 航行, 海域, 军舰, 岛礁, sea, 巡航** aircraft carrier, South China Sea, South Sea, navy, navigation, sea area, warship, island reef, 海, cruise 中国海军 *5* 月 *3* 日宣布，辽宁舰航母编队日前在西太平洋海域进行远海实战化训练。日本自卫队本周一观察到包括一艘航母在内的八艘中国海军舰艇日本南部冲绳群岛，并在东海起降了舰载直升机。 *https://t.co/YdyahekvTR* On May 3, the Chinese Navy announced that the Liaoning aircraft carrier group recently conducted open-sea combat training in the Western Pacific. The Japanese Self-Defense Forces observed eight Chinese naval vessels, including an aircraft carrier, near Japan's southern Okinawa Islands, and carrier-based helicopters took off and landed in the East China Sea.
8	7945	**关税, 贸易谈判, 加征, 贸易, trade, 第一阶段, tariffs, 贸易战, 貿易, 商品** tariffs, trade negotiations, impose, trade, 貿易, first phase, 关税, trade war, trade, goods 【中美达成第一阶段经贸协议】美国原定于 *15* 日启动的第 *4* 轮加征关税停止，部分已经加征关税的税率也将下调。中国将扩大进口美国农产品，并扩大金融市场开放……*https://t.co/PJLO708evZ https://t.co/1oCWoMeODW* China and the U.S. reach a phase one trade agreement" - The U.S. will halt the fourth round of tariff increases originally set to start on the 15th, and some existing tariff rates will be reduced. China will expand imports of American agricultural products and further open its financial markets...
9	7916	**客机, 航班, 航空, 航空公司, 飞机, 波音, 马航, 机场, 坠毁, 失事** airliner, flight, aviation, airline, airplane, Boeing, Malaysia Airlines, airport, crash, accident 亚航班机失联机上共有 *162* 人：总部在吉隆坡的廉价航空公司亚航宣布，该公司从印尼的泗水到新加坡的一架班机与空中交通管制系统失去联系。*http://t.co/FrUZVRKFmV* AirAsia flight missing with 162 people on board: The low-cost airline AirAsia, headquartered in Kuala Lumpur, announced that one of its flights from Surabaya, Indonesia, to Singapore has lost contact with air traffic control systems.

Figure 9: **Iteration 1:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids.

ID	N	**c-TF-IDF representation** and *medoid* with (GPT-4o translations)	O.
212	10602	**示威者, 游行, 示威, 抗议者, 遊行, 集会, 抗议, 催泪弹, 警民, 驱散** protester, march, demonstration, protester, march, assembly, protest, tear gas, police-civilian, disperse 数百名泛民政党和团体的抗议者周一晚在湾仔警署外聚集，对前往警署预约拘捕的 *9* 位占领运动发起人和主要组织者表达支持。包括占中三子戴耀廷、陈健民和朱耀明在内的 *9* 人，在警署外逐一发表讲话，坚持雨伞运动不屈不挠，公民抗命无畏无惧。他们随后在支持者的口号中走进警署。现场有大批媒体报道。*https://t.co/hGig2lE6OJ* Hundreds of protesters from pro-democracy parties and groups gathered outside the Wan Chai police station on Monday night to show support for the nine initiators and main organizers of the Occupy Movement who went to the police station for scheduled arrests. The nine, including the Occupy Central trio of Benny Tai, Chan Kin-man, and Reverend Chu Yiu-ming, each gave speeches outside the station, insisting on the indomitable spirit of the Umbrella Movement and the fearless nature of civil disobedience. They then entered the police station amid supporters' chants. A large number of media outlets reported on the scene.	0
213	9970	**孩子, 女儿, 母亲, 父亲, 爸爸, 儿子, 父母, 儿童, 妈妈, 她** child, daughter, mother, father, dad, son, parents, children, mom, she 姚立法两岁的女儿在经过父亲失踪、多次抄家、被断水断电砸门等一系列的事情之后，在火热的夏天全身长满痱子，容易哭闹，变得敏感胆小。"爸爸回家，要爸爸"成为孩子说得最多的一句话。 Yao Lifa's two-year-old daughter, after experiencing her father's disappearance, multiple home raids, being cut off from water and electricity, and having the door smashed, developed a rash all over her body in the hot summer, became easily irritable, and turned sensitive and timid. "Daddy come home, want daddy" became the child's most frequently spoken phrase.	0
214	8934	**贸易战, 挑战, 面临, 地缘, 不确定性, 风险, 战略, 大国, 竞争, 美中关系** trade war, challenge, face, geopolitical, uncertainty, risk, strategy, major power, competition, US-China relations 【郑永年：世界秩序危机及其未来】尽管今天的中国面临各种问题，但崛起难以避免。不过，做大国很不容易，需要面临各种挑战。如果要做大国，中国有很多有关大国的问题需要回答，有很多大国必须面临的挑战和大国所需要担负的责任。*http://t.co/yNK3Lk8jtP* Zheng Yongnian: The Crisis of World Order and Its Future" Despite the various issues China faces today, its rise is inevitable. However, being a major power is not easy and requires facing various challenges. If China wants to be a major power, it has many questions related to being a major power that need to be answered, many challenges that must be faced, and responsibilities that a major power needs to bear.	1
215	8800	**开庭, 法官, 法院, 驳回, 上诉, 中院, 开庭审理, 庭审, 不服, 裁定** court session, judge, court, dismiss, appeal, intermediate court, hold a court session, trial, dissatisfied, ruling 维权网：　无锡中院国赔委无视最高法院　*116*　号指导判例以"执行尚未终结"驳回张建平的国赔申请 *https://t.co/YR6etEdtSr https://t.co/gWxjPSByDW* Rights Protection Network: Wuxi Intermediate Court's State Compensation Committee Ignores Supreme Court's Guiding Case No. 116 and Rejects Zhang Jianping's State Compensation Application on Grounds of "Execution Not Yet Concluded	1
216	8373	**韩币, 现钞, 美元兑, 外汇牌价, 买入价, 顾客, 汇率, 股价指数, 帐号, 兑** Korean won, cash, USD to, exchange rate, buying price, customer, exchange rate, stock index, account number, to 许多进入乌克兰的俄罗斯军车上涂有拉丁字母 *Z*，尽管"*Z*"在俄文使用的基里尔字母中本不存在。它到底有何含义？有媒体推测，*Z* 是俄罗斯西部军区 *ZVO* 的第一个字母，因此出现在部分坦克军车上。按照俄国防部 *3* 月初的解释，*Z* 代表 *Za pobedu*（英文转写），俄语"为了胜利"*https://t.co/yPiRSptAnS* Many Russian military vehicles entering Ukraine are marked with the Latin letter Z, even though "Z" does not exist in the Cyrillic alphabet used in Russian. What does it mean? Some media speculate that Z is the first letter of the Western Military District ZVO in Russia, hence it appears on some tanks and military vehicles. According to the Russian Ministry of Defense's explanation in early March, Z stands for Za pobedu, which means "For victory" in Russian.	9
217	6984	**贬值, 日元, 下跌, 人民币, 跌幅, 升值, 美元, 货币, fall, 美元汇率** devaluation, yen, decline, renminbi, drop, appreciation, dollar, currency, fall, dollar exchange rate 【人民币兑美元跌至近八年最低水平 →*https://t.co/nbkvpcxruT*】- 人民币兑美元周二跌至近八年来最低水平。特朗普在美国总统大选中获胜后，人民币加速下跌，目前这种跌势仍在持续。*https://t.co/iJcMQJIMdB* The Chinese yuan fell to its lowest level against the US dollar in nearly eight years on Tuesday. After Trump's victory in the US presidential election, the yuan accelerated its decline, and this trend is still continuing.	3
218	5934	**炸弹, 爆炸, 丧生, 受伤, 炸死, 至少, 发生爆炸, 自杀, 自杀式, 死亡** bomb, explosion, killed, injured, killed by explosion, at least, explosion occurred, suicide, suicide-style, death *Eight killed when magnitude 5.5 earthquake hits China's northwest https://t.co/IaGE2Qf6b7* 八人遇难，5.5 级地震袭击中国西北地区	6
219	5912	**医院, 医生, 病人, 治疗, 手术, 医疗, 护士, 精神病, 精神病院, 醫院** hospital, doctor, patient, treatment, surgery, medical, nurse, mental illness, psychiatric hospital, hospital *RT @pufei:* 黄琦先生今日在华西医院进行冲击治疗，院方再次拒绝提供床位，将黄琦先生安置在病房厕所旁的行军椅上进行治疗。 RT @pufei: Mr. Huang Qi underwent shock treatment today at Huaxi Hospital. The hospital once again refused to provide a bed, placing Mr. Huang Qi on a folding chair next to the bathroom in the ward for treatment.	1
220	5828	**六四, 纪念, 悼念, 周年, 纪念日, 纪念活动, 烛光, 国葬, 哀悼, 葬礼** June Fourth, commemorate, mourn, anniversary, memorial day, commemorative event, candlelight, state funeral, mourning, funeral 六四 *26* 周年纪念活动在德国斯图加特展开：八九-六四天安门民主运动迎来了又一个纪念日。像以往一样，每年的六四前后，各种形式的纪念活动都会在全球各地展开。要求民主的呼声从未中断。今天，天安门学运 *26* 周年之际，欧洲六四纪念活… *http://t.co/bonTqv4gI6* The 26th anniversary of June 4th commemorative activities held in Stuttgart, Germany: The 1989 Tiananmen Democracy Movement has reached another anniversary. As in the past, around June 4th each year, various forms of commemorative activities are held around the world. The call for democracy has never ceased. Today, on the 26th anniversary of the Tiananmen student movement, the European June 4th commemoration…	1
221	5774	**峰会, 会晤, 首脑, 会谈, 十国集团, 举行会谈, 团长, 六方会谈, 商讨, 外长** summit, meeting, leader, talks, G10, hold talks, delegation leader, Six-Party Talks, discuss, foreign minister *G20* 財長和央行行長會議在成都市舉行 *https://t.co/WkXH5rQswN https://t.co/1ViZ6yC8gk* The G20 Finance Ministers and Central Bank Governors Meeting is held in Chengdu.	1

Figure 10: **Iteration 2:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids. Overlap describes the maximum number of tokens a component's representation has in common with any component from the first iteration.

ID	N	**c-TF-IDF representation** and *medoid* with (GPT-4o translations)	O.
417	31500	**开庭, 律师, 寻衅滋事, 看守所, 罪, 颠覆, 检察院, 派出所, 起诉, 开庭审理** court session, lawyer, picking quarrels and provoking trouble, detention center, crime, subversion, procuratorate, police station, prosecute, court trial 伯明翰警方周三证实，*48* 岁的孔琳琳已经被正式起诉。她预计于 *11* 月 *7* 日在英国伯明翰出庭。*https://t.co/lXL878HwUh* Birmingham police confirmed on Wednesday that 48-year-old Kong Linlin has been formally charged. She is expected to appear in court in Birmingham, UK, on November 7.	6
418	7918	**候选人, 初选, 共和党, 安哲秀, 支持率, 竞选, 党, 补选, 选举, 参选** candidate, primary election, Republican Party, Ahn Cheol-soo, approval rating, campaign, party, by-election, election, run for office 澳新星车手敲定下赛季加盟红牛车队：*F1* 红牛车队确认，来自澳大利亚珀斯的丹尼尔·里卡多将在下个赛季取代退役的马克·韦伯，成为车队的正式车手。里卡多将与塞巴斯蒂安·维特尔搭档，一起征战 *2014* 年的比赛。*http://t.co/pURtGeKfbL* Australian rising star driver confirmed to join Red Bull Racing next season: F1 Red Bull Racing confirms that Daniel Ricciardo from Perth, Australia, will replace the retiring Mark Webber as the team's official driver next season. Ricciardo will partner with Sebastian Vettel to compete in the 2014 races.	4
419	6319	**会晤, 举行会谈, 会谈, 会面, 首脑, 金正恩, 文在寅, 朴槿惠, 通电话, 青瓦台** meeting, hold talks, talks, meeting, leader, Kim Jong-un, Moon Jae-in, Park Geun-hye, phone call, Blue House 岸田首相與烏克蘭總統澤連斯基舉行電話會談 *https://t.co/MKUqATfOZ2 https://t.co/1JoEpuQD6F* Prime Minister Kishida holds a phone call with Ukrainian President Zelensky.	1
420	5762	**你们, 我, 你, 我要, 故事, 我会, 写, 右派, 问, 我家** you, I, you, I want, story, I will, write, rightist, ask, my home 如果是你，你会怎么做呢？ *1.* 报警 *2.* 猛按车笛 *3.* 算我倒霉，办完事再回来取车 *4.* 老实说，除了等我还能做些什么呢？*https://t.co/HV9mMH1A8g* If it were you, what would you do? 1. Call the police 2. Honk the horn repeatedly 3. Consider myself unlucky, finish my business and come back for the car 4. Honestly, what else can I do but wait?	1
421	4549	**社会主义, 马克思主义, 共产主义, 马克思, 思想, 共产党, 权力, 意识形态, 程晓农, 官媒姓** socialism, Marxism, communism, Marx, thought, Communist Party, power, ideology, Cheng Xiaonong, state media 【# 管中祥：避談政治的教育，才是不折不扣的 # 政治】全文：*https://t.co/md8vsIAUZh* 什麼是政治性內容？要絕口不提？學術研究、課堂知識無法真空存在，許多學門所關切的議題，不僅與所在環境相繫，也和世界相連，很難不碰觸「政治」。 有不同意見才能讓教室成為民主的意識與行動的養成所。*https://t.co/3R8vVtMXtu* What is political content? Should it be completely avoided? Academic research and classroom knowledge cannot exist in a vacuum; many disciplines are concerned with issues that are not only connected to their environment but also linked to the world, making it difficult to avoid 'politics'. Having different opinions allows the classroom to become a place for nurturing democratic awareness and action.	0
422	4458	**大学, 哈佛, 学术, 招生, 大学排名, 论文, 学院, 孔子, 哈佛大学, 高等教育** university, Harvard, academic, admissions, university ranking, thesis, college, Confucius, Harvard University, higher education 美国研究生院排名出炉耶鲁法学院战胜哈佛：根据《美国新闻与世界报道》发布的 *2015* 全美五类研究生院排名，宾州大学的沃顿商学院较去年上升了两位，与哈佛大学，斯坦福大学并列第一。工程学院的前三名分别为麻省理工学院，斯坦福大… *http://t.co/H02mlIfTFc* The ranking of American graduate schools is out: Yale Law School surpasses Harvard. According to the 2015 rankings of five types of graduate schools in the United States released by "U.S. News & World Report," the Wharton School of the University of Pennsylvania has risen two places from last year, tying with Harvard University and Stanford University for first place. The top three engineering schools are Massachusetts Institute of Technology, Stanford University, and...	2
423	3773	**奥巴马, 国情咨文, 奥巴馬, 白宫, 广岛, 讲话, 米歇尔, 会晤, 枪支, 内塔尼亚胡** Obama, State of the Union, Obama, White House, Hiroshima, speech, Michelle, meeting, firearms, Netanyahu 奥巴马获胜连任后飞回白宫，重建经济：刚刚在美国大选中获胜连任的美国总统奥巴马无暇享受胜利的滋味，于当地时间 *7* 日从竞选总部所在地芝加哥回到首都华盛顿，继续工作。*http://t.co/F4lTh7ur* After winning re-election, Obama flew back to the White House to rebuild the economy: The U.S. President Obama, who just won re-election in the U.S. election, had no time to enjoy the taste of victory and returned to the capital Washington from his campaign headquarters in Chicago on the 7th local time to continue working.	1
424	3681	**明哲, 李嘉诚, 李, 李旺阳, 李文亮, 李波, 李登辉, 李小琳, 李在, 李家** Mingzhe, Li Ka-shing, Li, Li Wangyang, Li Wenliang, Li Bo, Lee Teng-hui, Li Xiaolin, Li Zai, Li Jia 勒索女主张曾与李炳宪交往被分手心生恨意：利用私密视频勒索演员李炳宪数十亿韩元的模特李某再爆惊人内幕，称因接受不了李炳宪提出分手才出此下策。李某以敲诈勒索未遂的罪名已经被拘留。 据韩国一家媒体 *11* 日报道，李某的律师主张说：… *http://t.co/aVps9TurYv* A model named Li, who attempted to extort tens of billions of Korean won from actor Lee Byung-hun using private videos, revealed shocking details, claiming she resorted to this because she couldn't accept Lee Byung-hun's breakup. Li has been detained on charges of attempted extortion.	1
425	3337	**改变, 改革, 变化, 体制改革, 优先, 战略, 转变, 三中全会, 深化改革, 对华政策** change, reform, change, institutional reform, priority, strategy, transformation, Third Plenary Session, deepen reform, China policy 沙特阿拉伯王储穆罕默德周三说，沙特将在五年内"改头换面"，称要展开改革使国家经济多元化，不再依赖出口石油。*https://t.co/hzKFtl1vfT* Saudi Arabia's Crown Prince Mohammed said on Wednesday that Saudi Arabia will "transform" in five years, stating that reforms will be carried out to diversify the national economy and reduce reliance on oil exports.	1
426	3268	**博物馆, 文化遗产, 世界遗产, 名录, 文物, 纪念馆, 联合国教科文组织, 森林, 申遗, 故宫** museum, cultural heritage, world heritage, list, cultural relic, memorial hall, UNESCO, forest, apply for world heritage status, ... 緊鄰公園的除了烏克蘭最高學府謝甫琴科大學之外，尚有烏國館藏規模最大、有百年歷史的「東西方藝術博物館」(*Khanenko Museum*)，與一處兒童醫院。*8/9* Next to the park, besides Ukraine's top university, Shevchenko University, there is also the Khanenko Museum, the largest art collection in Ukraine with a history of a hundred years, and a children's hospital.	1

Figure 11: **Iteration 3:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids. Overlap describes the maximum number of tokens a component's representation has in common with any component from the first iteration.

C.5 Hausa Dataset: Components

ID	N	**c-TF-IDF representation** and *medoid* with (GPT-4o translations)
0	10509	**masha, insha, yasaka, allahu, jikansa, akbar, yajikansa, jikanshi, rahamarsa, jikan** blessing, will, you, God, your child, great, your child, your child, our mercy, child *afwan da allah wayece rahama sadau* May Allah forgive and have mercy on us.
1	10185	**musulunci, musulmai, musulmi, addinin, musulma, musulinci, islam, addini, muslim, islama** Islam, Muslims, Muslims, religion, Muslim, Islam, Islam, religion, Muslim, Islam *karya kike dama tun fari kintaba ridda wanna tufafi da kikasa sukuma namusulmaine musulumci yayadda da tabarrunne keda kanki kyasan bakida alaka da musulumci* You are lying, you have worn this outfit before and it is not Islamic. Islam does not agree with showing off, and you know you have no connection with Islam.
2	7341	**madrid, barcelona, liverpool, chelsea, arsenal, manchester, real, gasar, league, united** Madrid, Barcelona, Liverpool, Chelsea, Arsenal, Manchester, Real, Gasar, League, United *premier league an kama mai kalaman wariya a wasan manchester* Premier League arrests a player for racist remarks in the Manchester match.
3	7161	**coronavirus, covid, corona, virus, rigakafin, korona, kamu, cutar, allurar, kamuwa** coronavirus, covid, corona, virus, vaccine, corona, infection, disease, vaccine, infection *jumillar maaikatan lafiya sun kamu da cutar coronavirus a kano* Many people in Kano have contracted the coronavirus.
4	6244	**bindiga, garkuwa, hallaka, rundunar, sanda, cafke, gobara, sandan, harin, arin** gun, shield, destroy, army, police, arrest, fire, policeman, attack, attack *yan bindiga sun yi garkuwa da mutum a shirori jihar neja* Gunmen have kidnapped a person in Shiroro, Niger State.
5	5400	**bbc, bbchausa, hausa, news, tarihin, kubani, sashen, labarai, bakuda, haussa** bbc, bbchausa, hausa, news, history, give me, section, news, you don't have, hausa *kai bbc makaryatanebasu gayan aibin kasarsu senawasu* BBC is not giving them the opportunity to speak about their
6	4769	**gaskiyane, gaskiyar, gaske, gaskiya, true, gaskiyan, maganarka, adalci, adalcin, halinta** truth, truth, true, truth, true, truths, your statement, justice, justice, your character *gaskiya dai abun yfara katsanta* Truly, the situation is becoming
7	4672	**nigerian, niger, nigeria, nigeriya, nigeriar, talakan, kasata, nigerians, talakawan, republic** Nigerian, Niger, Nigeria, Nigeria, Nigeria, poverty, wealth, Nigerians, poor people, republic *wae ahakan ake so acanza nigeria ana irin wanan zalincin karara* Why do they want to change Nigeria with this kind of blatant oppression?
8	4167	**hmmm, hmm, hmmmm, hmmmmm, hm, uhmm, uhm, uhmmm, hmmmmmm, hummm** hmmm, hmm, hmmmm, hmmmmm, hm, uhmm, uhm, uhmmm, hmmmmmm, hummm *hmmmm tohhh dan masani jikan masanawa ae sae kaje ka sakosu kawae* Hmmm, well, an expert knows the offspring of experts, so just go and ask them.
9	3953	**yarjejeniyar, israila, bukaci, syria, iran, turkiya, majalisar, dokar, sudan, fadarsa** diplomacy, Israel, request, Syria, Iran, Turkey, council, law, Sudan, his/her place *majalisar najeriya ta bukaci buhari ya kafa dokartabaci* The Nigerian parliament has asked Buhari to declare a state of emergency.

Figure 12: **Iteration 1:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids.

ID	N	**c-TF-IDF representation** and *medoid* with (GPT-4o translations)	Overl.
102	19015	**shiru, magana, gaya, kanku, maganar, fada, said, naka, miki, kike** silent, talk, tell, yourselves, the talk, say, said, yours, to you, you are *kaji wata magana ta banxa da kake fada wadane manya kasa ku kenan kune manya kasa dole kace hakan tunda abin bashi kai kanka ba* Understand a useless talk that you are saying, you elders of the land, you are the elders of the land, you must say that since the matter does not concern you.	0.0
103	6995	**allahumma, hai, summa, makomarsa, jikanshi, jikansa, subhanallah, makoma, innaka, yajikansa** Oh Allah, yes, then, his place, his body, his body, glory be to Allah, place, indeed you, his body *assalamu alekum barkammudawarhaka* Peace be upon you	0.3
104	6212	**yasan, san, gane, sani, fahimci, sanin, fahimta, fahimtar, yasani, sane** know, know, understand, know, understand, knowing, understanding, understanding, knows, aware *wanda yasan mene kano shine xai fahimci abinda jagoran arewa yake nupi sannan ya gane nupinsa* Whoever understands what the northern leader is saying will also understand his own language.	0.0
105	6077	**gani, kallo, see, kallon, ganin, tunanin, ganina, duba, samodara, kalli** look, see, see, seeing, seeing, thinking, my sight, check, samodara, look *rahma sadau kinyi gaba mahasada sai kallo magulmata sai gani* Rahma Sadau, keep moving forward, let the envious watch, and let	0.1
106	4846	**matsala, talauci, wahala, matsalar, matsalan, matsaloli, matsalolin, kalubale, problem, wuya** problem, poverty, trouble, problem, problem, problems, problems, challenge, problem, difficulty *ya zamto cikin manyan matsalolin kasar da yana da wahalar samuwaidan kuma ya samu ga yawan alala da karancin nagarta matukar basu nemo bakinzaren ba to dole su kasance cikin wahala da rudani yau an wayi gari matsalar ta shiga gidan kowa* We are facing major problems in the country that are difficult to address, and with the abundance of challenges and lack of quality, if they do not find solutions, they will inevitably remain in hardship and confusion; today, the problem has entered everyone's home.	0.1
107	4809	**kotu, hukunci, doka, hukuncin, kotun, hukunta, yanke, sharia, koli, dokar** court, judgment, law, judgment, the court, to judge, to decide, Islamic law, large, the law *wannan hukunci ne bisa doran doka wannda ya ga bai yi daidai ba to se ya daukaka kara* This is a judgment based on the law, and whoever feels it is not right should appeal.	0.1
108	4594	**dawo, koma, dawowa, maimaita, mayar, komawa, back, sake, maida, mashekiya** return, go back, return, repeat, give back, return, back, redo, return, return *tabbijanledoci sai kasata nageriya* The government will not allow Nigeria	0.1
109	4141	**jikansu, tona, kubutar, garesu, yabasu, asirinsu, maganinsu, taimakesu, gafarta, sharrinsu** sacrifice, light, to save, to help, to speak, their secrets, their words, to assist, to forgive, their evil *qoqari su kayi da zaa jinjina musu* Efforts that deserve commendation.	0.2
110	3879	**aikin, aiki, aikinka, aikinku, aikinsa, office, business, job, aikinsu, work** work, work, your work, your work, his work, ofis, kasuwanci, aiki, their work, work *aiki aiki aiki wani aikin sai kano* Work, work, work, what work except Kano.	0.0
111	3779	**musulmai, musulunci, musulmi, islam, musulinci, addinin, musulman, musulmin, khadimul, musulunchi** muslims, islam, muslim, islam, islam, religion, muslim, muslims, servant, muslim Muslim, Islam, Muslim, Islam, Islam, religion, Muslim, Muslim, servant, Muslim *komi kafiri xaiyi makiyin musulunci ne da musulmai* The infidels are the enemies of Muslims.	0.6

Figure 13: **Iteration 2:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids. Overlap describes the maximum number of tokens a component's representation has in common with any component from the first iteration.

ID	N	**c-TF-IDF representation** and *medoid* with (GPT-4o translations)	Overl.
253	13223	**talauci, kidnappers, talaka, yunwa, talakawa, bandits, sannan, rashin, kidnapping, tausayin** suffering, kidnappers, poor person, hunger, poor people, bandits, then, lack, kidnapping, compassion *daman munsani kuma mundaina binsa* We have seen the light and we have	0.2
254	7468	**tuba, manzon, kika, miki, kiyi, kiji, saw, yafe, w, kikayi** repent, messenger, you did, to you, do, you know, see, forgive, and, you did *kaga maiyi don allah kayi naka kayina rogo malam munan daram muna tayaka daaddua allah kareka daga sharrin makiaya malam inda tun abaya gwamanati natafiya da ireirenku da sai allah yasan matakin damuka taka* Please, for the sake of Allah, do your part and help us, we are in need of your assistance. May Allah protect you from the evil of the oppressors. Since the time of the government, you have been like this, and only Allah knows the extent of your actions.	0.2
255	7068	**babu, zamuce, bace, bazan, no, baiyi, taba, dace, daceba, bazai** father, we will, lost, I will not, no, he/she/it did not, touch, good, it is good, he/she/it will not *babu wanda zai yayemuna wannan sai allah addua kawai zamuyi janaa* There is no one who can do this for us except God, we will only pray.	0.0
256	6758	**kina, kike, kanku, ruwanku, kunyi, kika, zaki, uwarku, kukeyi, baku** your, you, your, your water, you, you, lion, your mother, you are doing, not you *kana shugaba bakasan me zakai* You don't know what you will do.	0.0
257	5190	**zaiyi, zasuyi, zatayi, zamuyi, will, insha, bige, sabuwa, chacha, zaiga** we will, they will do, they will sell, we will do, will, if God wills, beat, new, new, he will go *najeriya zata lallasa afirka takudu* Nigeria will dominate Africa.	0.1
258	4818	**koshin, acikin, lauje, nadi, yantattun, shiga, mati, muje, fckng, keep** outside, inside, leaf, pound, free, enter, mat, let's go, fckng, keep *to kai meye ruwanka a ciki munafuki* Why are you involving yourself in this, hypocrite?	0.1
259	4622	**ronaldo, league, tottenham, united, psg, juventus, mourinho, messi, bayern, champions** Ronaldo, liga, Tottenham, United, PSG, Juventus, Mourinho, Messi, Bayern, şampiyonlar *kasuwar cinikin yan kwallon kafa makomar ronaldo sterling jesus nunez romero giroud* football transfer market future of Ronaldo Sterling Jesus Nunez Romero Giroud	0.2
260	3477	**bindiga, garkuwa, cafke, neja, sanda, karamar, rundunar, sandan, hallaka, hatsarin** gun, shield, capture, snake, fence, small, battalion, soldier, ambush, accident *wani dan bindiga ya kai hari wata makaranta a birnin kazan da ke kasar rasha da safiyar ranar talata ya kashe mutum takwas bakwai daliban da ke aji takwas da kuma malami guda ya kuma ji wa wasu raunukan da suka kai su ga kwanciya a asibiti a cewar jamian rasha ap* A gunman attacked a school in the city of Kazan, Russia, on Tuesday morning, killing eight people, seven of whom were eighth-grade students and one teacher, and injuring others who were hospitalized, according to Russian officials.	0.7
261	2797	**congratulations, hai, allahumma, main, mein, congrats, innaka, pyaar, tum, hoon** congratulations, hi, O Allah, I, in, congratulations, indeed, love, you, am *fakuluma kaddararrahmanu mafulunwamaiyatawakkalu alallahu fahuwa hasbuk* Speak about the destiny of the Most Merciful	0.2
262	2651	**said, gwara, sunce, gaya, kauye, fada, bahaushe, nace, makaho, aniya** said, told, they said, tell, village, say, Hausa person, I said, blind person, intention *kauye garin su kare yace nima kaini akai kaidai ai shaani danbirni yacuci na kauye* The village said take me there too, but the city deceived the villager.	0.0

Figure 14: **Iteration 3:** TOP10 components by number of samples N together with their c-TF-IDF representations and medoids. Overlap describes the maximum number of tokens a component's representation has in common with any component from the first iteration.

C.6 Performance of Transformed Embeddings on MTEB

We benchmark the performance of the embeddings generated by a pipeline of SBERT and the SCA transformation against the original SBERT embeddings and an embedding consisting of the first 190 principal components found through PCA, which is equal to the number of components found by SCA trained on the full Trump dataset. Furthermore, we include the same benchmark on components from just the first iteration and a 55-dimensional PCA.

As a linear decomposition, SCA is a transformation of the embeddings into an explainable space, where each dimension corresponds to the score along one semantic component as defined by formula (2) (see Appendix A.2 for details). We use this interpretation of the component scores to evaluate the topic distribution in terms of its performance as a representation of samples in downstream tasks defined in the MTEB benchmark (Muennighoff et al., 2023) and compare against the vanilla embeddings as well as a

Ground truth	Sample	SCA	LDA*	TopicGPT*
Education	*Bills Perkins Fund for Equity and Excellence. This bill amends the Carl D. Perkins Career and Technical Education Act of 2006 to replace the existing Tech Prep program with a new competitive grant program to support career and technical education. Under the program, local educational agencies and their partners may apply for grant funding to support: career and technical education programs that are aligned with postsecondary education programs, dual or concurrent enrollment programs and early college programs, certain evidence-based strategies and delivery models related to career and technical education, teacher and leader experiential [...]*	leas, students, student, school, schools, ihes, 1965, education, elementary, academic	City infrastructure: city, building, area, new, park	Education: Mentions policies and programs related to higher education and student loans.

Table 3: Example from the Bills dataset from Pham et al. (2024). The SCA method is able to assign the correct topic relating to education. Results marked with * are copied from Pham et al. (2024).

PCA transformation of the same dimension.

The main results of all chosen benchmarks are shown in table 15. SCA as well as PCA perform uniformly well over all benchmarks and match the performance of the base model with minor variations.

C.7 Example from Pham et al. (2024)

Pham et al. (2024) provide an example topic assignment for a sample from the Bills dataset to illustrate the qulitative validity of their approach. Looking at the topic assigned by SCA, we argue that our method is also able to match the correct topic relating to education, see 3.

task name	type	metric	SBERT	SCA	PCA	SCA'	PCA'
AmazonCounterfactualClassification	Class.	accuracy	0.758	0.752	0.757	0.703	0.737
AmazonPolarityClassification	Class.	accuracy	0.764	0.768	0.768	0.778	0.781
AmazonReviewsClassification	Class.	accuracy	0.385	0.385	0.385	0.370	0.390
ArguAna	Retr.	ndcg_at_10	0.489	0.486	0.479	0.456	0.457
ArxivClusteringP2P	Clust.	v_measure	0.379	0.371	0.374	0.352	0.350
ArxivClusteringS2S	Clust.	v_measure	0.317	0.314	0.315	0.299	0.302
AskUbuntuDupQuestions	Rerank.	map	0.602	0.609	0.606	0.595	0.599
BIOSSES	STS	spearman	0.763	0.754	0.765	0.711	0.750
Banking77Classification	Class.	accuracy	0.811	0.802	0.808	0.768	0.789
BiorxivClusteringP2P	Clust.	v_measure	0.330	0.326	0.318	0.291	0.287
BiorxivClusteringS2S	Clust.	v_measure	0.294	0.292	0.292	0.270	0.269
CQADupstackRetrieval	Retr.	ndcg_at_10	0.313	0.309	0.314	0.268	0.278
ClimateFEVER	Retr.	ndcg_at_10	0.153	0.151	0.156	0.132	0.126
DBPedia	Retr.	ndcg_at_10	0.262	0.259	0.262	0.217	0.226
EmotionClassification	Class.	accuracy	0.459	0.454	0.457	0.438	0.440
FEVER	Retr.	ndcg_at_10	0.568	0.553	0.553	0.469	0.458
FiQA2018	Retr.	ndcg_at_10	0.230	0.224	0.231	0.183	0.194
HotpotQA	Retr.	ndcg_at_10	0.370	0.350	0.372	0.248	0.263
ImdbClassification	Class.	accuracy	0.646	0.647	0.646	0.647	0.644
MSMARCO	Retr.	ndcg_at_10	0.571	0.555	0.568	0.537	0.541
MTOPDomainClassification	Class.	accuracy	0.892	0.880	0.887	0.839	0.863
MTOPIntentClassification	Class.	accuracy	0.687	0.658	0.675	0.595	0.632
MassiveIntentClassification	Class.	accuracy	0.693	0.684	0.688	0.652	0.667
MassiveScenarioClassification	Class.	accuracy	0.762	0.749	0.761	0.715	0.741
MedrxivClusteringP2P	Clust.	v_measure	0.319	0.317	0.320	0.305	0.307
MedrxivClusteringS2S	Clust.	v_measure	0.315	0.311	0.317	0.305	0.306
MindSmallReranking	Rerank.	map	0.302	0.300	0.292	0.301	0.289
NFCorpus	Retr.	ndcg_at_10	0.255	0.257	0.259	0.230	0.234
NQ	Retr.	ndcg_at_10	0.336	0.329	0.340	0.277	0.306
QuoraRetrieval	Retr.	ndcg_at_10	0.864	0.862	0.863	0.850	0.853
RedditClustering	Clust.	v_measure	0.457	0.455	0.457	0.414	0.416
RedditClusteringP2P	Clust.	v_measure	0.521	0.513	0.516	0.481	0.484
SCIDOCS	Retr.	ndcg_at_10	0.140	0.140	0.141	0.122	0.125
SICK-R	STS	spearman	0.796	0.793	0.798	0.783	0.787
STS12	STS	spearman	0.779	0.784	0.759	0.785	0.755
STS13	STS	spearman	0.851	0.849	0.852	0.829	0.842
STS14	STS	spearman	0.808	0.808	0.797	0.793	0.783
STS15	STS	spearman	0.875	0.872	0.868	0.856	0.859
STS16	STS	spearman	0.832	0.833	0.820	0.826	0.812
STS17	STS	spearman	0.870	0.866	0.858	0.840	0.843
STS22	STS	spearman	0.635	0.637	0.608	0.632	0.599
STSBenchmark	STS	spearman	0.868	0.866	0.858	0.853	0.847
SciDocsRR	Rerank.	map	0.781	0.778	0.774	0.761	0.754
SciFact	Retr.	ndcg_at_10	0.503	0.500	0.502	0.429	0.442
SprintDuplicateQuestions	PClass.	similarity_ap	0.906	0.912	0.911	0.892	0.897
StackExchangeClustering	Clust.	v_measure	0.530	0.531	0.535	0.493	0.501
StackExchangeClusteringP2P	Clust.	v_measure	0.331	0.333	0.332	0.328	0.325
StackOverflowDupQuestions	Rerank.	map	0.468	0.462	0.466	0.449	0.452
SummEval	Summ.	spearman	0.316	0.321	0.318	0.312	0.314
TRECCOVID	Retr.	ndcg_at_10	0.379	0.363	0.383	0.345	0.370
Touche2020	Retr.	ndcg_at_10	0.174	0.176	0.177	0.168	0.173
ToxicConversationsClassification	Class.	accuracy	0.656	0.655	0.656	0.656	0.657
TweetSentimentExtractionClassification	Class.	accuracy	0.592	0.604	0.604	0.595	0.604
TwentyNewsgroupsClustering	Clust.	v_measure	0.444	0.449	0.450	0.434	0.419
TwitterSemEval2015	PClass.	similarity_ap	0.667	0.663	0.667	0.647	0.664
TwitterURLCorpus	PClass.	similarity_ap	0.851	0.847	0.851	0.831	0.844
		Average	0.552	0.549	0.550	0.524	0.529

Figure 15: MTEB evaluation results