Bidirectional Topic Matching: Quantifying Thematic Intersections Between Climate Change and Climate Mitigation News Corpora Through Topic Modelling

Raven Adam and Marie L. Kogler
Department of Environmental Systems Sciences / University of Graz, Austria
raven.adam@uni-graz.at marie.kogler@uni-graz.at

Abstract

Bidirectional Topic Matching (BTM) is a novel method for cross-corpus topic modeling that quantifies thematic overlap and divergence between corpora. BTM is a flexible framework that can incorporate various topic modeling approaches, including BERTopic, Top2Vec, and Latent Dirichlet Allocation (LDA). It employs a dual-model approach, training separate topic models for each corpus and applying them reciprocally to enable comprehensive cross-corpus comparisons. This methodology facilitates the identification of shared themes and unique topics, providing nuanced insights into thematic relationships. A case study on climate news articles illustrates BTM's utility by analyzing two distinct corpora: news coverage on climate change and articles focused on climate mitigation. The results reveal significant thematic overlaps and divergences, shedding light on how these two aspects of climate discourse are framed in the media.

1 Introduction

Topic modeling is widely used to analyze and structure large textual corpora (Churchill and Singh, 2022), with a key application being the identification of latent topics that experts can evaluate for quantitative insights (Grundmann, 2021). Beyond single-corpus analysis, topic modeling also facilitates comparisons across multiple corpora, enabling the examination of thematic similarities and differences (Bystrov et al., 2022).

In climate discourse research, cross-corpus methods can reveal how different aspects of climate change and mitigation are framed in the media. While corpus linguistics has traditionally applied similarity measures during corpus creation or selection, studies have demonstrated their value for discourse analysis (Taylor, 2018). Recent research has leveraged such approaches to compare narratives across policy debates, social media discussions, and news coverage in various contexts, including migration, elections, and economic development (Shaikina and Funkner, 2020; Bystrov et al., 2024; Hellwig et al., 2024; Taylor, 2018).

This study introduces Bidirectional Topic Matching (BTM), a novel method for cross-corpus topic modeling, to analyze thematic overlaps and distinctions in climate change and mitigation news articles. BTM identifies shared and corpus-specific topics, enabling both quantitative comparisons and deeper qualitative exploration of how these issues are framed.

Existing cross-corpus topic modeling approaches typically rely on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or language embedding models to compute topic similarities via cosine similarity (Carniel et al., 2022; Hellwig et al., 2024). Others merge corpora into a single model and analyze topic distributions separately (Wang et al., 2023). In contrast, BTM trains distinct topic models for each corpus and applies them reciprocally, allowing topics to be assigned across corpora. This approach enhances the identification of both shared and unique themes, providing deeper insights into the evolving discourse on climate change and mitigation.

2 Method

2.1 Topic modelling

BTM is a flexible framework for cross-corpus analysis that can incorporate various topic modeling approaches. For assessing corpus similarity, any method capable of inferring topics for new data is suitable. However, analyzing unique or corpus-specific topics requires a method that can identify intraclass outliers—documents that do not align with any topics generated by the chosen topic modeling approach. Language embedding-based methods, such as BERTopic (Grootendorst, 2022) or Top2Vec (Angelov, 2020), are particularly well-

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 208–217

suited for this purpose as they inherently support outlier detection. Traditional approaches like Latent Dirichlet Allocation (LDA), which assign a topic to every document, can also be adapted through post-processing techniques such as HDBSCAN (McInnes and Healy, 2017) or Local Outlier Factor (Breuniq et al., 2000) to identify outliers. Given BERTopic's state-of-the-art performance and its built-in outlier detection capabilities, this study demonstrates the application and efficacy of BTM using BERTopic as the underlying topic modeling approach.

BERTopic presents an innovative method for topic modeling, capitalizing on recent advancements in embedding models. Derived from Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), this approach involves the representation of documents as points within a high-dimensional vector space. In this space, each coordinate represents contextual information corresponding to the respective document. As a result, semantically analogous documents will be in proximity to each other. Subsequently, dimensionality reduction and clustering algorithms are employed to identify compact clusters of documents with shared thematic content. Each of these clusters can then be interpreted as individual topics that are found within the investigated collection of documents and are represented by a set of keywords that are most indicative of the underlying theme. An outlier refers to a document that cannot be assigned to any of the identified topics due to its lack of thematic similarity. This occurs when the document does not align well with any of the topics, often because it is too different or semantically distant from the other documents in the model. In BERTopic, both topics and outliers can be easily accessed and handled, where outliers are grouped together under an outlier topic, often with a special identifier like -1. As a final step, a class-based term frequency inverse document frequency measure (c-TFIDF) is applied to extract the most salient terms from each topic and create interpretable topic representations (Grootendorst, 2022).

2.2 Cross-Corpus Topic Assignment

For BTM, which is schematically depictured in Figure 1, two independent topic models are trained on two thematically related corpora, corpus 1 and corpus 2. Each model is used to identify the main themes within the respective corpus, generating topics T1 for corpus 1 and topics T2 for corpus 2. Individually, these native topic models provide a comprehensive understanding of the thematic structures specific to each dataset.

To explore thematic alignment between the corpora, each model was applied to the corpus, it was not trained on. For this, the semantic similarity between the document's embedding and the topic embeddings of the model trained on the other corpus was calculated. Specifically, each document in corpus 2 gets matched to a topic from T1, and each document in corpus 1 gets matched to a topic from T2, based on the highest similarity score. This process produced cross-corpus topic assignments, resulting in T12 (topics from T1 assigned to Corpus 2) and T21 (topics from T2 assigned to Corpus 1).

Subsequently, topic pairs are generated by assigning each document from one corpus to the most similar topic from the opposite corpus. Specifically, for each document, the topics assigned by the corpus 1 model (T11 and T12) and the topics assigned by the corpus 2 model (T22 and T21) are combined into cross-corpus topic pairs.

For a comprehensive cross-corpus analysis, both the main set of topics and outliers are considered. Outliers, while exhibiting atypical or low similarity scores within their own topic model, are included in the pairing process if they represent the highest similarity match for a document. Thus, topic similarity is calculated across all topics (0, 1, 2, ..., n), with outliers treated as an additional category (-1). This approach ensures that all thematical aspects are represented, even if the relationships involving outlier topics require further scrutiny in subsequent analyses. This becomes especially crucial when working with documents that are split into smaller units, like paragraphs, where certain sections may show unexpected topic associations, increasing the likelihood of outliers that require careful attention.

2.3 Cross-Corpus Topic Pair Analysis

The topic pairs from both corpora were analyzed through co-occurrence analysis to identify frequently paired topics between the two models. Specifically, we calculate how often each pair — composed of one topic from the corpus 1 model (T1, T12) and one from the corpus 2 model (T2, T21) — is assigned to the same document. The cross-topic co-occurrence is given by aggregation of these pairs across all documents. This process allows us to assess the frequency with which specific cross-corpus topic combinations occur together,

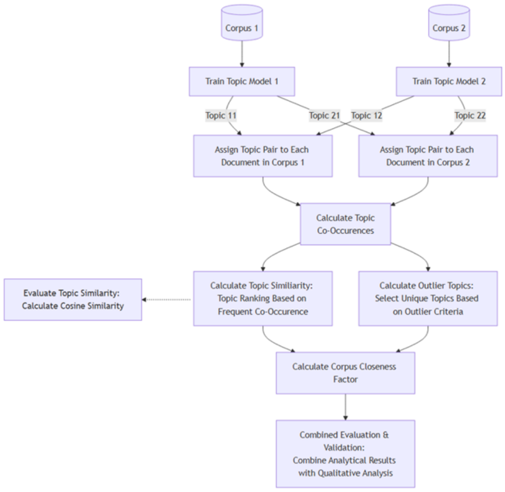


Figure 1: Schematic Outline of Bidirectional Topic Matching Procedures for Calculating the Thematic Closeness Factor of Corpus 1 and Corpus 2. Optional additional analysis of topic similarity may be conducted via cosine similarity.

providing insights into their thematic relationships. High-frequency pairs indicate topics from both models that were commonly associated with similar documents, reflecting thematic alignment between the corpora. Although the co-occurrence analysis itself remains undirected, focusing solely on the frequency of simultaneous topic occurrences within the documents, the subsequent exploration of relationships between topics from Corpus 1 and Corpus 2 is framed in a directed context. This directed approach enables a detailed investigation of the interactions and semantic linkages between the topics across the two corpora.

The interpretation of topic pairs helps clarify patterns of topic co-occurrence between corpora. High co-occurrence between a native topic and main cross-topics suggests strong thematic alignment, whereas alignment with smaller cross-topics indicates a more nuanced or niche connection. If a native topic aligns with outlier topics from the cross corpus, it may reflect themes unique to the native corpus. Similarly, when outlier topics from both corpora co-occur, it suggests a shared lack of thematic focus, while low co-occurrence between outliers is unexpected and may indicate inconsistencies in topic modeling or heterogeneity within the outlier topics.

3 Topic and Corpus Measures

For a corpus containing T native topics, a series of measures can be calculated to describe its relationship with a second corpus containing $\tilde{T}$ cross topics. A pairing strength is introduced as a quantitative measure of the degree of association between a topic from the native corpus and a topic from the cross corpus. This measure is based on the frequency of co-occurrence of the two topics within the same documents. For a topic pair $(t_i, \tilde{t}_j)$, where $t_i|i \in \{-1, \ldots, T\}$ belongs to the native topics and $\tilde{t}_j|j \in \{-1, \ldots, \tilde{T}\}$ belongs to the cross topics, the pairing strength $S(t_i, \tilde{t}_j)$ can be defined as:

$$S(t_i, \tilde{t}_j) = \frac{n(D_{ij})}{n(D_i)} \tag{1}$$

where $n(D_{ij})$ denotes the size (or cardinality) of the set of documents D_{ij} to which both topics t_i and $\tilde{t}_j$ are assigned. Respectively, $n(D_i)$ denotes the size of the set of documents D_i associated with the native topic t_i.

For the cross topics $\tilde{t}_j|j \in \{0, \ldots, \tilde{T}\}$, the pairing strength is referred to as topic closeness and represents the degree of alignment between each cross-topic and a specific native topic t_i. A special case of pairing strength involves the outlier topic $\tilde{t}_{-1}$ called topic uniqueness. Topic uniqueness quantifies the extent to which a native topic is distinct from the cross corpus. Native topics with a topic uniqueness value of 0.5 or higher are classified as unique topics.

3.1 Corpus Closeness and Corpus Uniqueness

Based on the topic closeness of all native topics, we define the corpus closeness C, which quantifies the overall thematic relatedness between the two corpora:

$$C = \frac{\sum_{i=0}^{T} \sum_{j=0}^{\tilde{T}} S(t_i, \tilde{t}_j)}{T} \tag{2}$$

as well as its weighted variant C_w, which gives higher importance to larger and more relevant native topics:

$$C_w = \frac{\sum_{i=0}^{T} n(D_i) \sum_{j=0}^{\tilde{T}} S(t_i, \tilde{t}_j)}{\sum_{i=0}^{T} n(D_i)} \tag{3}$$

Both closeness measures reflect the thematic overlap between the two corpora, while the weighted measure assigning greater significance to larger and thus more prominent topics within the native corpus. Generally, low closeness indicates that the two corpora are largely thematically independent. The difference $C_w - C = \theta; \theta \in [-1, 1]$ can be used to assess whether the relationship between the corpora is evenly distributed across all native topics or predominantly concentrated within a subset of native topics:

$$f(x) = \begin{cases} \theta \sim 1; & \text{corpus closeness is proportionally influences by larger native topics} \\ \theta \sim 0; & \text{corpus closeness is not influenced by native topic size} \\ \theta \sim -1 & \text{corpus closeness is proportionally influenced by smaller native topics} \end{cases} \tag{4}$$

The corpus uniqueness U and its weighted equivalent U_w are alternatives to the corpus closeness to indicate the level of independence between the corpora:

$$U = 1 - C = \frac{\sum_{i=0}^{T} S(t_i, \tilde{t}_{-1})}{T} \tag{5}$$

$$U_w = 1 - C_w = \frac{\sum_{i=0}^{T} n(D_i) \cdot S(t_i, \tilde{t}_{-1})}{\sum_{i=0}^{T} n(D_i)} \tag{6}$$

Here, $S(t_i, \tilde{t}_{-1})$ represents the topic uniqueness of each native topic. As with the corpus closeness factor, a high positive difference $U_w - U$ indicates that most of the corpus uniqueness is explained by larger native topics while a large negative difference sees most of it covered by smaller native topics.

3.2 Corpus Alignment

Both closeness and uniqueness fail to account for the specificity of topic matches and topic size distribution of the native corpus. The topic alignment strength $SA(t_i)$ of a native topic quantifies the concentration of topic closeness values with respect to the topics of the cross corpus. This indicates whether a native topic is associated with a single theme (focused) or to multiple themes (scattered) in the other corpus. To achieve this, the highest topic closeness of the native topic is selected:

$$\begin{aligned} SA(t_i) &= \max_{j \in \{0,\ldots,\tilde{T}\}} S(t_i, \tilde{t}_j) \qquad (7) \\ &= \max\{S(t_i, \tilde{t}_0), S(t_i, \tilde{t}_1), \ldots, S(t_i, \tilde{t}_{\tilde{T}})\} \qquad (8) \end{aligned}$$

A high topic alignment strength indicates that a native topic aligns with a single cross topic, whereas a low value suggests a wider variety of important pairings.

The corpus alignment A serves as an overall metric that captures the average alignment strength across all native topics. It quantifies whether the topic alignments between the two corpora are focused on specific topic pairs or spread over multiple combinations.

$$A = \frac{\sum_{i=0}^{T} SA(t_i)}{T} \tag{9}$$

$$A_w = \frac{\sum_{i=0}^{T} n(D_i) \cdot SA(t_i)}{\sum_{i=0}^{T} n(D_i)} \tag{10}$$

Here, the difference $A_w - A$ is useful to indicate whether the distribution of topic alignment strength is skewed towards larger or smaller native topics.

By comparing the corpus uniqueness factor U and the corpus alignment factor A, we identify three key relationships between corpora. Low uniqueness and low alignment indicate thematic overlap, with the cross corpus exploring similar topics in greater depth or from multiple perspectives. Low uniqueness and high alignment suggest that the corpora are closely related, likely subsets of a larger parent corpus. High uniqueness and low alignment imply that the corpora are largely independent, as many topics in the native corpus are not present in the cross corpus. A scenario with both high uniqueness and high alignment is not possible due to their inherent relationship.

4 Validation through related Methods

Since both topic models are generated from the same embedding model, the resulting embedding vectors for each topic are located in the same vector space. Therefore, to validate the effectiveness of the proposed method, we introduced an additional analysis by measuring the cosine similarity between the topic embeddings of the two BERTopic models. In this validation process, cosine similarity scores were first calculated between the topic embeddings of the corpus 1 and corpus 2 models to quantify the semantic overlap between their topics. Higher cosine similarity scores indicated greater alignment between topics. These scores were then compared to the distribution of observed topic pairs, with the goal of finding the most similar topics across the corpora. To assess the consistency between the two methods, Cohen's kappa was calculated, providing a measure of agreement between the cosine similarity-based approach and the topic pair distribution.

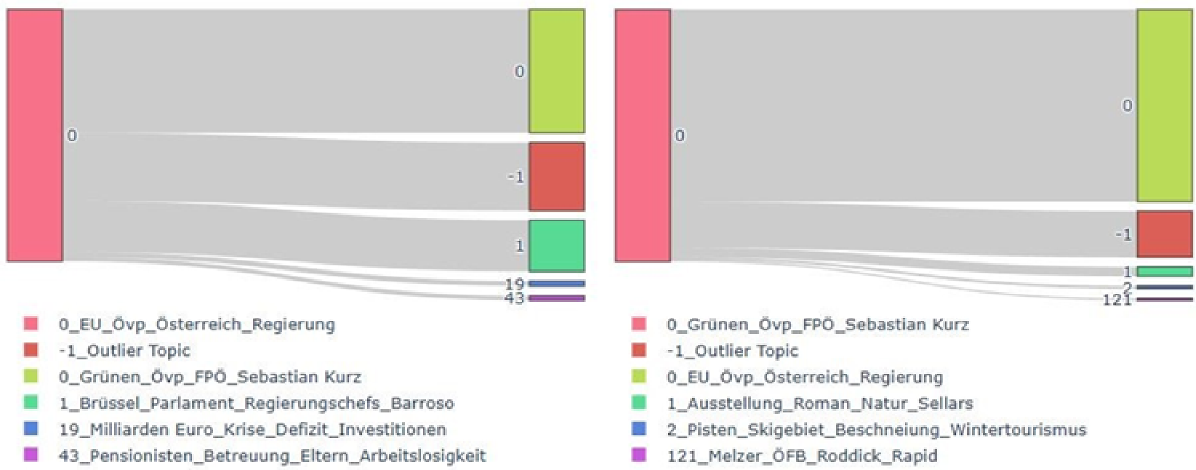


Figure 2: Left side – The largest native topic from corpus 1 along with the five most prominent cross topic pairs from corpus 2. They gray area indicates the pairing strength for each pair. Right side – The largest native topic from corpus 2 along with the five most prominent cross topic pairs from corpus 1. They gray area indicates the pairing strength for each pair.

5 Case Study Climate News

5.1 Dataset

To showcase BTM, two sets of digitized print articles were extracted from the WISO database that provides a repository for online newsarticles in the German-speaking region. According to Adam, Scholger, and Kogler (2023), the regional climate debate is characterized by two largely independent subject areas: climate change, which encompasses information on natural and physical impacts, dangers, and risks, and climate mitigation, which focuses on actions, socio-economic strategies, and technological solutions. The search terms climate change (“klimawandel*” where the asterisk serves as a wildcard symbol that matches any suffixes or word endings attached to the German root word "klimawandel") and climate action (“klimaschutz*”) were used to create the climate change dataset (corpus 1) and the climate action dataset (corpus 2), respectively. The investigated period spans from 2002 until 2022 and includes 21.753 articles in corpus 1 and 20.135 articles in corpus 2, with an overlap of 3.111 articles.

To account for the limited encoding length of embedding models, all articles were split into smaller parts of up to 150 words, which corresponds to the average length of German paragraphs (Altpeter et al., 2015). This was done with the help of the *gsd model* available in the stanza library (Qi et al., 2020). The final dataset therefore consisted of 124.500 paragraphs.

Both BERTopic models were trained based on the *German Semantic STS V2* embedding model. For corpus 1 a topic model consisting of 122 topics was generated, while corpus 2 produced a topic model with 88 topics.

6 Results

6.1 Case Study

6.1.1 Topic Pairs and primary Relationships between Topics

Tables 1 and 2 provide qualitative evidence supporting BTM’s ability to identify meaningful relationships between topics across corpora. By examining paired topics, corpus-specific nuances emerge. For example, a comparison of topics focused on forests and glaciers reveals differences in thematic emphasis: Corpus 1 highlights specific results of climate change, such as increased bark beetle infestations and rockfalls in the Alps, while Corpus 2 emphasizes the state of forests or national parks and the impact of climate change on alpine temperatures. This capacity to reveal varying degrees of specificity allows researchers to understand how distinct datasets prioritize or converge on shared themes. Such insights are critical for comparative discourse analyses, such as political communication or cross-cultural studies.

6.1.2 Subpairing Topics – Quantifying Secondary Thematic Relationships

Whether individual topics are directly shared between corpora or whether one corpus discusses certain topics more diversely can be analyzed us-

Native Topics Corpus 1 (T1)	Cross Topics Corpus 2 (T12)	SA
EU ÖVP Austria Government	Greens ÖVP FPÖ Sebastian_Kurz	0.44
Trees Bark_Beetle Federal_Forestry Spruce	Woods Hectare Federal_Forestry National_Park	0.60
Fridays Greta_Thunberg Streets Youths	Fridays Greta_Thunberg Movement Humans	0.69
Glacier Alps Rockfall Dachstein	Degree Glacier Temperatures Climate_Change	0.58
Diesel Electric_Cars Vehicles Automobile_Industry	Electric_Cars Vehicles BMV Diesel	0.41

Table 1: Five native topics of corpus 1 along with their respective main cross topic pair from corpus 2 and topic alignment strength SA (highest pairing strength). Each topic is represented by four topic words or phrases (connected with an underscore), which is the standard output of BERTopic. The topic representations were translated from German to English.

Native Topics Corpus 2 (T2)	Cross Topics Corpus 1 (T21)	SA
Greens ÖVP FPÖ Sebastian_Kurz	EU ÖVP Austria Government	0.70
Brussels Parliament Head_of_Government Barroso	EU ÖVP Austria Government	0.61
Renovation Residential_Construction Housing_Subsidies Buildings	Passive_House Residential_Construction Energy_Efficiency Real_Estate	0.31
ÖBB Million_Euro Truck Commuter	ÖBB Vienna Mobility Means_of_Transport	0.68
Baerbock Merkel CSU Greens	Laschet Baerbock Greens Coalition	0.74

Table 2: Five native topics of corpus 2 along with their respective main cross topic pair from corpus 1 and topic alignment strength SA (highest pairing strength). Each topic is represented by four topic words or phrases (connected with an underscore), which is the standard output of BERTopic. The topic representations were translated from German to English.

ing topic alignment strength, as shown in Tables 1 and 2. For instance, the politics topic in Corpus 1 exhibits a moderate topic alignment strength of 0.44. This indicates that several topics from Corpus 2, beyond the most similar cross-topic, address relevant aspects of this native topic. The left side of Figure 2 visually showcases this distribution across different cross-topic pairings. This suggests that political discourse is more granular in Corpus 2, allowing its topic model to recognize distinctions within documents assigned to a single topic in Corpus 1.

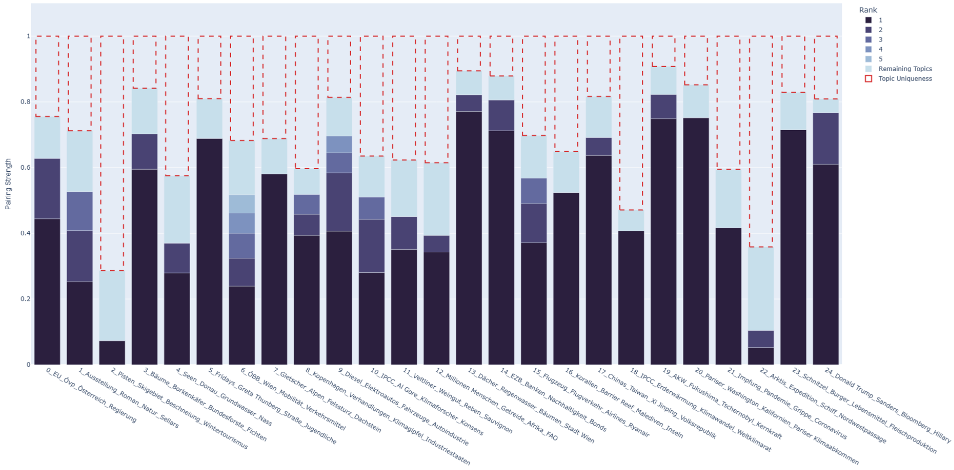
Conversely, Table 2 reveals that both national and EU-level politics topics in Corpus 2 exhibit high topic alignment strength with the same politics topic in Corpus 1. This supports the hypothesis that political discourse in Corpus 2 is more detailed, encompassing multiple perspectives that align with a broader theme in Corpus 1.

A broader overview is provided in Figure 3, which illustrates the pairing strength composition for the 25 largest topics in each corpus. For most native topics, the most similar cross-topic alone does not account for the majority of topic closeness. This highlights thematic asymmetries, where one corpus tends toward generality while the other emphasizes specificity. Such analyses are instrumental in uncovering where thematic overlaps or divergences occur, enabling nuanced interpretations of the data

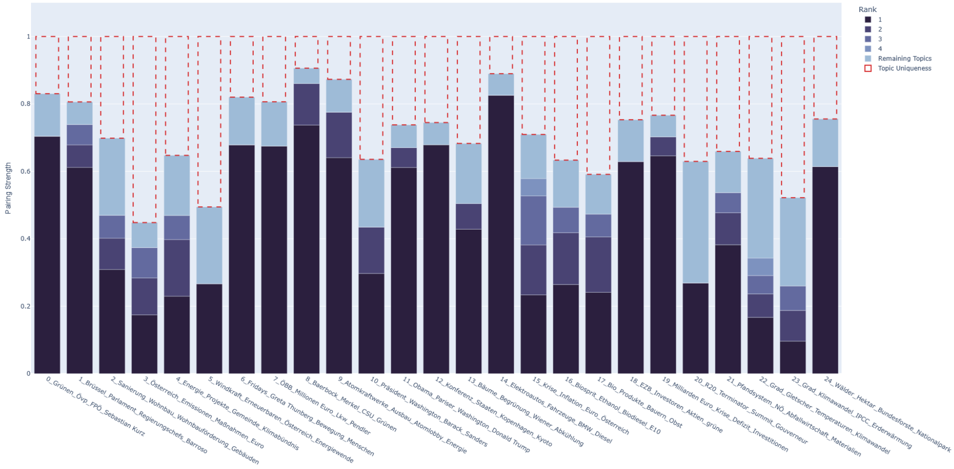
6.1.3 Identifying Unique Topics

One of BTM's most compelling features is its ability to identify topics unique to each corpus. This is achieved by extracting topics with a topic uniqueness value above 0.5. In this case study, 23 unique topics were identified in Corpus 1, while Corpus 2 contained 15 unique topics.

Table 3 illustrates examples of unique topics from each corpus. Corpus 1 focuses on science communication and geographic impacts, such as water supply, while Corpus 2 emphasizes actionable measures, including renewable energy and local initiatives. Such differentiation is especially valuable for corpora with overlapping themes, as it enables researchers to discern distinct areas of fo-



(a)



(b)

Figure 3: The pairing strength composition for the 25 largest native topics. The shading of the bars indicates the ranking of the topic pairing strengths, where the most prominent pair is represented by the darkest color. Topic pairs with a pairing strength below 0.05 were merged into the “remaining topic” category. The outlier topic pairing strength or topic uniqueness is indicated by the red dashed bars. a): Corpus 1. b): Corpus 2.

cus. For example, in interdisciplinary studies, this capability bridges gaps between problem-oriented and solution-oriented approaches, fostering more comprehensive analyses.

6.1.4 Corpus Level Relationship

Table 4 reveals that both corpora exhibit notable distinctions, with approximately one-third of the content in each corpus not described by the other. Both show a corpus uniqueness factor of 0.34, indicating a significant level of thematic independence. The corpus closeness factor of 0.66 suggests major thematic overlaps, while the low difference between weighted and general corpus uniqueness factors (< 0.1) implies that neither corpus is skewed toward unique topics of particular sizes. However, Corpus 2 displays slightly more pronounced topic uniqueness in smaller topics compared to Corpus 1.

Similarly, both corpora have comparable corpus alignment factors (0.45 for Corpus 1 and 0.44 for Corpus 2). The minor influence of native topic sizes indicates that alignment is not disproportionately driven by larger topics. Together, these metrics suggest that while the corpora share substantial thematic overlap, they focus on different thematic subsets in more detail. This is consistent with the low corpus uniqueness and low corpus alignment case, where native topics frequently pair with multiple relevant cross-topics, as observed in Figure 3.

Corpus 1 Unique Topics	Corpus 2 Unique Topics
Slopes Ski_Area Snow_Making Wintertourism	Austria Emissions Measures Euro
Lakes Donau Groundwater Water	Energy Project Municipality Climate_Alliance
IPCC Al_Gore Climate_Researcher Consensus	Wind_Power Renewable Austria Energie_Transition

Table 3: Selection of three unique native topics from corpus 1 and corpus 2 respectively based on a topic uniqueness above 0.5.

Native Corpus	C	$C_w - C$	U	$u_w - U$	A	$A_w - A$
Corpus 1	0.66	0.02	0.34	-0.02	0.45	-0.01
Corpus 2	0.66	0.04	0.34	-0.04	0.44	0.04

Table 4: Values for the corpus closeness factor C, the corpus uniqueness factor U, the corpus alignment factor A and the difference between the three factors and their respective weighted variants for corpus 1 and corpus 2.

6.2 Validation - Comparison with Cosine Similarity

We demonstrate the agreement between BTM and cosine similarity-based methods for climate news articles to highlight the validity of the proposed approach. When identifying the most similar topic from corpus 2 for each topic in corpus 1, Cohen's kappa was calculated at 0.75. Conversely, when determining the most similar topic from corpus 1 for each topic in corpus 2, Cohen's kappa increased to 0.81. These values reflect a strong level of agreement, affirming the reliability of BTM (Mchugh, 2012).

Discrepancies between BTM and cosine similarity approaches were most evident when BTM assigned the outlier topic as the closest match. Since this topic encompasses documents that do not fit into any defined clusters, its inclusion is inherently challenging for methods relying solely on cosine similarity. Beyond the outlier topic, the remaining discrepancies (approximately 20%) lacked clear evidence favoring one method over the other, suggesting that both approaches offer comparable utility for calculating topic similarity.

7 Discussion and Conclusion

BTM provides a robust framework for cross-corpus topic modeling. By leveraging BERTopic's interpretable topic representations and employing reciprocal topic assignments, BTM facilitates a nuanced exploration of thematic relationships across corpora. This approach not only captures shared topics but also highlights unique themes, offering a comprehensive lens through which to analyze corpora with overlapping or divergent thematic structures.

7.1 Methodological Contributions

BTM addresses key limitations in traditional cross-corpus topic modeling approaches. By training separate topic models for each corpus and applying them reciprocally, BTM ensures that each model's native structure is preserved while enabling cross-corpus comparisons. This dual approach allows for the identification of both shared and unique topics, a capability that is particularly valuable in interdisciplinary or comparative discourse studies.

Validation through cosine similarity underscores the reliability of BTM. Strong agreement between BTM and cosine similarity-based methods (Cohen's kappa scores of 0.75 and 0.81) demonstrates the robustness of the approach, while the discrepancies observed with outlier topics highlight areas where BTM's methodological strengths are most apparent. These findings suggest that BTM can serve as a reliable alternative or complement to existing methods, particularly for datasets with significant thematic variability.

7.2 Insights from the Case Study

The application of BTM to climate news articles revealed meaningful thematic distinctions and overlaps between two corpora focused on climate change and climate action. The results demonstrate that while both corpora share substantial thematic overlap (corpus closeness factor of 0.66), they also exhibit notable differences, with approximately one-third of the content in each corpus being unique (corpus uniqueness factor of 0.34).

Corpus 1 prioritizes broad environmental and scientific discussions, such as the geographic impacts of climate change, while Corpus 2 focuses on actionable measures like renewable energy and local initiatives. This differentiation underscores the value of BTM in identifying thematic nuances that may be overlooked by less granular methods. Moreover, the ability to quantify topic alignment and uniqueness provides a structured way to assess thematic relationships, facilitating more targeted qualitative analyses.

8 Limitations

There are a few notable limitations in the suggested approach. First of all, BTM provides direction dependent results. Comparing Corpus 1 with Corpus 2 can lead to different results than comparing Corpus 2 with Corpus 1. For example, if Corpus 2 were to be a highly specific sub-corpus of Corpus 1. In this case, Corpus 1 would exhibit high Uniqueness values, as only a limited number of its native topics would be covered by Corpus 2. However, Corpus 2 would have low Uniqueness as all of its native topics are present in Corpus 1.

Secondly, the presented case study uses the same embedding model for both corpora. While this is necessary to compare the results with cosine similarity, there are cases where it might be preferable to use different embedding models for each corpus. Especially when domain-specific models are available such as in the medical or financial domain. BTM can, theoretically, still be employed in such a case, it is however unclear how valid the results would be. Such an investigation would be an important aspect of future research.

A third limitation is that using topic merging methods after creating topic models will result in different corpus level measures than using unmerged topics. The topic level measures of a merged topic will be the averages calculated from the topic level measures of each individual topic that was used to create the merged topic. And as the corpus level measures are either weighted or unweighted averages of the used topic level measures, averaging some of them beforehand will naturally change the final results.

9 Further Research

Future research could apply BTM to dissect the complex interplay between scientific understanding and policy formulation. For instance, a systematic comparison of academic literature on specific climate solutions, such as carbon capture technologies or nature-based solutions, with corresponding governmental policy documents or legislative proposals could quantitatively reveal how scientific findings are translated, prioritized, or re-framed within policy-making arenas (Ibarra et al., 2022). Similarly, BTM offers a robust methodology to analyze the critical interface between expert communication and public discourse. By comparing outputs from climate science organizations, like IPCC summaries or national climate assessments, with the vast textual data generated on social media platforms or in public commentary on news articles, researchers could identify unique public concerns, pinpoint areas of scientific misunderstanding, or highlight divergent thematic emphases, thereby informing the development of more effective and resonant climate communication strategies.

Furthermore, BTM can facilitate nuanced comparisons across diverse geopolitical and ideological landscapes. It could be employed to systematically examine climate narratives within Nationally Determined Contributions (NDCs) submitted by developed versus developing nations, or to contrast climate impact reporting styles and thematic priorities between media outlets in the Global North and Global South (Hase et al., 2021). Such analyses could illuminate shared thematic ground alongside areas of significant contention or differing national priorities, which is crucial for international climate negotiations and cooperation.

Beyond governmental and public spheres, BTM can also shed light on corporate engagement with climate change. Applying the framework to analyze corporate sustainability or Environmental, Social, and Governance (ESG) reports across various industry sectors, or between companies with different stated climate commitments, could identify common and unique themes related to perceived climate risks, adopted mitigation strategies, and planned adaptation efforts (Dahl and Fløttum, 2019). Through these varied applications, BTM promises to provide researchers with a powerful tool for a deeper, more quantified understanding of the multifaceted and evolving discourses surrounding climate change, its impacts, and the global response.

References

Elke Karin Altpeter, Tobias Marx, Nhung Xuan Nguyen, Aline Naumann, and Susanne Trauzettel-Klosinski. 2015. Measurement of reading speed with standardized texts: a comparison of single sentences and paragraphs. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 253(8):1369–1375.

Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics. pages 1–25.

David Meir Blei, Andrew Yan-Tak Ng, and Michael Irwin Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Markus M. Breuniq, Hans Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: Identifying density-based local outliers. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 29(2):93–104.

Victor Bystrov, Viktoriia Naboka, Anna Staszewska-Bystrova, and Peter Winker. 2022. Cross-Corpora Comparisons of Topics and Topic Trends. *Journal of Economics and Statistics*, 242(4):433–469.

Victor Bystrov, Viktoriia Naboka-Krell, Anna Staszewska-Bystrova, and Peter Winker. 2024. Comparing Links between Topic Trends and Economic Indicators in the German and Polish Academic Literature. *Comparative Economic Research. Central and Eastern Europe*, 27(2):7–28.

Théophile Carniel, Leo Cazenille, Jean Michel Dalle, and José Halloy. 2022. Using natural language processing to find research topics in Living Machines conferences and their intersections with Bioinspiration & Biomimetics publications. *Bioinspiration and Biomimetics*, 17.

Rob Churchill and Lisa Singh. 2022. The Evolution of Topic Modeling. *ACM Computing Surveys*, 54(10).

Trine Dahl and Kjersti Fløttum. 2019. Climate change as a corporate strategy issue. *Corporate Communications: An International Journal*, 24(3):499–514.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Reiner Grundmann. 2021. Using large text news archives for the analysis of climate change discourse: some methodological observations. *Journal of Risk Research*, 0(0):1–13.

Valerie Hase, Daniela Mahl, Mike Steffen Schäfer, and Tobias Keller. 2021. Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). *Global Environmental Change*, 70(December 2020).

Nils Constantin Hellwig, Jakob Fehle, Markus Bink, Thomas Schmidt, and Christian Wolff. 2024. Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election. *International Journal of Speech Technology*.

Cecilia Ibarra, Guadalupe Jiménez, Raúl O'Ryan, Gustavo Blanco, Luis Cordero, Ximena Insunza, Pilar Moraga, Maisa Rojas, and Rodolfo Sapiains. 2022. Scientists and climate governance: A view from the south. *Environmental Science & Policy*, 137:396–405.

Mary L Mchugh. 2012. Interrater reliability : the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Leland McInnes and John Healy. 2017. Accelerated Hierarchical Density Based Clustering. In *IEEE International Conference on Data Mining Workshops, ICDMW*, pages 33–42.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 101–108.

Alevtina A. Shaikina and Anastasia A. Funkner. 2020. Medical Corpora Comparison Using Topic Modeling. *Procedia Computer Science*, 178:244–253.

Charlotte Taylor. 2018. Similarity. In Charlotte Taylor and Anna Marchi, editors, *Corpus Approaches to Discourse*, chapter 2, pages 19–37. Routledge, London.

Zhongyi Wang, Jing Chen, Jiangping Chen, and Haihua Chen. 2023. Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 129(11):7359–7384.