From Semantic to Thematic analysis: Extracting Research trends and topics from Text Data

Husain Vaghjipurwala**, Katharina Borgmann*, Maryam Hosseinzadeh*, Agota Barabas*, Jörg Noennig*
*HafenCity Universität. Hamburg
**husain.vaghjipurwala@hcu-hamburg.com

Abstract
Topic modelling and extraction is not a new field in synthesis-research, but with the current advancements in LLMs and Transformer-based models, such processes can as well be advanced with better accuracy and efficiency. In this paper, we explore the application of Transformer based methods, firstly to validate the domain experts based manual identification of topics and secondly, to create a benchmark method for machine-based extraction of Topics and Themes related to sustainable urban development. The research uses a corpus of documents, including Project proposals and Project profiles, from ten projects in the field of Sustainable Development. The process builds upon established methodology whereby the tokenized text is used to create embeddings with BERT (Bidirectional Encoder Representations from Transformers) to create a vector space. As part of the new proposed architecture, the embeddings are then clustered and labelled according to their cosine similarity with domain expert themes, which allows a comprehensive view of themes/topics and sub-topics.

Keywords: Topic modelling, BERT based transformers, Sustainable development, Semantic clustering, Thematic clustering

1. Introduction

1.1. Background

Sustainable urban development policies are essential to address global challenges such as climate change, rapid urbanization, and population growth. Cities, housing over half of the world's population, contribute approximately 70% of global carbon emissions and 60% of resource consumption (United nations, 2019). With urban populations projected to increase by 2.5 billion by 2050, primarily in Asia and Africa (UN-Habitat, 2020), the strain on resources and ecosystems will intensify, necessitating effective sustainable policies. Sustainable urban policies aim to mitigate environmental impacts while enhancing quality of life, resilience, and inclusivity. Climate change exacerbates extreme weather events in urban areas, posing risks to infrastructure and populations (IPCC, 2021). Effective policies must incorporate adaptation and mitigation strategies, such as green infrastructure and renewable energy, to reduce risks and vulnerability.

Furthermore, sustainable development addresses socio-economic disparities by promoting equitable access to resources and green spaces in densely populated areas (OECD, 2020). Policies encouraging resource efficiency and pollution control are pivotal for urban resilience and long-term viability. Trend analysis is critical for monitoring policy changes and literature themes in sustainable urban development. It detects patterns to identify significant shifts and emerging priorities, providing insights into policy evolution in response to societal needs and environmental pressures (Olawumi and Chan, 2018). Advancements in Artificial Intelligence (AI), particularly Natural Language Processing (NLP) and Large Language Models (LLMs), have enhanced trend analysis by automating data extraction from vast textual datasets. These tools enable real-time analyses, improving the relevance of policy insights (Levitt, 2018; Kashi and Shah, 2023). Thus, trend analysis is invaluable in navigating the complexities of sustainable urban development and ensuring that policies remain responsive and adaptive.

1.2. Rationale

Qualitative trend analysis offers significant advantages by enabling policymakers to detect shifts in priorities, assess policy impacts, and anticipate future needs. It captures nuanced changes that quantitative methods may overlook, such as emerging themes in resilience, social equity, and environmental sustainability within urban policies. As cities adopt sustainability-focused frameworks, qualitative trend analysis reveals the evolution of policy goals in response to societal and environmental demands, supporting more adaptive policymaking (Olawumi

and Chan, 2018). It also provides insights into the long-term efficacy of sustainable development policies. However, sustainable urban development involves complex interactions between social, environmental, and economic systems. Rapid urbanization and climate change present challenges not easily captured by numerical data. Moreover, the growing volume of grey and academic literature—such as sustainability reports, policy frameworks, and research papers—creates information overload, hindering decision-making. Addressing these challenges, this paper reports on the experiment undertaken to explore the application of Trend analysis in Sustainable Urban Development. The experiment aims to bring for the methodology and adaptation employed for extracting Topics from texts and converting them to themes for trend monitoring.

The paper begins with a comprehensive literature review of the scope of Topic modelling, followed by the experiment conducted with a modified version of the Transformer based models. The research is conducted within the context of the SURE Facilitation and Synthesis Research project, focusing on sustainable urban development in Southeast Asia and China.

2. Literature review

2.1. Evolution of Topic modelling

Topic modelling is a popular technique used for text analysis that provides an automatic approach for coding a collection of documents into a set of meaningful categories, known as topics (Thompson and Mimno, 2020; Lore, Harten and Boeing, 2024; El-Gayar et al., 2024). It is considered an unsupervised machine learning technique primarily used to uncover hidden patterns and structures from large sets of unstructured data (El-Gayar et al., 2024). The goal of topic modelling is to organize, search, and summarize unstructured data, making it a valuable tool in various fields (El-Gayar et al., 2024). Topic models aim to identify hidden topical patterns within a collection of texts (Egger and Yu, 2022).

Different techniques and approaches have been used to extract topics using topic modelling (El-Gayar et al., 2024). Among the most established and frequently adopted techniques is Latent Dirichlet Allocation (LDA) (Egger and Yu, 2022; El-Gayar et al., 2024). Introduced by Blei et al., LDA is a probabilistic generative model (Churchill and Singh, 2022; Egger and Yu, 2022; El-Gayar et al., 2024) that aims to discover the topics in a document based on the words it contains. It views each document as a mixture of topics, and each topic as a distribution over words (Churchill and Singh, 2022; Lore, Harten and Boeing, 2024; Kapoor et al., 2024). LDA became a widely used algorithm for extracting meaningful topics from various types of text data (Churchill and Singh, 2022; Egger and Yu, 2022; El-Gayar et al., 2024). It is often considered a classic or traditional approach to topic modelling (Mu et al., 2024; Kapoor et al., 2024).

2.2. The promise of Transformer based models

Transformer based Large Language Models (LLMs) offer a significant promise for the field of topic modelling, presenting a compelling alternative to traditional methods like Latent Dirichlet Allocation (LDA) (Mu et al., 2024). The promise of these models for topic modelling stems from their advanced language understanding capabilities and flexible application (Mu et al., 2024). Classic topic modelling approaches, such as LDA, may lack semantic understanding and struggle to capture nuanced semantics and contextual understanding (Mu et al., 2024; Kapoor et al., 2024). LLMs, on the other hand, leverage deep neural networks to learn rich, contextual representations (Kapoor et al., 2024) and can capture subtle semantic, pragmatic features, nuances, and subtle thematic undertones in language (Mu et al., 2024; Kapoor et al., 2024). This enables a richer and more detailed categorization of topics (Egger and Yu, 2022; Mu et al., 2024).

LLM-based models can generate more cohesive and interpretable topics compared to LDA (El-Gayar et al., 2024).

BERTopic and RoBERTa, representing Transformer-based models, generated more cohesive topics in a case study, where associated words were closely related thematically (El-Gayar et al., 2024). This increased cohesiveness makes the topic labelling process easier and faster compared to topics generated by LDA, which often contain a large number of focused words. While LDA may achieve a higher labelling rate, the process is more challenging due to the word count and potential overlapping labels (El-Gayar et al., 2024). As opposed to the traditional approach, LLMs or Transformer based models can produce topics with superior clustering and improved semantic coherence.

Unlike traditional methods that often require text preprocessing (like stemming and lemmatization) and post-processing like manual topic labelling and interpretation, LLMs can also aid in reducing the need for human involvement (Mu et al., 2024). They can automatically extract (or generate) topics and provide human-understandable explanations. LLMs can also be used off the shelf to automatically assess the coherence of topic word collections and their ratings have shown to highly correlate with human annotations (Mu et al., 2024). Studies have shown that using LLM-extracted key phrases for clustering can reduce noise, avoid manual data cleaning, and result in semantically concentrated clusters (Kapoor et al., 2024).

2.3. Semantic clustering

Transformer based models, such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pre-training Approach), utilize contextualized word representations (Thompson and Mimno, 2020; Churchill and Singh, 2022). Unlike earlier models that assigned a single vector to each word regardless of its usage, BERT generates embeddings that capture the dynamic meaning of words in their specific context (Thompson and Mimno, 2020; Churchill and Singh, 2022; Lore, Harten and Boeing, 2024; El-Gayar et al., 2024). This capability allows BERT to understand nuanced phrases and word relationships by considering the surrounding sentence (Thompson and Mimno, 2020; Lore, Harten and Boeing, 2024). These contextual embeddings provide a better representation and performance compared to traditional word embedding techniques, especially when working with large sparse matrices in NLP (El-Gayar et al., 2024).

One significant application of these contextual embeddings is in semantic clustering. This involves grouping textual data (like words, tokens, or documents) based on their meaning or semantic similarity, as captured by the embeddings (Thompson and Mimno, 2020). Clustering token-level contextualized word representations can produce output that shares many similarities with traditional topic models (Thompson and Mimno, 2020; Churchill and Singh, 2022). Since each word type is modelled as a single vector in vocabulary-level embeddings, those methods cannot easily account for polysemy (words with multiple meanings) or leverage local context for disambiguation (Thompson and Mimno, 2020). In contrast, token-level clustering, grounded in specific documents, supports a wide range of analysis techniques (Thompson and Mimno, 2020).

A key example of a BERT-based model implementing semantic clustering for topic extraction is BERTopic (Egger and Yu, 2022; Sy et al., 2024; El-Gayar et al., 2024). BERTopic uses BERT embeddings and a technique called class-based TF-IDF (c-TF-IDF) to generate dense clusters of documents and extract topic representations (Egger and Yu, 2022; Sy et al., 2024; El-Gayar et al., 2024). The process begins by representing each document as a set of numerical values using BERT embeddings and these embeddings capture contextual information about words (Sy et al., 2024). After obtaining the high-dimensional embeddings, dimensionality reduction techniques like UMAP can be applied to transform them into a lower-dimensional space, which helps in data visualization and clustering (Egger and Yu, 2022; Sy et al., 2024). Clustering algorithms, such as HDBSCAN, are then applied to the lower-dimensional space to identify dense clusters of data points, effectively grouping documents into topics (Egger and Yu, 2022; Sy et al., 2024).

Other approaches also leverage BERT embeddings for clustering. For instance, clustering tokens based on their contextual vectors drawn from BERT using k-means clustering has been proposed as a method to produce topics (Thompson and Mimno, 2020; Churchill and Singh, 2022). Clustering vocabulary-level embeddings has been shown to produce semantically related word clusters (Thompson and Mimno, 2020). Furthermore, hybrid techniques combine manual labelling with BERT-based models, using manually created labels as training data to fine-tune a pre-trained BERT model for sequence classification tasks to identify topics (Lore, Harten and Boeing, 2024). LLMs, including BERT-based ones, combined with clustering techniques can provide a scalable approach to identifying themes and patterns in unstructured text data by leveraging LLMs' semantic understanding and clustering algorithms' ability to group similar responses (Kapoor et al., 2024). BERT-based models provide rich, contextual embeddings of text, enabling semantic clustering approaches that group words, tokens, or documents based on their context-aware meaning (Thompson and Mimno, 2020; Sy et al., 2024).

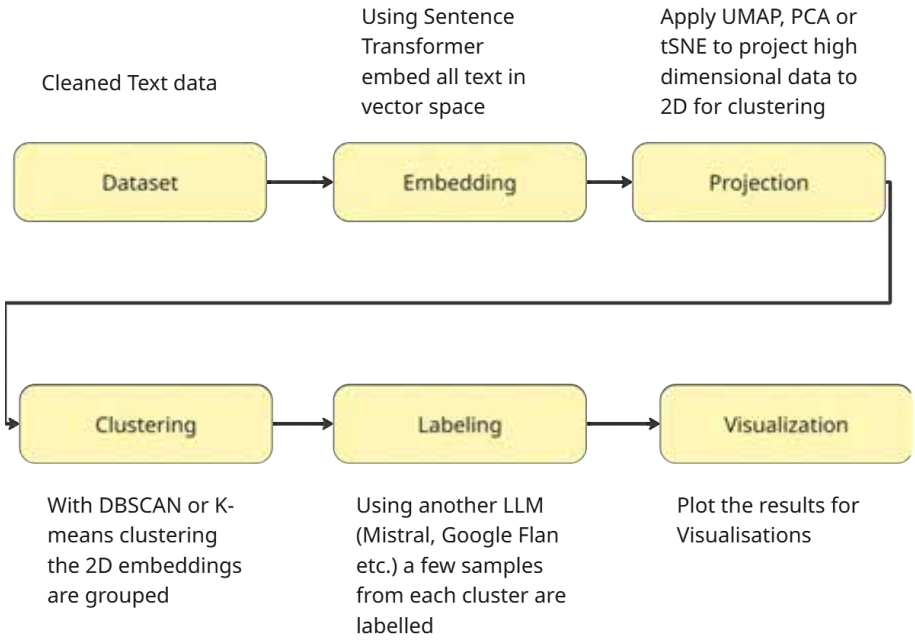


Figure 1 Typical BERT based topic modelling architecture, source: Author

2.4. Thematic Analysis

The rise of machine-based Topic modelling and analysis of text documents has provided immense help in understanding the semantic dimensions of text. This, however, is limited in some respects especially when compared with a more abstract level organization or classification of the said text. Such higher-level abstraction is where the Domain Expertise of humans becomes more valuable. Added to this, are the questions of traceability, reproducibility of knowledge and recognition of patterns over time that have been some of the obvious limitations of Machine based topic – modelling.

At the same time, Qualitative researchers have always focused upon the Thematic Analysis of texts as a qualitative analytic method (Clarke and Braun, 2017). Thematic analysis is defined as a method for identifying, analysing and interpreting patterns of meaning ('themes') within large text data (Clarke and Braun, 2017). According to Boyatzis (Boyatzis, 1998), such thematic analysis minimally organises and describes the text data in detail, but frequently goes further to interpret various aspects of the topic. Although it is one of the foundational method for qualitative analysis, there is no clear agreement about what thematic analysis is and how you go about conducting it (Braun and Clarke, 2006). Different fields of study, have approached the process of qualitative the-

matic analysis differently, but for the purposes of this research, we lean heavily on the fundamental work done by Braun & Clarke since the early 2000's (Braun and Clarke, 2006; Clarke and Braun, 2017; Braun and Clarke, 2021).

A simplistic structure proposed by (Clarke and Braun, 2017) for conducting a Thematic Analysis, involves Six main phases: Familiarising with data, generating/ extracting initial codes, searching for themes, reviewing themes, defining and labelling themes and lastly, reporting on the themes. In such a process, the researcher is not a passive participant but rather actively directs the identification of what's relevant in the text and what's not. As Braun and Clarke also argue that the notion that Themes are "Emerging" or "discovered", as a passive account of the process of analysis, denies such a role of the researcher. If anything, they conclude, Themes reside in our heads from our thinking about our data and creating links as we understand them (Braun and Clarke, 2006). There are usually two approaches to identifying themes as discussed below

2.4.1. Inductive versus Theoretical approach to Thematic analysis

In an Inductive approach, the themes are identified in a bottom-up way and are strongly linked to the text data under study. They do not necessarily need to reflect the researcher's theoretical interest in the subject (Braun and Clarke, 2006). Such an analysis involves coding or extracting key phrases from the text without any preconceptions of the researcher nor a pre-existing coding framework. As one can imagine, the Theoretical approach is driven by a researcher's analytical interest in the subject matter. The approach aligns with coding for a specific research question which is decided prior to the analysis (Braun and Clarke, 2006). In the context of Urban Development, this would be, for example, extracting phrases or coding the text on themes related to Ecosystem Services which deals with nature-based solutions.

Thematic analysis has however remained a rarely acknowledged method owing to its lack of transparency and lengthy ritual of execution. The flexibility of Thematic Analyses also allows for its drawbacks when compared to other qualitative approaches like Grounded Theory and Discourse analyses which could be considered 'Off- the -Shelf' methodologies.

2.5. Research Design

In qualitative research, a "trend" refers to a recurring pattern or tendency within data that reveals shifts over time, particularly in themes, priorities, or policy directions. A "trend signal" represents specific indicators or events that mark the emergence of these trends, such as policy revisions, increased funding in certain areas, or shifts in public discourse (Olawumi and Chan, 2018). These signals are vital as they enable researchers to identify the early stages of a trend, offering foresight into possible future developments. Key elements of trend analysis include "trend extrapolation" where researchers project current trends into the future to hypothesize potential directions (Sneegas et al., 2021). "Trend impact analysis" involves assessing the likely effects of these trends on a field, such as the impact of urban sustainability policies on climate resilience (Zhang et al., 2019). These components make trend analysis a comprehensive tool for understanding complex, qualitative data, facilitating a forward-looking perspective essential for adaptive and informed policy development.

For the purposes of our experiment, we derive the methodology for the first step of extracting "trend signals" and ascribe it the terminology of "topic extraction" which is often used in the topic modelling literature. We shall, use a modified BERT based methodology to extract semantic topics from the texts from ten projects under the SURE (Sustainable Urban Regions) funding. As we shall see, and is also advocated by (Braun and Clarke, 2006), themes do not just emerge from text but requires a certain view. Hence, we supplement the usual BERT based topic modelling architecture with a Theme mapping architecture as shown in Figure 2. Thereby, we propose a methodology for converting Semantic topics into Thematic topics.

3. Methodology

3.1. Project Data

The methodology outlines the approach used to review and analyse documents of ten research projects conducted under the Sustainable Urban Regions (SURE) initiative. The documents collected cover various phases of the projects, including the initial project proposals, the research and development phases, and the final reports. Each of these sources provides a comprehensive view of the projects, offering detailed insights into the participating institutions and collaborators, project funding structures, research objectives, identified problems, and the policy impacts. They also elaborate on the work packages, including the internal structure of the project and the distribution of tasks among partners. In addition, they contain project timelines, planning documents, academic profiles of team members, relevant references, and the academic literature that supports the research design and implementation.

Exclusion Criteria:

In order to ensure data confidentiality and maintain a focus on information directly relevant to the analysis, several categories of data were deliberately excluded from the review. First, the academic and professional curricula vitae of project partners were omitted, as they primarily contain personal backgrounds and qualifications that are not central to the scope of this analysis. Second, any personal contact information—such as email addresses, telephone numbers, or mailing addresses—was excluded in order to protect the privacy of the participants. Third, all financial documentation was removed from the analysis. This includes detailed funding information, cost breakdowns, and budgets.

Included Information:

The analysis focuses specifically on information that contributes directly to understanding the structure, aims, and scholarly contributions of each project. This includes general overviews and summaries that contextualize the research within broader goals, as well as detailed descriptions of the objectives, methodological frameworks, and policy relevance. In addition, the documents' descriptions of how work is divided through structured work packages, the scheduling and sequencing of tasks across timelines, and the incorporation of key academic references were all central to the analytical process. By narrowing the scope of the review to these components, this methodological approach ensures that sensitive personal and financial data are excluded, while allowing a comprehensive understanding of how the SURE projects contribute to sustainable urban regions´ development.

3.2. Focus Topics / Research Themes

In the initial stages of the project SURE, the team of qualitative researchers from the Synthesis team, had conceptualised 6 areas of Focus called Focus Topics. These are the research themes that each of the projects would fall under depending on their project proposal and expertise of research. The application of this methodology at the early stages of the project mandated that a limited number of documents were used for the manual thematic analysis. The documents used in this instance were the ones that we also use for the current experiment i.e, the project proposals and the shorter project profiles.

A systematic review was conducted to extract and identify the central themes from these documents. This process was complemented by open coding, a qualitative data analysis technique, to ensure an exhaustive identification of key topics. Following the coding process, thematic analysis was employed to group related topics, enabling the identification of overarching themes and priority areas. Table 1 shows the results of the initial experiment to derive the Research themes / Focus topics and their lower-order representations driven by domain expertise of the researchers involved.

Table 1 Overarching Themes or Focus Topics devised by domain expertise

Themes / Focus Topics	Lower-order Abstraction	Extracted from Documents
Integrated planning and development	Land use and urban development; Infrastructure and services	Strategic planning and governance tools; Build back better strategies; Integrated urban development; Integrated urban neighbourhood development; Co-development and decision support planning tools; Participatory urban development; Efficient municipal services of general interest; Reliable drinking water and sanitation supply; Integrated urban development
Urban Rural Nexus	Integrated territorial development	Urban-regional cooperation; Sustainable urban-rural development
Resource Efficiency and Mitigation	Strengthening circular development; Sustainable supply and waste infrastructure; Increase of indoor comfort; Energy and resource-efficient building	Urban -rural material flows
Risk management and Risk reduction	Concepts of multiple risk prevention; Strategic risk reduction; Risk transfer; Securing of critical infrastructure	Culturally adapted risk prevention; Informal settlements; Flood risk reduction; Improving (reduction) socio-economic vulnerability
Sustainable Behaviour and Practices	Migration; Environmentally friendly behaviour and sustainable living; Social and environmentally sustainable technologies; Sustainable utilization concepts for historic wooden houses; Environmentally friendly behaviour and sustainable lifestyles;	
Ecosystem Services and nature-based solutions	Blue and green infrastructures; Adaptation to flood risk	Sustainable water management; Strategic planning of green infrastructure; Modelling of nature-based solutions and blue-green infrastructure; Heat adaptation through green infrastructure and air quality improvement; Promotion of Urban Greening and Urban Climate; Improve of Air Quality

3.3. Proposed Architecture

Our proposed architecture for the experiment involves leveraging the original BERT workflow and integrating an additional flow that allows us to map the Semantic clusters on to a Thematic Analysis. Figure 2 shows the structure of the Integrated workflow, where the top part synthesizes all the Project documents like proposals and profiles, while the lower part synthesizes the Focus topic or Theme texts employing the same algorithms. We discuss each of these sections and the steps in the next section.

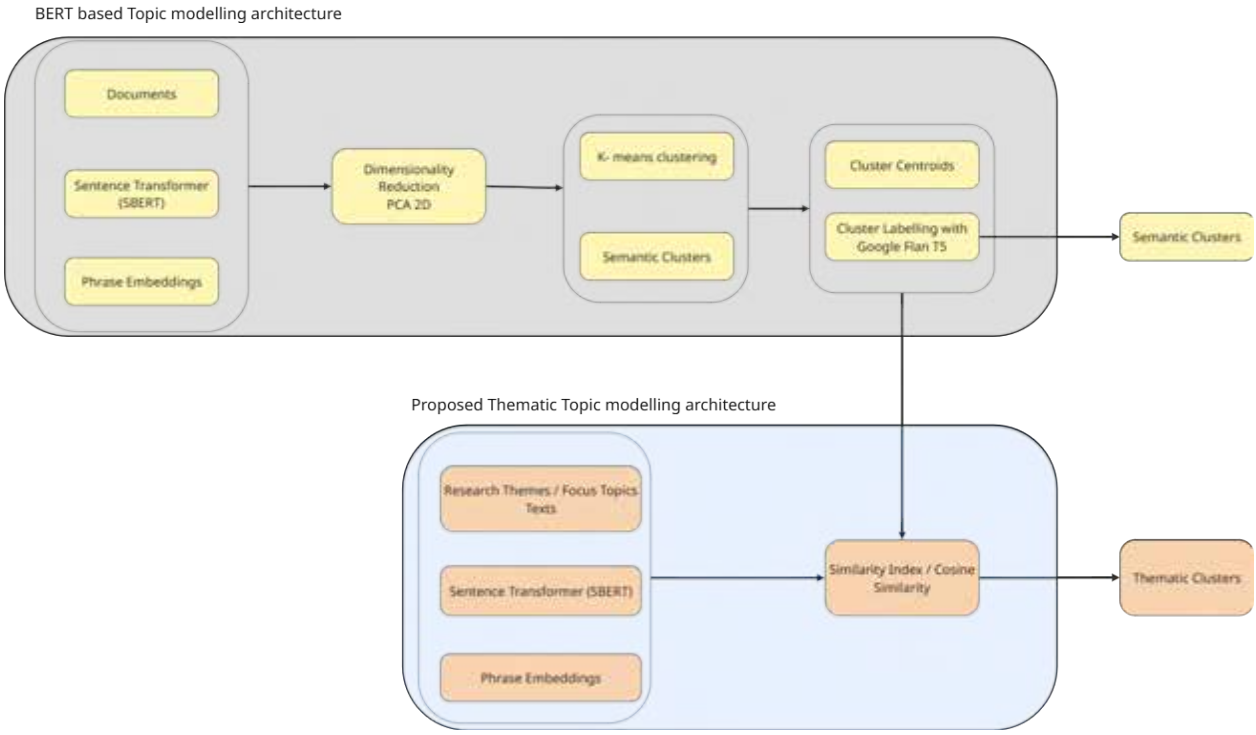


Figure 2 Proposed Architecture for Thematic clustering

4. Analysis

4.1. Phrase extraction

SBERT or Sentence BERT is a modification of the original BERT architecture, specifically designed to produce semantically meaningful phrase embeddings in a computationally efficient way. SBERT employs a Siamese Network structure map, which we shall not go into the details of in this paper, to map sentences into a vector space where semantically similar phrases lie close together and dissimilar ones far apart. As is evident form the study of (Cheng et al., 2023), SBERT's embeddings have been shown to align much better with human judgements of similarity and work effectively in downstream unsupervised tasks like clustering and topic modelling (Cheng et al., 2023).

We use the all-MiniLM-L6-v2 which is part of the SBERT family and used for its efficiency and small relative size. A total of 96,232 tokens were processed from the Project documents and these embeddings were saved with their metadata onto a ChromaDB vector database.

4.2. Dimensionality Reduction and Clustering

We use the PCA (Principal Component analysis) for transforming higher dimensional vector embeddings to 2D vectors to capture the highest variance. Here, as shown in Figure 3, the X-axis represents the highest variance and is the most important feature direction. The Y-axis represents the second highest variance in the data and together they aim to represent the highest possible variance in the data. This allows us to cluster the data in a more efficient way by using the standard k-means approach.

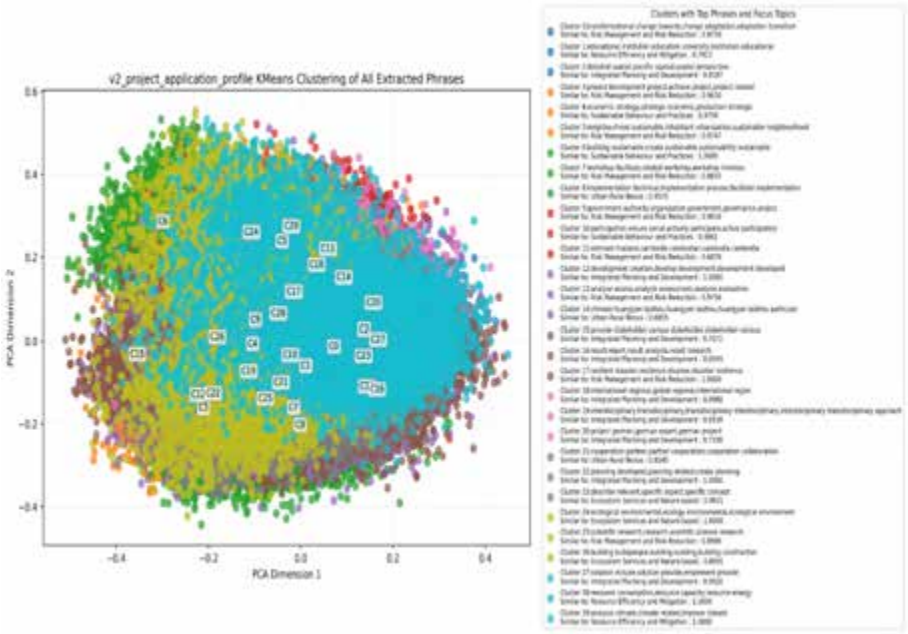


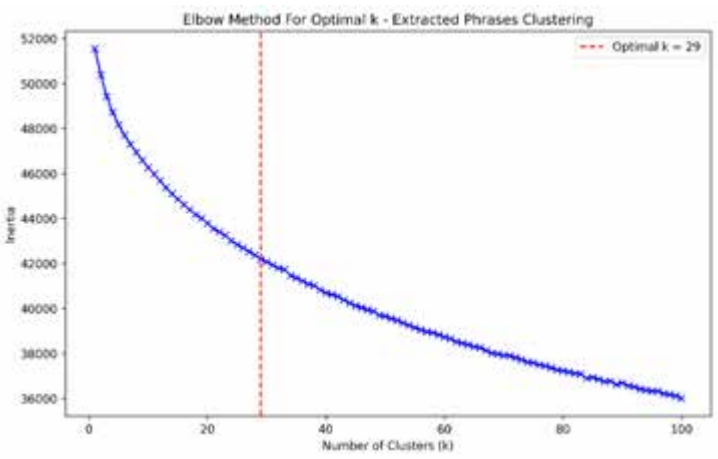
Figure 3 K-means Clustering of embeddings after PCA

Using K-means for clustering, our goal is simply to partition n data points into k clusters so that points within each cluster are as semantically similar as possible, and the clusters themselves are as distinct as possible. The challenge, however, is finding the optimal k for such large number of embeddings. We employ a heuristic approach of, what is commonly referred to as the 'Elbow Curve', where we plot for a loss function against every value (integer) of k iteratively until the change in the loss function is

minimal or stagnant. We use the function Inertia which is the total Within-Cluster Sum of Squared distances between each data point and iterate it over k.

$$Inertia = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

To understand intuitively, a lower Inertia represents tighter clusters which means that the data points are closer to the cluster centroids. The eventual aim of the curve is to find the highest k for which the Inertia is the smallest and further reduction is not significant. Figure 5 shows the curve as applied in our experiment resulting in an optimal value of k = 30.



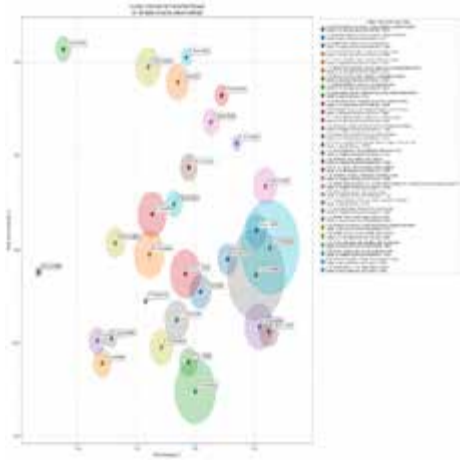


Figure 4 Deriving optimal K clusters
Figure 5 Clusters represented with their Centroids

A simpler visualisation of the resulting clusters is offered in Figure 5 which highlights the clusters, their sizes by the amount of data points they contain and their centroids. Table 2 shows some of the clusters and the data points or phrases extracted from the text.

In order to make these clusters human readable, we apply an open source LLM Google flan-t5-large, on the top 50 phrases in every clusters. This results in a stricter semantic topic creation, especially for larger clusters. Table 2 shows some sample clusters, the embedded phrases it contains and the LLM labelling of the semantic topic.

Table 2 Sample Cluster data points and their LLM labels

Cluster	Top 10 Phrases	Google flan-t5-large labelling
Cluster 0	transformational change towards; change adaptation; adaptation transition; transformational change; transformative change; transformation process towards; transformation process toward; change transformation process; adaptation transformation; transformative change towards	Transformational change and towards adaptation
Cluster 2	detailed spatial; specific spatial; spatial perspective; spatial structure; spatial mapping; spatial typology; spatial aspect; spatial perception; realistic spatial; investigates spatial	Spatial representations of a realism
Cluster 10	participation ensure social; actively participate; active participatory; analysis participatory; establishes participatory; realize participatory; involves participatory; effective participatory; community participation; approach participatory	Participatory analysis and the role of participation in social change

4.3. Converting Semantic Clusters into Thematic clusters

As is clearly seen from Table 2, the clusters have a significant semantic similarity whereby phrases with similar meaning or structure have been clustered efficiently. Yet, this erodes the thematic understanding of a 'topic' from the text while focusing only on parts of language as topics. To circumvent this issue, we propose the addition of the second part of our proposed architecture as shown in Figure 2 where we map our BERT based seman-

tic clusters onto the Thematic clusters derived by domain experts.

We follow the same methodology as used for the project documents, now with the Focus Topics/Themes that were devised by domain expertise as shown in Table 1. Each of the Focus topic has a short text that conveys the theoretical background and the current research directions in that theme. These texts are processed with all-MiniLM-L6-v2 and the embeddings for each theme are stored separately in a ChromaDB vector database. The number of key-phrases identified under each focus topic is shown in Table 3.

Table 3 number of Phrase embeddings in Focus Topics

Focus Topic / Theme	Number of key-phrases extracted
Integrated planning and development	146
Urban Rural Nexus	98
Resource Efficiency and Mitigation	167
Risk management and Risk reduction	147
Sustainable Behaviour and Practices	149
Ecosystem Services and nature-based solutions	168

We then undertook a cosine similarity to check the similarity of all the clusters obtained from the project documents with each of the six themes separately. Thus, we map each cluster from Figure 5 onto the theme embeddings and assign a cosine similarity score between 0 and 1, where 0 is no similarity to 1 is complete similarity with embeddings within the themes.

We noticed, that many a cluster would be less relevant to a certain theme and hence needed human intervention to design a threshold of similarity for inclusion. This was done manually based on the domain expertise and fixed at 0.75 which means that the clusters under the 0.75 score threshold were not considered relevant for the theme. Figures 6 – 11 show the different clusters belonging to each theme or Focus topic.

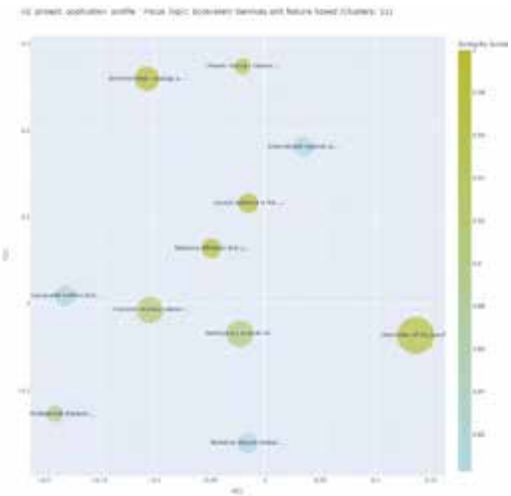


Figure 6 Ecosystem Services

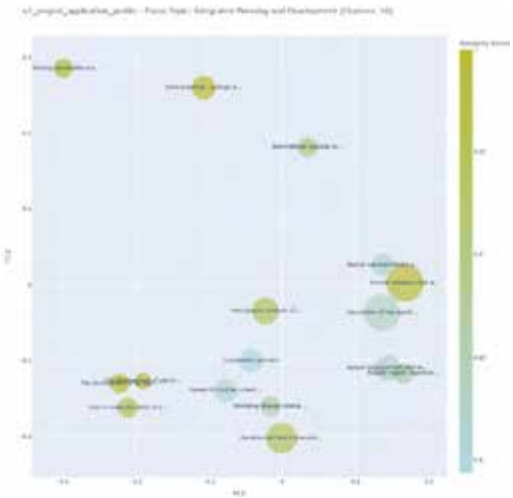


Figure 7 Integrated Planning and Development

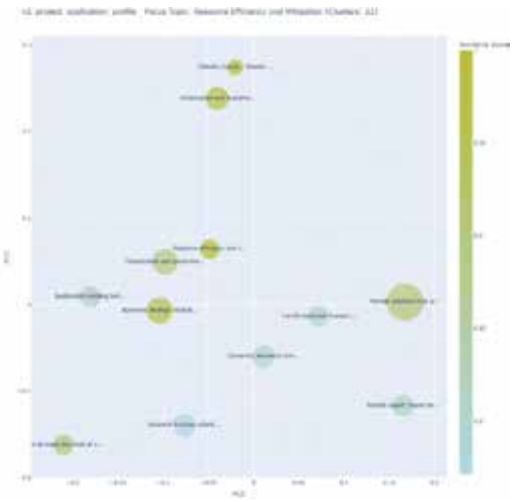


Figure 8 Resource Efficiency and Mitigation

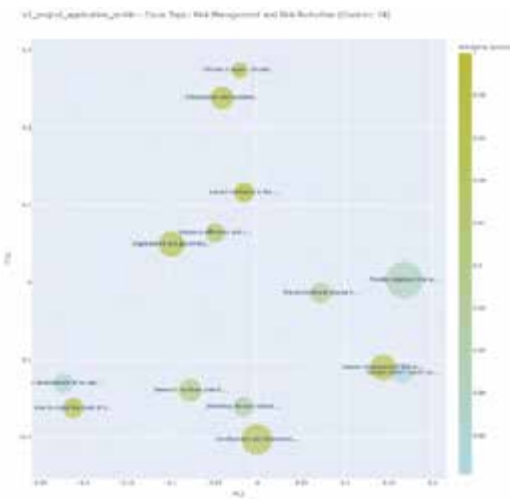


Figure 9 Risk management and Risk reduction

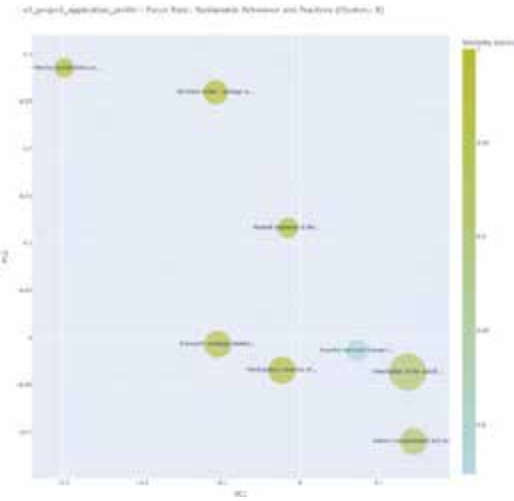


Figure 10 Sustainable Behaviour and Practices

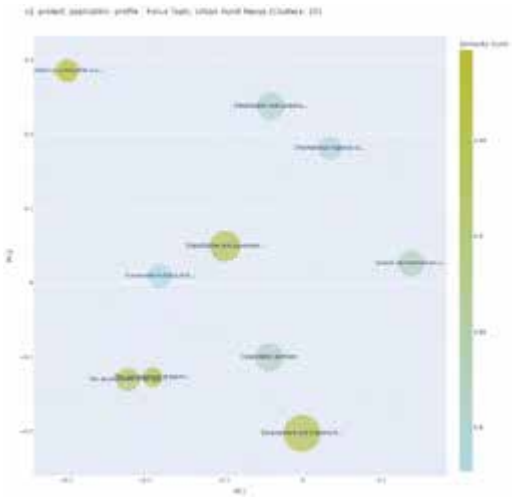


Figure 11 Urban Rural Nexus

5. Conclusion

5.1. Results

Our experiment focussed on generating thematic clusters from semantic clusters with the help of transformer-based models. The figures 6 – 11 show the details of thematic belonging of semantic clusters of our text data. To simplify our understanding, Figure 12 shows a Sankey chart referring to the relationship between semantic topics and their thematic focus.

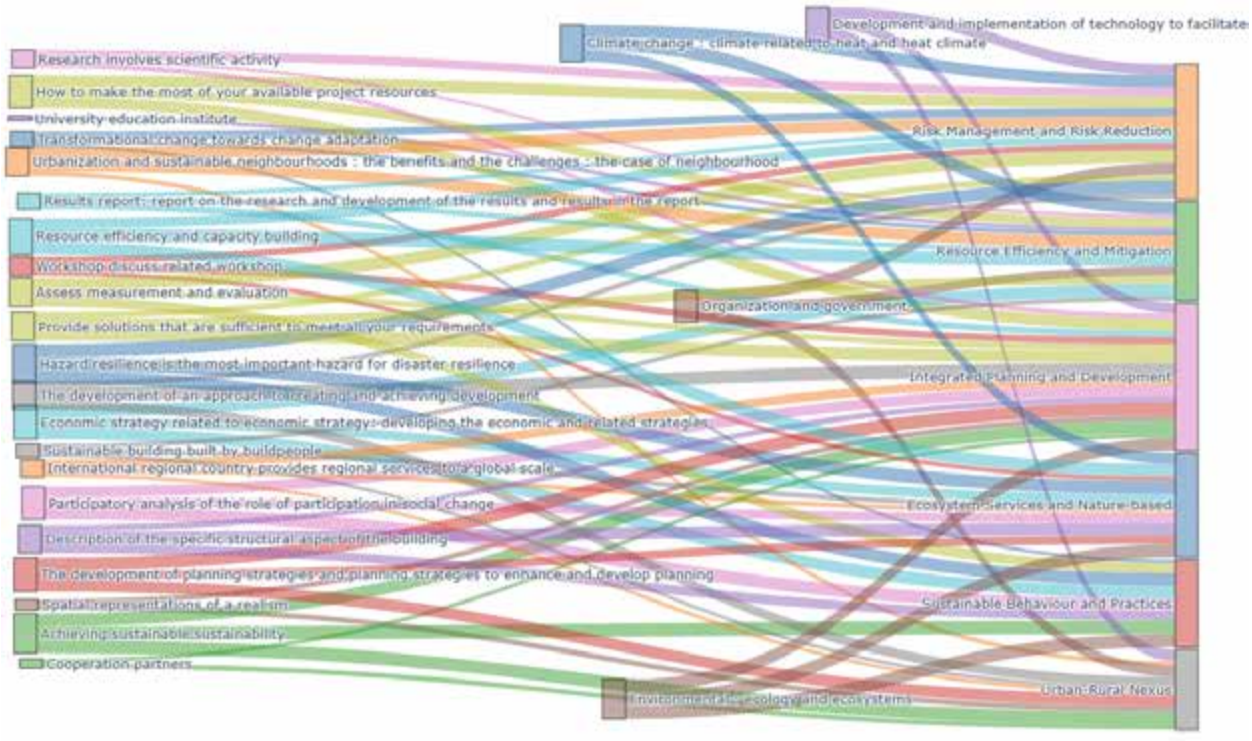


Figure 12 Sankey chart for understanding the links between Thematic and Semantic clusters

Some of the topics generated from the clusters, for example, 'Climate change related to heat stress mitigation' belongs to three different thematic groups viz. Risk Management and Risk Reduction, Resource Efficiency and Mitigation and Ecosystem Services and Nature-based solutions in varying proportions indicating its importance under these themes. In a similar vein, the topic 'Environmental: ecology and ecosystems' belongs in almost equal proportions to Integrated Planning and Development, Ecosystem Services and Nature based solutions and Sustainable behaviour and Practices. Yet, the accuracy overall can be argued to be a bit nuanced as it may not be entirely clear why a certain cluster should belong to a certain theme without the domain expertise present at our disposal.

5.2. Future pathways

This experiment was conducted as part of the Synthesis research project which aims to establish a sustainable contribution to the emerging field of knowledge management in the field of Sustainable Urban Development. It deals with a part of the larger question of monitoring trends in Urban Development research. As is our position, a topic is a subset of a Theme which mapped over time becomes a Trend, we aim to further this experiment with a larger corpus of structured Thematic texts onto which to map more documents and research and to monitor the consistent themes as trends.

5.3. Limitations

The experiment here is limited by the nature and sensitivity of the documents involved in within the larger SURE project. Thus far we have not used publicly available datasets like research papers, for our analysis and hence the architecture remains a proof of concept. Additionally, the text for the Focus topics or Themes devised by domain expertise is nimble and should be expanded to include more depth and breadth around a certain theme for better analysis.

5.4. Acknowledgements
This research is funded by the German Ministry of Education and Research (BMBF) and is part of its funding priority Research for Sustainability (FONA).

References

Boyatzis, R.E. (1998) Transforming qualitative information: Thematic analysis and code development. (Transforming qualitative information: Thematic analysis and code development). Thousand Oaks, CA, US: Sage Publications, Inc.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', Qualitative Research in Psychology, 3(2), pp. 77–101. doi: 10.1191/1478088706qp063oa

Braun, V. and Clarke, V. (2021) 'Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern based qualitative analytic approaches', Counselling and Psychotherapy Research, 21(1), pp. 37–47. doi: 10.1002/capr.12360

Cheng, H. et al. (2023) 'A Neural Topic Modeling Study Integrating SBERT and Data Augmentation', Applied Sciences, 13(7), p. 4595. doi: 10.3390/app13074595

Churchill, R. and Singh, L. (2022) 'The Evolution of Topic Modeling', ACM Computing Surveys, 54(10s), pp. 1–35. doi: 10.1145/3507900

Clarke, V. and Braun, V. (2017) 'Thematic analysis', The Journal of Positive Psychology, 12(3), pp. 297–298. doi: 10.1080/17439760.2016.1262613

Egger, R. and Yu, J. (2022) 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts', Frontiers in Sociology, 7, p. 886498. doi: 10.3389/fsoc.2022.886498

El-Gayar, O. et al. (2024) A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps. Americas Conference on Information Systems (AMICS).

IPCC (2021) Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of teh Intergovernmental Panel on Climate Change.

Kapoor, S. et al. (2024) Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling. Available at: http://arxiv.org/pdf/2409.15626v1.

Kashi, A. and Shah, M.E. (2023) 'Bibliometric Review on Sustainable Finance', Sustainability, 15(9), p. 7119. doi: 10.3390/su15097119

Levitt, H.M. (2018) 'How to conduct a qualitative meta-analysis: Tailoring methods to enhance methodological integrity', Psychotherapy Research : Journal of the Society for Psychotherapy Research, 28(3), pp. 367–378. doi: 10.1080/10503307.2018.1447708

Lore, M., Harten, J.G. and Boeing, G. (2024) 'A hybrid deep learning method for identifying topics in large-scale urban text data: Benefits and trade-offs', Computers, Environment and Urban Systems, 111, pp. 102–131. doi: 10.1016/j.compenvurbsys.2024.102131

Mu, Y. et al. (2024) Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. LREC-COLING. Torino. Available at: http://arxiv.org/pdf/2403.16248v2.

OECD (2020) Cities in the World: A New Perspective on Urbanization. Paris.

Olawumi, T.O. and Chan, D.W. (2018) 'A scientometric review of global research on sustainability and sustainable development', Journal of Cleaner Production, 183, pp. 231–250. doi: 10.1016/j.jclepro.2018.02.162

Sneegas, G. et al. (2021) 'Using Q-methodology in environmental sustainability research: A bibliometric analysis and systematic review', Ecological Economics, 180, p. 106864. doi: 10.1016/j.ecolecon.2020.106864

Sy, C.Y. et al. (2024) 'Unsupervised Machine Learning Approaches in NLP: A Comparative Study of Topic Modeling with BERTopic and LDA', Intelligent Systems And Applications In Engineering, 12(21s), pp. 3276–3283.

Thompson, L. and Mimno, D. (2020) Topic Modeling with Contextualized Word Representation Clusters. Available at: http://arxiv.org/pdf/2010.12626v1.

UN-Habitat (2020) World CIties Report 2020: The value of Sustainbale Urbanization.

United nations (2019) World Urbanization Prospects: The 2018 Revision.

Zhang, C. et al. (2019) 'Bibliometric Analysis of Trends in Global Sustainable Livelihood Research', Sustainability,

11(4), p. 1150. doi: 10.3390/su11041150