Unlocking the potential of NLP in text data analysis for sustainable urban development

Conference Paper · June 2023

CITATIONS
0

READS
103

5 authors, including:

Chintan Patel
HafenCity University Hamburg
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Katharina M. Borgmann
HafenCity University Hamburg
11 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Barabas Agota
HafenCity University Hamburg
6 PUBLICATIONS **0** CITATIONS
SEE PROFILE

IFKAD 2023 - Managing Knowledge for Sustainability
Matera, Italy 7-9 June 2023
ISBN 978-88-96687-16-1 || ISSN 2280-787X

Unlocking the Potential of NLP in Text Data Analysis for Sustainable Urban Development

Chintan Patel*

Digital City Science, Hafencity Universität
Henning-Voscherau-Platz 1
20457 Hamburg, Germany

Mohammed Sohanur Rahman

DigitalCity Science, Hafencity Universität
Henning-Voscherau-Platz 1
20457 Hamburg, Germany

Katharina M. Borgmann

DigitalCity Science, Hafencity Universität
Henning-Voscherau-Platz 1
20457 Hamburg, Germany

Ágota Barabás

DigitalCity Science, Hafencity Universität
Henning-Voscherau-Platz 1
20457 Hamburg, Germany

Jörg R. Noennig

Digital City Science, Hafencity Universität
Henning-Voscherau-Platz 1
20457 Hamburg, Germany

** Corresponding author*

Abstract

This paper reports on results of the SURE facilitation and synthesis research (FSR) project for the funding priority SURE (Sustainable Development of Urban Regions) of the German Federal Ministry of Education and Research (BMBF). SURE engages ten collaborative projects which develop concepts and test locally implementable solutions and strategies for sustainable transformation of fast-growing urban regions in Southeast Asia and China. SURE aims to create conceptual, theoretical, methodological, and translational innovations that integrate and move beyond discipline-specific approaches to address issues of sustainable urban development. The paper discusses the application of Natural Language

ISBN 978-88-96687-16-1 || ISSN 2280-787X

Processing (NLP) as one form of Artificial Intelligence (AI) to support data and knowledge synthesis in sustainable urban development research. The abundant urban data and recent advancements in the field of AI have the potential to transform how urban researchers perceive and tackle sustainable development-related problems of cities. The research team employs various NLP algorithms to assess text data with the goal to analyse patterns in order to explore intra-project synergies and research intelligence on future trends. NLP has exhibited an ability to digest copious textual data and improve the usability of urban corpora, improving study scope and reducing resources required for research. However, the implementation of NLP to study issues related to sustainable urban development is a relatively novel. Predominantly used NLP modules are unable to identify contextual relations amongst multiple words which is essential in urban region study. To overcome this issue, algorithms employed were trained to identify various word classifications related to urban study discipline for precise output. We discuss the preliminary results of the ongoing exploration and show how it could contribute to an understanding of large text-based data sets in urban knowledge management. We examine the possibilities and limitations of such an approach and discuss the implications of AI as part of a multi-methodological approach to carry out a synthesis of sustainable urban development research efforts across an entire region covered under SURE framework. The paper also gives an outlook on utilising new AI based algorithms to generate text-based data analysis channel as well as indicate the limits, successes, challenges and constraints of such approaches.

Keywords – Sustainable Urban Development, Natural Language Processing, Artificial Intelligence, Knowledge Management

Paper type – Academic Research Paper

1 Background and Introduction

Rapidly growth of urban dwellings in cities located in South-East Asia and China has posed a challenge to scholars in dealing with complex issues threatening the societal livelihood of community and sustainable development. Researchers Kates et al. (2001), Binder et al. (2015), Zscheischler et al. (2014) discovered that dependency on inputs from single discipline is not enough to address intricate and complex problems related to sustainable development of urban and rural areas. Rather a transdisciplinary framework encompassing academics, non-academics and mutual learning amongst stakeholders is necessary to determine a practice-oriented solution that could be implemented locally. On such premise, the German Federal Ministry of Education and Research (BMBF) sponsors the funding priority SURE 'Sustainable Development of Urban Regions' that promotes transdisciplinary research towards SDG localization and an accompanying synthesis research with a primary focus to explore new

ISBN 978-88-96687-16-1 || ISSN 2280-787X

collaborative approaches that enable societal contribution, future topics and challenges in urban and rural development.

The SURE funding priority comprises ten collaborative projects oriented towards development of concepts and testing implementable solutions locally in fast-growing urban regions in Southeast Asia and China. These ten projects and funding priority are accompanied by the SURE Facilitation and Synthesis Research (SURE FSR) project that aims to create conceptual, theoretical, methodological, and translational innovations that integrate and move beyond discipline-specific approaches to address the issue of sustainable urban development. Such framework enables vast tangible and intangible knowledge generation, that is captured from project documents (e.g. proposals, reports, results) and through facilitation activities (workshops, peer-to-peer meetings, interviews) to identify transdisciplinary and sustainable urban development challenges. The SURE FSR aims for transdisciplinary knowledge synthesis in sustainable urban research, with one of its pillars being 'Research intelligence ', where the team explores the digital tools and data science approaches to establish synergies between project via content comparison, topic modelling and exploration of future trends and challenges. This paper is embedded in the larger research effort of the SURE FSR, where the knowledge is being synthesized with the employment of AI tools. The IFKAD'23 paper by Agota Barabas et al., titled *'sustainable knowledge synthesizer: a modular tool for urban research'*, from the HafenCity University Hamburg (ID 245) shows how databases, as well as data services for knowledge management and communication and collaboration, are provided, adapting solutions from business intelligence (e.g. project dashboard, monitors, cockpits), and then aggregated in the functional tool "Synthesizer". The paper discusses the concept of such a synthesizing system and its application in a meta-research environment of transdisciplinary sustainable urban development approaches, sheds light on the opportunities and challenges of the development of such a synthesizing tool, and draws a first picture of the complexity accompanying the development of a "synthesizer" as part of the SURE FSR.

Recent strides in the field of artificial intelligence (AI) and machine learning give an opportunity to analyse such tangible project outputs. AI applications has been prevalent in energy efficiency and mobility since early 2010s while recently developed AI tool's applicability in sustainability science has proven effective in various sectors including water management, sanitation, agriculture, pollution, urban development as well as tracking progress of sustainable development goals (SDGs) (Goralski & Tan, 2020; WU et al., 2022).

ISBN 978-88-96687-16-1 || ISSN 2280-787X

A sub-branch of Machine Learning, denoted as Natural Language Processing (NLP), has a capability of understanding human language structures and assist in analysing big text corpora. It is a form of computational algorithms which perform various tasks such as text analysis, text categorisation, sentiment analysis, topic modelling, network analysis, or text summarisation. In context of transdisciplinary research, NLP bears large potential to perform qualitative analyses on large volumes of data collected through interviews, focus groups, or survey responses as well as identify the main topics or themes within a large amount of text. Furthermore, NLP has wide applicability in field of sustainability research due to wide range of customisation and possibility of creating tailored neural networks in accordance to given research objectives. Such AI-assisted qualitative data analysis can help researchers identify patterns and trends that might otherwise go unnoticed, providing a more comprehensive understanding of the research topic. A bibliometric study conducted by Fosso Wamba et al., (2021) observed a significant increase in the utilisation of neural networks and other Machine Learning approaches for addressing ecosystem challenges, infrastructure management, and stakeholder management. Although trend of using AI-related tools to tackle challenges related to urban development is still in its infancy, researchers have applied NLP techniques in various topics related to urban governance and management, public health, land use and functional zones, mobility, and urban design (Cai, 2021). Li et al. (2023) demonstrated the capability of NLP using social media data to predict urban flood susceptibility in Chengdu, China; while Bardhan et al. (2019) developed a more sophisticated NLP algorithm capable of performing sentiment analysis as a mean to document various aspects in integrated urban planning of slum housing rehabilitation projects in Mumbai. Riding this wave of exploration of AI applicability relevance in area of sustainability and urban development, SURE FSR attempts to explore the sphere of content comparison and qualitative research in order to find cross-cutting topics and establish synergies between diverse topics. Exploring the applicability of AI in the area of sustainability and urban development, SURE FSR employs it for content comparison, to validate findings (e.g. focus topics), search for synergies between projects and knowledge synthesis. For this end, the SURE funding priority has been a rich source of tangible data to be analysed with NLP in order to create and test hypotheses.

ISBN 978-88-96687-16-1 || ISSN 2280-787X

2 Research Methodology

2.1 Research Objective

The SURE Facilitation and Synthesis research Project explores the challenges of synthesis research in cross-cultural settings and institutional set-ups, with a particular focus on sustainable development in the built environment in South-East Asia and China. The activity of this complex project is divided in two main "wings": facilitation and synthesis research. The approach used in the SURE Facilitation and Synthesis research project is transdisciplinary, problem-driven, and solution-oriented. The project employs a meta-study methodology that synthesizes and consolidates existing conceptual, methodological, and empirical knowledge from literature and the ten collaborative projects within the SURE funding priority. Overall, the approach used in the SURE Facilitation and Synthesis research project is designed to support individual research projects and to systematically leverage the cross-project synergy potentials at the funding priority level. The goal is to create conceptual, theoretical, methodological, and translational innovations that integrate and move beyond discipline-specific approaches to address the issue of sustainable urban development.

The project looks to synthesize and consolidate existing conceptual, methodological, and empirical knowledge from literature and the ten SURE collaborative projects. This qualitative meta-study aims to contribute to the transdisciplinary sustainable research discourse by providing a scientific contribution to the third epistemic way (Lang, Wiek et al. 2012). The research project is designed to support individual research projects on the one hand and to systematically leverage the cross-project synergy potentials at the funding priority level based on the collected and structured knowledge from the projects and beyond. Conducting meta-research across disciplines and across cultural borders requires a management of knowledge that is sensitive towards these challenges (Ioannidis, Fanelli et al. 2015). Within the SURE facilitation and synthesis research approach, we developed a research and knowledge architecture that allows for constant reflection to improve the applied concepts. The overall structure of the SURE FSR is described in more detail in the IFKAD'23 paper 'Meta-study: research architecture for sustainable knowledge synthesis' (ID 231) by Katharina M. Borgmann et al. from the HafenCity University Hamburg.

Although the funding priority targets inherently non-linear development dynamics that are difficult to capture by computation, trained algorithms still may provide a substantial support. For this reason, we put forward the concept of a

ISBN 978-88-96687-16-1 || ISSN 2280-787X

technology-aided knowledge synthesizer and outline its conceptual design and technical implementation in a parallel paper to this conference (Barabas et al., 2023). The synthesizer is conceived as a digital infrastructure for knowledge creation and decision support, based on thorough project analysis and content elaboration. Assisting with scientific classification, among others, it aims to indicate future research topics and challenges in sustainable urban development. The system's idea is rooted in knowledge theory and knowledge lifecycles models with a focus on modes of knowledge synthesis – the integration and refinement of existing knowledge into new insights. The conceptual design of the synthesizer is composed of modules for the access, structuring and integration of knowledge captured from the transdisciplinary research projects in SURE.

This paper presents an exploratory inquiry's initial findings with NLP that is part of synthesizer´s module development. This component is envisioned to conduct the cross-project content analysis and topic modelling to aid the synthesis research. Previously the SURE FSR team has synthesised and clustered without algorithmic support six focus topics based on projects and funding priority contents:

1. Urban-Rural Nexus
2. Resource Efficiency and Mitigation
3. Sustainable Behaviour and Practices
4. Ecosystem Services and Nature Based Solutions
5. Risk Management and Risk Reduction
6. Integrated Planning and Development.

Our initial exploration of NLP aims to validate these six focus topics generated in the researchers´ previous discourse, and to understand the opportunities and challenges of automated text analysis and topic modelling within the overall process of knowledge synthesis. For the validation, we included the original documents of the BMBF funding call and project proposals that were used for the initial topic synthesis. In addition, we extended the document list with relevant policy frameworks and strategy documents to generate still better results.

Albeit the investigation is still embryonic, it provides valuable insights in development of topic modelling and discovery of future trends and challenges in sustainable urban development.

ISBN 978-88-96687-16-1 || ISSN 2280-787X

Table 1: Documents included in Dataset

Name	Type	Publication Year
Project 1	Project Proposal	2017
Project 2	Project Proposal	2017
Project 3	Project Proposal	2017
Project 4	Project Proposal	2017
Project 5	Project Proposal	2017
Project 6	Project Proposal	2017
Project 7	Project Proposal	2017
Project 8	Project Proposal	2017
Project 9	Project Proposal	2017
Project 10	Project Proposal	2017
IPCC 5th Edition	International Policy Framework	2014
Paris Agreement	International Policy Framework	2015
New Urban Agenda	International Policy Framework	2016
WBGU	German Policy Framework	2016
FONA3	German Policy Framework	2016
BMBF Call	Call for Proposal	2016
FONA Strategies	German Policy Framework	2021

*Abbreviations
IPCC - *Intergovernmental Panel on Climate Change*
WBGU - *Wissenschaftlicher Beirat der Bundesregierung Globale Umweltveränderungen (German Advisory Council on Global Change)*
FONA - Forschung für Nachhaltigkeit (Sustainability Research)
BMBF - Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)

2.1 Data – Proposals and Policies

The database of this study consists of proposals prepared by all ten projects and policies that has served as foundation for the SURE funding priority. Additionally, text extracted from webpage dedicated to all for proposals is also included in study. *Table 1* enlists documents included in analysis along with the publication years and other attributions. Due to privacy measures project names are not disclosed. The policy framework documents and SURE call for proposals are open-source and available online.

ISBN 978-88-96687-16-1 || ISSN 2280-787X

2.3 Data Preparation

Eliciting textual data and pre-processing of documents for NLP analysis is necessary to remove the irrelevant information such as headers, footers, and references along with images. While manually pre-analysing the project proposal documents, irrelevant data without objective value were omitted. They include sensitive information such as partner information, personal details along with other numerical and graphical data.

3 Data Analysis and Results

3.1 NLP Analysis

3.1.1 Document Cleaning

Fig 1 illustrates the workflow of the methodology encompassing the step-by-step process for NLP analysis. The data pre-processing and removal of sensitive data from the documents has been conducted manually without any computation, resulting in converted pdf files to ensure better accessibility across various platforms. Python is chosen as a programming language to perform the research analysis, as a significant number of open-source algorithm libraries are offered in that language that are required at various stages of research. An iterative process is required to validate and check the compatibility at each stage of such analysis to ensure the accurate extraction of the desired output. Thus, in every stage of the process represented in decision tree (*fig 1*), a compatibility check is performed to maintain the homogeneity of the data.

ISBN 978-88-96687-16-1 || ISSN 2280-787X

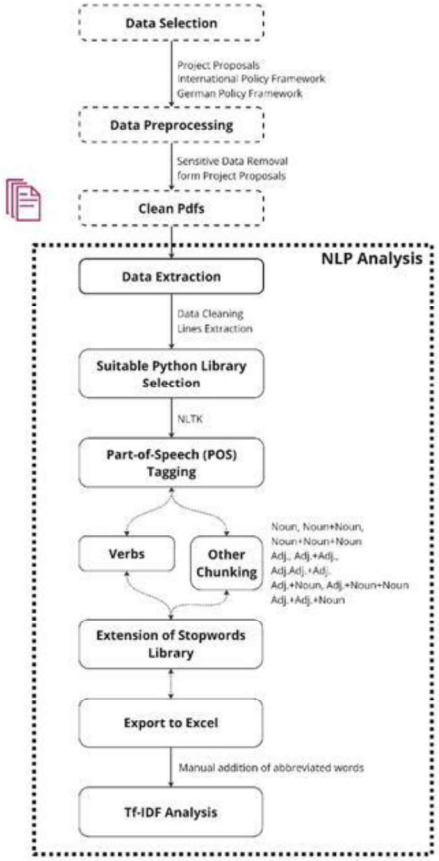

Fig 1: NLP Analysis Decision Tree

3.1.2 Approach and Customisation of the Natural Language Processing

Numerous open-source NLP libraries are available that are capable of performing basic tasks such as tokenisation, word tagging to advanced Machine Learning frameworks for natural language understanding and generation. Choosing a suitable library for required analysis could be a daunting task as these libraries offer specialised computation to perform specific tasks. Although the flexibility of Python scripts allows for the usage of multiple libraries in one algorithm, doing so also increases risk of errors and incompetence in analysis. Part-Of-Speech (POS) tagging emerged as essential process of computation from NLP analysis scope and output requirements study. One of the prominent complexities observed during an initial check is that word roots and its affixes usually hold contextually different meaning in the sentence formation of texts related to urban development and sustainability science (Jha et al., 2021). Thus, the standard linguistics processes of text lemmatisation and text stemming are

ISBN 978-88-96687-16-1 || ISSN 2280-787X

avoided in our algorithm to preserve the textual context as much as possible in NLP results. Versality of the library also emerges as a deciding factor in the selection. As this study is to serve as basis for topic modelling and future challenges predication, the chosen library should offer tools and functionality to perform tasks such as named entity reorganisation (NER), machine translation, text classification and sentiment analysis. Primary requirement for this study is to perform an extensive POS tagging and identification of n-grams variations combining noun, adjective, adverb as well as long expressions. The Natural Language Toolkit (NLTK) was chosen as primary Python library to perform task related to NLP, as it offers tools to perform the afore-mentioned requirements and ease of customisation (Bird et al., 2009). In order to extract lines and sentences from the cleaned documents, removal of punctuations and word wrapping is necessary to achieve accurate results (Goralski & Tan, 2020). Python offers various pre-programmed libraries to perform line and sentence extractions from pdf documents. After conducting an iterative process conducted on assorted libraries and respective outputs, PdfPlumber library was chosen in the further computation as it offers better functionality in processing texts from diverse sources (GitHub, 2023). Due to diversity of data, it is observed that predefined stop words and punctuations used by the library are not adequate to provide satisfying results. Therefore, additional custom punctuations (see Annex 1) are amalgamated with existing stopword directory of NLTK after iteration.

To achieve the indicated research objective and uncover potential correlation between policy frameworks and the established project purpose, the NLP algorithm was trained to identify and extract keywords from each document. A bilateral approach was integrated in the Python script to differentiate POS tagging of verbs and various combinations of nouns and adjectives. While existing studies claim to have completely replaced traditional analysis methods in few fields of urban planning, existing models available for NLP analysis of text related to sustainability and related urban development lack cohesion of results (Cai, 2021; Kölbel et al., 2022). To avoid this, a rather novel approach is necessary that analyses the text holistically using an assorted combination of POS chunks to decrease the probability of overlooking phrases that determine the context. A NLP technique denoted as chunking is customised and assimilated in the algorithm. Chunking essentially identifies sequences of words that belong together in a sentence based on their grammatical role and extracts the contextually important sequences of words as units of meaning. Rather than utilising predefined rules and patterns, the algorithm is adapted to identify

ISBN 978-88-96687-16-1 || ISSN 2280-787X

grammatical combinations up to three consecutive words including noun, noun+noun, noun+noun+noun, adjective (adj.), adj.+adj., adj.+adj.+adj., adj.+noun, adj.+noun+noun adj.+adj.+noun. Initial iteration stages of NLP computation showed inadequacy in identifying different cases of alphabets and punctuations (see Annex 1). A computation loop was added to counter and identify same words in different case-sensitivity and avoid duplication of same keyword in output. A more coherent output is obtained in further iterations of the NLP algorithm runtimes after adaptive measures were added. In the final stages of the NLP process, a term-frequency (TF) analysis is carried out on identified keywords. In essence, TF measures the frequency of occurrence of each term in a document or corpus representing the importance of the term in the analysed document. It provides intrinsic value to each keyword that are necessary in an effort to establish correlation or to compare diverse set of documents with varying length and text structure. A well-structured output in spreadsheet format is obtained by using ExcelWriter library of Python for the further manual analysis and validation.

3.1.3 Manual Iteration of NLP output

A manual interpretation of results obtained from NLP analysis is vital for a tenacious approach towards the establishment of topical synergies and correlations. A prominent discrepancy in frequent term analysis was discovered as the NLP engine lacked capacity to incorporate abbreviated words in the account. Notably the NLTK library requires extensive modifications before being able to consider and account for abbreviated wordings in documents. Thus, a rather simplified manual approach was chosen to avoid discrepancies in existing output from NLP analysis. Furthermore, an additional manual iteration was undertaken to indicate the relevance of each keyword within respective text of document. To define the relevance and assign a comparable scale with respect to other data extracted from document, a rather rudimentary version of TF-IDF (term frequency-inverse document frequency) approach was taken which involves proportionating the frequency of term in reference to total word count of corresponding document. The remaining manual iterations involved the sorting of data for the purpose of visualisation and convenience of usage.

3.2 Focus Topics

The clarification of focus topics is an essential part of the SURE funding priority, as they serve as a foundation for subsequent knowledge exchange and

ISBN 978-88-96687-16-1 || ISSN 2280-787X

dissemination, as well as for the exchange of case studies and approaches across projects. By way of non-algorithmic analysis, the SURE FSR team had established six such focus topics, as being woven into each project´s efforts of developing locally adaptable solutions and strategies (see section 2.1). In that 'non-technical ' attempt, a rather tenacious Grounded Theory-approach had been taken to identify and classify the most frequent keywords in each project proposal, and to derive an initial project ontology and list of focus topics (BMBF, 2023; FONA, 2023; Slawski et al., 2022; SURE, 2021).

3.3 Results of NLP Analysis

In the process of this NLP analysis, we used all ten project proposals and policies (see Table 1) to identify the most frequent keywords or topic clusters to determine the primary trends and foci of the respective documentation. Additionally, to explore potential links between project proposals and policy framework, a combined text of all proposals was created and analysed with the NLP algorithm to find most frequent terms and notions. Considering the vast number of keywords and their corresponding occurrence in documents, only the top 15 most frequently occurring keywords were retained from the project proposals for scrutiny, to establish consistency and reconcilability for critical analyse of findings. For the policy framework and analyses of all proposal texts, the top ten most frequently occurring keywords were retained.

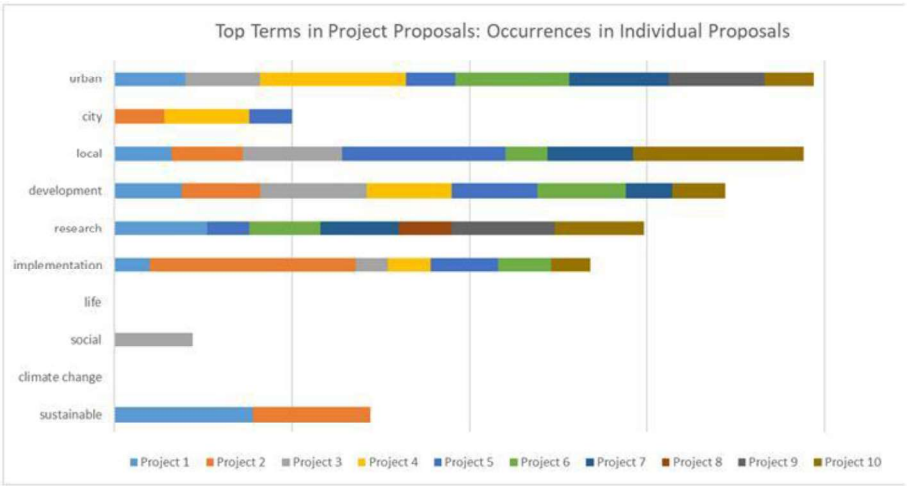

Fig 2: Top terms in Project Proposals: Occurrences in individual proposals

ISBN 978-88-96687-16-1 || ISSN 2280-787X

Fig. 2 represents a cross correlation between ten most frequent terms found in the combined text of the ten project proposals and their occurrences in the individual project proposal's 15 most frequent terms. The chart reveals several interesting traits and patterns. Most terms appear frequently in majority of project proposals, while terms 'urban' and 'local' are more prominent than others. Terms 'life' and 'climate change' are surprisingly missing from all ten project proposal's 15 most frequent terms. This could suggest these terms are mentioned as overarching topics and projects tackle them indirectly through their focus topic. Notably there are two outlier terms 'social' and 'sustainable' that appear in only one or two project proposals. Although since text lemmatisation and stemming is intentionally avoided in NLP analysis, other forms of root word sustainable such as sustainability, sustainable, sustainable development are interpreted differently. Thus, it could be that single word occurrence of these outlier terms is accounted differently in other project proposals. This comparison provides valuable insights into key themes and challenges that are undertaken by projects. For example, *project 2* emphasises 'sustainable' and 'implementation' while *project 1* puts more emphasis on 'sustainable', 'research' and 'development'.

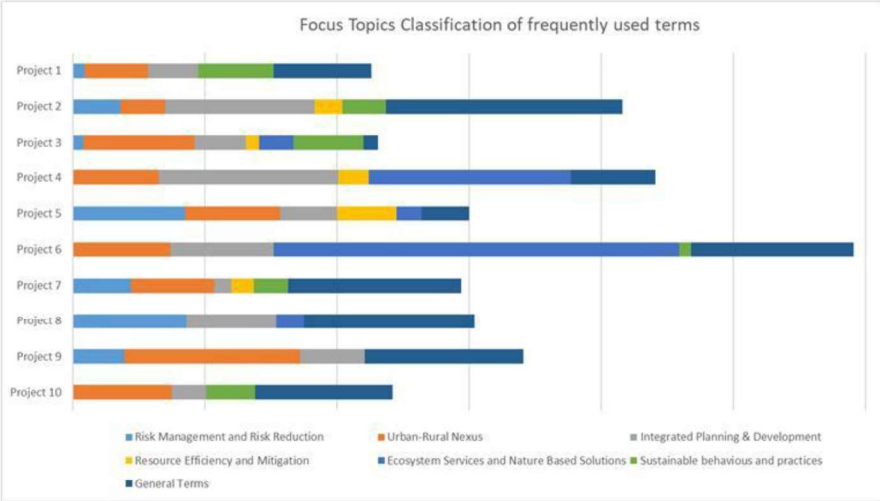

Fig 3: Focus Topics Classification of frequently used terms in project proposals

Fig. 3 illustrates the distribution of focus topic related terms in individual projects keywords. General terms were used to categorise keywords that are difficult to classify within the ontological boundaries of one single focus topic. Findings from Slawski et al., (2022) suggest that focus topics is reflection of BMBF's call for proposal and projects´ thematic focus and skills. NLP analysis

ISBN 978-88-96687-16-1 || ISSN 2280-787X

provides a more detailed insight into each projects´ focus. *Project 4* and *project 6*, for example, use 'Ecosystem Services and Nature Based Solutions' related strategies significantly higher than other projects, while all projects utilise 'Urban-Rural Nexus' related strategies in their quest for sustainable urban development solutions. As expected, targeted focus of each project varies greatly thanks to heterogeneous challenges undertaken by each project in combination with diversity in local and cultural environment.

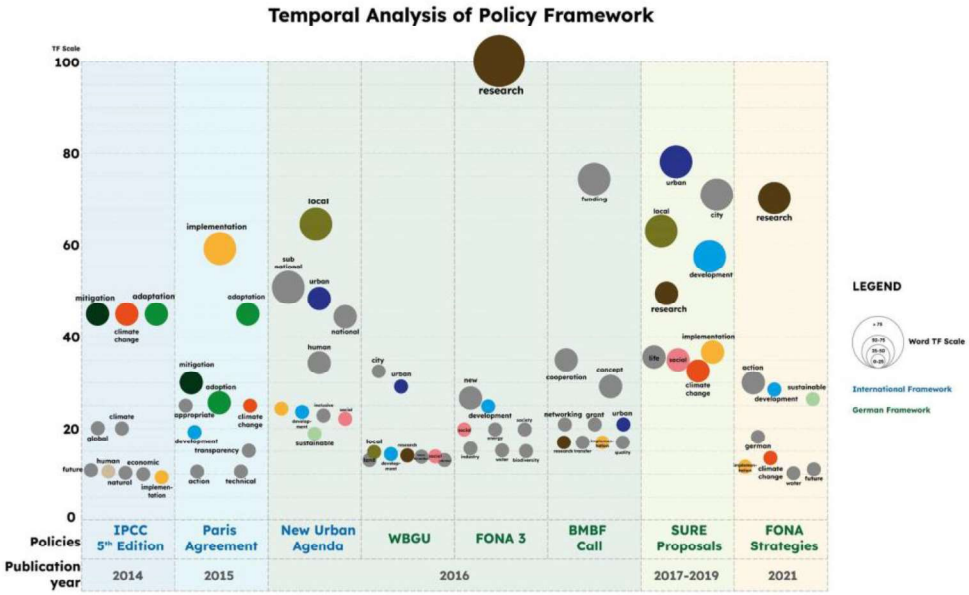

Fig 4: word occurrence over the years (Temporal Analysis)

Fig 4 shows the in-depth interpretation of output from all documents analysed by NLP. Ten Most prominent keywords from each document or webpage are represented according to their relative importance (TF-IDF). This visualization narrates the evolvement of foci of national and international policy framework and in what capacity the existing policy framework has influenced the challenges being tackled within SURE framework. The existing international policy framework has given priority to highlighting and narrating environmental challenges while German policy framework has shown more emphasis on research and development of solutions. The prominent words present in multiple policy documents are highlighted in graph by using bright colour theme while the other words are in grey.

ISBN 978-88-96687-16-1 || ISSN 2280-787X

4 Conclusions and Further Discussion

Transdisciplinary research is intrinsically complex, targeting multiple objectives simultaneously. The incorporation of innovative research directives along with traditional research methodology thus benefits the objective of achieving knowledge creation and dissemination across projects in the SURE funding priority. Although sustainable urban development related research puts emphasis on the local environment and context, an overarching and generic framework opens up valuable opportunities for the knowledge synthesis across the individual projects and the funding priority – for the currently running activities as well as for prospective future research. The explorative content comparison conducted via NLP techniques presented in this paper contributes to such objective. The linguistic analysis of project proposals highlights the priorities that each project gives to certain focus topics and challenges. An analogical NLP overview of the national and international policies that have led to SURE funding priority in the past, may line out the most prominent challenges overtime – as well as upcoming ones in the future.

The results show the application potential of NLP analytics in research related to urban development and sustainability science. The contextual preservation of identified keywords and chunks of words has been key for carrying out an effective and meaningful analysis of the given text corpora. The NLTK library has exhibited high confidence while performing POS tagging, chunking and context preservation when performing the preliminary text analysis with the relatively small set of given data from the SURE projects. Although the NLP analysis conducted in this paper is rudimentary, it has exhibited promising insights in projects´ primary focus and challenges. The implications of AI in sustainable urban development research are significant, and our study underscores the need to continue exploring and evaluating the possibilities and limitations of such an approach. An NLP analysis of rich database containing more publications, research articles and other tangible outputs from projects could provide valuable insights and serve as testing platform for various hypotheses. SURE FSR aims to exploit the opportunity to explore various NLP applications including topic modelling and entity recognition. It would support the overall meta-study to consolidate existing knowledge from literature and collaborative projects. Furthermore, NLP model could be trained with enough manual data-tagging to

ISBN 978-88-96687-16-1 || ISSN 2280-787X

identify specific challenges in field of sustainable urban development potentially providing insights into most recent challenges and prediction of future trends.

References

Bardhan, R., Sunikka-Blank, M., & Haque, A. N. (2019). Sentiment analysis as tool for gender mainstreaming in slum rehabilitation housing management in Mumbai, India. Habitat International, 92, 102040. https://doi.org/10.1016/j.habitatint.2019.102040

Binder, C. R., Absenger-Helmli, I., & Schilling, T. (2015). The reality of transdisciplinarity: A framework-based self-reflection from science and practice leaders. Sustainability Science, 10(4), 545–562. https://doi.org/10.1007/s11625-015-0328-2

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python (1st ed.). O'Reilly Media. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=443090

BMBF. (2023, April 12). Umwelt und Klima. https://www.bmbf.de/bmbf/de/forschung/umwelt-und-klima/umwelt-und-klima_node.html

Cai, M. (2021). Natural language processing for urban research: A systematic review. Heliyon, 7(3), e06322. https://doi.org/10.1016/j.heliyon.2021.e06322

FONA. (2023, April 12). FONA-Strategie Übersicht. https://www.fona.de/de/fona-strategie/

Fosso Wamba, S., Bawack, R. E., Guthrie, C., Queiroz, M. M., & Carillo, K. D. A. (2021). Are we preparing for a good AI society? A bibliometric review and research agenda. Technological Forecasting and Social Change, 164, 120482. https://doi.org/10.1016/j.techfore.2020.120482

GitHub. (2023, April 12). jsvine/pdfplumber: Plumb a PDF for detailed information about each char, rectangle, line, et cetera — and easily extract text and tables. https://github.com/jsvine/pdfplumber

Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. The International Journal of Management Education, 18(1), 100330. https://doi.org/10.1016/j.ijme.2019.100330

Jha, A. K., Ghimire, A., Thapa, S., Jha, A. M., & Raj, R. (2021). A Review of AI for Urban Planning: Towards Building Sustainable Smart Cities, 937–944. https://doi.org/10.1109/ICICT50816.2021.9358548

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., McCarthy, J. J., Schellnhuber, H. J., Bolin, B., Dickson, N. M., Faucheux, S., Gallopin, G. C., Grübler, A., Huntley, B., Jäger, J., Jodha, N. S., Kasperson, R. E., Mabogunje, A., Matson, P., . . . Svedlin, U. (2001). Environment and development. Sustainability science. Science, 292(5517), 641–642. https://doi.org/10.1126/science.1059386

Kölbel, J. F., Leippold, M., Rillaerts, J., & Wang, Q. (2022). Ask BERT: How Regulatory Disclosure of Transition and Physical Climate Risks Affects the CDS Term Structure. Journal of Financial Econometrics, Article nbac027. Advance online publication. https://doi.org/10.1093/jjfinec/nbac027

ISBN 978-88-96687-16-1 || ISSN 2280-787X

Lang, D. J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., Swilling, M., & Thomas, C. J. (2012). Transdisciplinary research in sustainability science: Practice, principles, and challenges. Sustainability Science, 7(S1), 25–43. https://doi.org/10.1007/s11625-011-0149-x

Li, Y., Osei, F. B., Hu, T., & Stein, A. (2023). Urban flood susceptibility mapping based on social media data in Chengdu city, China. Sustainable Cities and Society, 88, 104307. https://doi.org/10.1016/j.scs.2022.104307

Mondejar, M. E., Avtar, R., Diaz, H. L. B., Dubey, R. K., Esteban, J., Gómez-Morales, A., Hallam, B., Mbungu, N. T., Okolo, C. C., Prasad, K. A., She, Q., & Garcia-Segura, S. (2021). Digitalization to achieve sustainable development goals: Steps towards a Smart Green Planet. The Science of the Total Environment, 794, 148539. https://doi.org/10.1016/j.scitotenv.2021.148539

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. Journal of Information Science, 43(1), 88–102. https://doi.org/10.1177/0165551515617393

Slawski, A., Schwartze, F., & Dietrich, K. M. (2022). Transdisciplinary Synthesis Research: Challenges and Approaches of Impact-Oriented Urban and Spatial Research, 115–134. https://doi.org/10.18154/RWTH-2022-05188 (Pnd - rethinking planning 2022(1), 124-143 (2022). special issue: "Transformatives Forschen trifft Stadtentwicklung : Einführung und Reflexion / herausgegeben von Laura Brings, Lea Fischer, Agnes Förster und Fee Thissen" / pages 124-143).

SURE (2021, November 10). Integrated Planning and Development | SURE Initiatives. Sustainable Urban Regions. https://www.sustainable-urban-regions.org/themes/focus-topics/

WU, S. R., Shirkey, G., Celik, I., Shao, C., & Chen, J. (2022). A Review on the Adoption of AI, BC, and IoT in Sustainability Research. Sustainability, 14(13), 7851. https://doi.org/10.3390/su14137851

Zscheischler, J., Rogga, S., & Weith, T. (2014). Experiences with Transdisciplinary Research. Systems Research and Behavioral Science, 31(6), 751–756. https://doi.org/10.1002/sres.2274

ISBN 978-88-96687-16-1 || ISSN 2280-787X

ANNEX 1

Custom Punctuations list

‘. ‘, ‘, ‘, ‘/ ‘, ‘! ‘, ‘? ‘, ‘; ‘, ‘: ‘, ‘(‘, ‘) ‘, ‘[‘, ‘] ‘, ‘- ‘, ‘_ ‘, ‘% ‘, ‘et ‘, ‘al ‘, ‘et al ‘, ‘a ‘, ‘b ‘, ‘c ‘, ‘d ‘, ‘e ‘, ‘f ‘, ‘g ‘, ‘h ‘, ‘i ‘, ‘j ‘, ‘k ‘, ‘l ‘, ‘m ‘, ‘n ‘, ‘o ‘, ‘p ‘, ‘q ‘, ‘r ‘, ‘s ‘, ‘t ‘, ‘u ‘, ‘v ‘, ‘w ‘, ‘x ‘, ‘y ‘, ‘z ‘, ‘ ‘ ‘, ‘□ ‘, ‘ ‘ ‘, ‘ ‘ ‘, ‘ ‘ ‘, ‘- ‘, ‘WP ‘, ‘wp ‘, ‘Wp ‘, ‘| ‘, ‘• ‘, ‘.nr ‘, ‘„ ‘, ‘ ‘ ‘, ‘□ ‘, ‘ws ‘, ‘e.g ‘, ‘eg ‘, ‘e g ‘, ‘▪ ‘, ‘… ‘, ‘§ ‘, ‘à ‘, ‘± ‘, ‘ ‘ ‘, ‘°c ‘, ‘ns ‘, ‘hoc ‘, ‘aalml ‘, ‘aatse ‘, ‘– ‘, ‘·’ ‘, ‘. ‘, ‘ ‘ ‘, ‘→ ‘, ‘· ‘, ‘- ‘, ‘i.e ‘, ‘% ‘, ‘ ‘ ‘, ‘ ‘ ‘, ‘[‘, ‘] ‘, ‘{ ‘, ‘} ‘, ‘• ‘, ‘© ‘, ‘< ‘, ‘> ‘, ‘€ ‘

Additional Custom Stopwords

accordance, account, appropriate, available, change, commit, conference, consideration, contribution, country, decision, determined, different, due, example, first session, guidance, high, high confidence, high agreement, importance, information, likely, many, medium confidence, medium evidence, meeting, need, new, order, paragraph, period, possible, ppm, process, project, range, rate, relative, relevant, secretariat, support, table, tion, view, work