

Applying Text Analytics Methodology to Analyze Project Reports

Irina Arhipova¹^a, Liga Paura¹^b, Nikolajs Bumanis¹, Gatis Vitols¹^c, Vladimirs Salajevs¹^d,
Aldis Erglis²^e, Gundars Berzins² and Evija Ansonska²^f

¹*Faculty of Engineering and Information Technologies, Latvia University of Life Sciences and Technologies,
Liela Street 2, Jelgava, LV 3001, Latvia*

²*Faculty of Business, Management and Economics, University of Latvia, Aspazijas bulv. 5, Riga, LV 1050, Latvia*

Keywords: Text Mining, Text Gap Analysis, Word Co-Occurrence Analysis, Unique Terms Identification.

Abstract: The goal of this article is to develop a support methodology for Driving Urban Transition (DUT) partnership to ensure that the knowledge gathered from ERA-NET Urban Accessibility and Connectivity (EN-UAC, 2023) projects, to repeatedly identify the requirements of programme entities and define specific topics for future calls. Fifteen projects under the Horizon 2020 ERA-NET initiative have been analysed to detect similarity between projects, uniqueness of the projects, project compliance with DUT and SRIA, and gap between projects and DUT, SRIA methodology. A particular focus in the analysis was on the project “Individual Mobility Budgets as a Foundation for Social and Ethical Carbon Reduction” (MyFairShare). Text mining methods were used for documents analysis. The similarity between the documents detected by the cluster algorithm and they were compared using words, as a result, the documents were combined into three clusters: “Strategy implementation and network infrastructure”; “Transport accessibility and policy” and “Urban city mobility”. The identification of unique terms shown the terms energy, ecosystem and climate are unique for DUT&SRIA and are not found in 15 EN-UAC project applications and the next specific topics for future calls can be within the subject of energy, climate and ecosystem.

1 INTRODUCTION

The objective of numerous European, national, regional, and local transportation strategies and measures is to offer economical, reachable, secure, and dependable urban access and linkage while imposing minimal or no harm on the environment. Over time, the European Commission, along with member states, regional and local governing bodies, private enterprises, and other parties involved, have diligently worked to propel the progression of mobility and transportation systems. Nevertheless, despite these endeavours, sustainability predicaments persist, underscoring the immediate requirement for a profound shift toward a carbon-neutral, equitable, and

emission-free transportation network to climate-neutral and smart cities (Clerici Maestosi et al., 2021). The 15-Minute City (15minC) (Moreno, 2019), Downsizing District Doughnuts (DDD) and Positive Energy Districts (PED) are three innovation pillars or new approaches for urban greening (Clerici Maestosi et al., 2021). In 121 European metropolitan areas the 15-minute walking city (15-MWC) index was evaluated and authors concluded the cities in the UK, Portugal, France, and Italy have a lower 15-MWC index and worst accessibility to their daily needs compared to West Germany (Bartzokas-Tsiompras & Bakogiannis, 2023). Along with the research on the 15minC, the researches on walking accessibility (Bartzokas-Tsiompras et al., 2023; Ione Avila-Palencia et al., 2022) and the behaviour of the

^a <https://orcid.org/0000-0003-1036-2024>

^b <https://orcid.org/0000-0002-6625-9475>

^c <https://orcid.org/0000-0002-4131-8635>

^d <https://orcid.org/0000-0001-8545-4690>

^e <https://orcid.org/0000-0002-1302-527X>

^f <https://orcid.org/0000-0002-7029-2745>

inhabitants (Arhipova et al., 2023) remains even more relevant.

The urgent and essential transformation of our planet demands an immediate and critical shift in society. This shift necessitates individuals to re-evaluate their behaviours and outlooks, which in turn involves a process of learning. Within the framework of transitioning toward entirely new socio-economic practices as envisioned by sustainable urban mobility, learning goes beyond the mere acquisition of knowledge.

Establishing an environment that encourages the exploration of diverse forms of knowledge and various ways of understanding is vital to cultivate learning centred on sustainability. This learning revolves around the notion of accessibility and connectivity (ACUTE, 2023).

The goal of this article is to develop support methodology for Driving Urban Transition (DUT) partnership to ensure that the knowledge gathered from ERA-NET Urban Accessibility and Connectivity (EN-UAC, 2023) projects and beyond is accessible as assistance for all stakeholders within the forthcoming Horizon Europe Partnership Program (DUT, 2023). This support aims to drive the urban transition toward a future that is both sustainable and conducive to a high quality of urban living.

In order to fulfil the goal of the article, it is essential to repeatedly identify the requirements of programme entities and define specific topics. This involves addressing research gaps, implementation deficiencies, using existing outcomes as a foundation for future calls, staying abreast of new developments and trends in urban accessibility and connectivity innovations and social practices, and exploring potential synergies with other funding programmes.

2 METHODOLOGY

Usually the text is viewed as a means of conveying meaning, but it can also be seen as a structured sequence of words or an unstructured collection of words. These words can be represented in ways that enable analysis without being limited by grammatical structure. Texts can be subjected to statistical examination by exploring the interrelationships between words within a specific text and comparing these relationships to those observed in other texts (Robinson & Silge, 2017). The utilization of computational methods involves:

- complementary enhancement to qualitative analysis approaches;

- scalability, capable of processing numerous documents (such as proposals, mid-term reports, final reports, etc.);
- application of text analysis method: unordered bag of words;
- employment of data mining techniques to unearth novel insights for upcoming calls: ordered string of words;
- manual customization options, allowing for the definition of stop-words and precision;
- while not flawless, it operates in a semi-automated manner.

The set (=corpus) of 15 ERA-NET Urban Accessibility and Connectivity (EN-UAC) project' proposals (Basecamp, 2023), Driving Urban Transitions to a Sustainable Future Roadmap 2022 (DUT, 2023) and JPI Urban Europe's Strategic Research and Innovation Agenda texts (=documents) is analysed (SRIA 2.0., 2023).

The process of text analytics or text mining involves multiple sequential steps (see Figure 1). The initial step is *text normalization* by eliminating punctuation, cases, and numbers. It is essential to remove common words, often referred to as stopwords, as they can obscure the more significant and relevant ones. This includes removing standard English stopwords (Vijayarani et al., 2015).

The next step is *words frequency analysis*: data visualization by generating Wordclouds. After that follows *text similarity analysis*: the idea is to measure the similarity between two words or texts in order to cluster documents into distinct groups. The ultimate goal is to pinpoint a specific collection of documents that maximizing the level of similarity. A similarity model can be employed not only for clustering documents but also for clustering words.

Word co-occurrence analysis step is necessary to check co-occurrence between DUT/SRIA and projects' proposals the combinations of files (entities in corpus) are created, for example, DUT – MyFairShare project proposal text, SRIA 2.0 – MyFairShare project proposal text (Arhipova et al., 2022), etc. and for each combination:

- a corpus, containing two files is created;
- out of each file, 10 most used terms are selected;
- the list of unique words is created, using 10 most used terms from each file;
- unique term combinations are created using unique words;
- each combination is then checked for belonging to corpus entities;

- the results are stored as matrix where $X=Y=\text{unique word}$ and XY value equals to the number of occurrences of particular combination.

As the result of word co-occurrence analysis, the matrix, as .csv file, is passed as input file to visualization software Gephi (Gephi, 2023).

The last step of the process of text analytics is *unique terms identification and text gap analysis*. The word frequency provides representations of content that include commonly occurring terms. However, to identify topics that are specific to each text, it is necessary a different approach.

The most commonly employed method for achieving this is by utilizing the term frequency-inverse document frequency (TF-IDF) measure, which stands for term frequency-inverse document frequency (Robinson & Silge, 2017).

Word frequency, text similarity analysis, unique terms identification, and text gap analysis was done using R (R Core Team, 2021) and R Text Mining Package *tm* (Feinerer, Hornik, 2020).

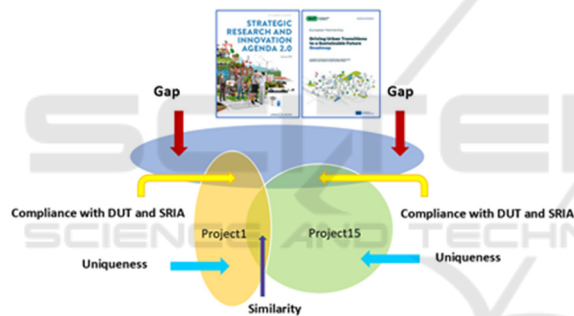


Figure 1: The text analytics scheme.

3 RESULTS

Fifteen projects under the Horizon 2020 ERA-NET Cofund Urban Accessibility and Connectivity (EN-UAC) initiative have been analysed, with particular focus on the MyFairShare project titled “Individual Mobility Budgets as a Foundation for Social and Ethical Carbon Reduction” (MyFairShare, 2023).

The research devised a theoretical notion of “equitable” mobility budgets, considering an individual's social and spatial circumstances (such as socioeconomic status and proximity to places of interest). These budgets assist individuals in comprehending the consequences of their mobility decisions, and also provide policymakers with insights into necessary interventions to alleviate excessive strain on local mobility budgets (Arhipova



Figure 3: Created wordcloud of DUT document.

The most frequently five words for DUT are urban (4.2%), the Positive Energy Districts Transition Pathway or ped (2.0%), city (2.0%), energy (1.5%) and innovation (1.4%) (see Figure 3).



Figure 4: Created wordcloud of SRIA 2.0. document.

The most frequently five words for SRIA 2.0. are urban (8.0%), innovation (2.0%), city (1.5%), agenda (1.4%), and strategic (1.4%) (see Figure 4).

3.2 Similarity Analysis

The goal of the text similarity analysis is to measure the similarity between words or texts, enabling the grouping of documents into distinct clusters. The objective is to identify a specific set of documents that maximise similarity levels. In order to measure the similarity between texts, the hierarchical cluster algorithm is used for clustering words and documents. Ward's or minimum variance method was used to analyse distance between clusters, which is popular in linguistic field (Szmrecsanyi, 2012).

The hierarchical clustering results by project proposals text is shown by dendrogram (see Figure 5).

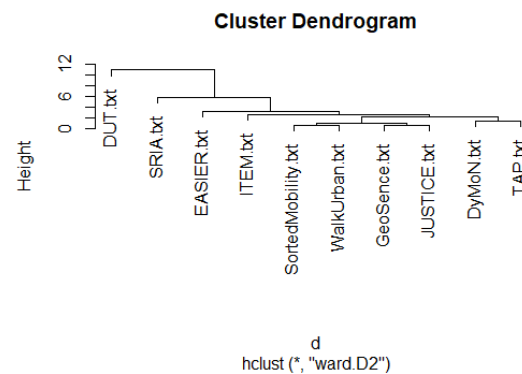


Figure 5: Dendrogram of 15 EN-UAC projects proposal, DUT and SRIA 2.0. texts.

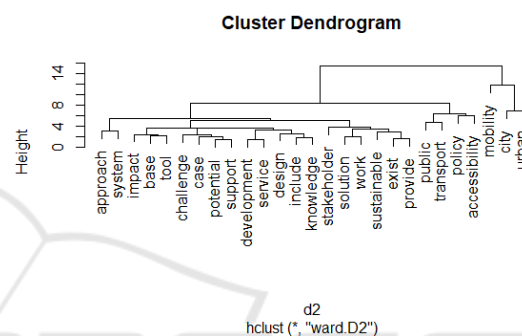


Figure 6: Dendrogram of 15 EN-UAC projects proposal, DUT and SRIA 2.0. words.

Based on the comparison with DUT and SRIA 2.0, it can be concluded that the projects exhibiting the highest similarity are as follows: CATAPULT and SmartHub, then Ex-TRA, ASAP, DyMoN, GeoSense, MyFairShare, TuneOurBlock, COCOMO, EASIER, WalkUrban, TAP, SortedMobility, ITEM, JUSTICE, SRIA and DUT (see Figure 4).

Similarity models can be used not only for document clustering but also for word clustering. The clustering result by EN-UAC project proposals, DUT and SRIA 2.0 words is shown by dendrogram (see Figure 6).

According to the clustering result the projects proposals, DUT and SRIA 2.0 words are divided to the tree clusters. Terms grouped near to each other are more frequently found together. It can be inferred that words sharing the greatest similarity can be grouped together into the following clusters of similarity:

- 1st cluster: approach, system, impact, base, tool, challenge, case, potential, support, development, service, design, include, knowledge, stakeholders, solution, work, sustainable, exit, provide;

- 2nd cluster: public, transport, accessibility, policy;
- 3rd cluster: mobility, city, urban (see Figure 6).

The names of the clusters' can be: 1st cluster "Strategy implementation and Network infrastructure"; 2nd cluster "Transport accessibility and policy" and 3rd cluster "Urban city mobility".

Based on the words analysis between DUT and SRIA documents three clusters in total with ten most frequent words was found (see Figure 7).

Word Urban is higher in the plot compare to other words that appear urban are more common within the corpus. Word PED (Positive Energy Districts Transition Pathway) is only occurs in DUT document and as well as word energy is more frequent in DUT.

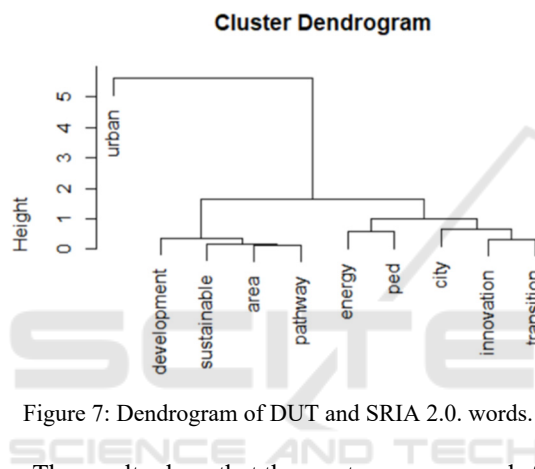


Figure 7: Dendrogram of DUT and SRIA 2.0. words.

The results show that the most common words for DUT and SRIA are urban, city, innovation, energy & transition, but for ENUAC projects - mobility & city. The compliance or main subjects for DUT&SRIA&EN-UAC projects are Strategy and model development, implementation & network infrastructure; Transport accessibility & policy; Urban city mobility.

3.3 Word Co-Occurrence Analysis

The word co-occurrence analysis (Lancia, 2007) between DUT, SRIA 2.0 and 15 projects' proposals are executed. Co-occurrence analysis is focused on to find out words relationship network and words, which are frequently used together. Force Atlas algorithm is used for graphs' layout calculation.

The weight of each connection is the visually differentiated: the stronger the line, the stronger the connection (i.e., more occurrences), while the position of the term depends on its authority/importance.



Figure 8: Compliance of MyFairShare project proposal to DUT.

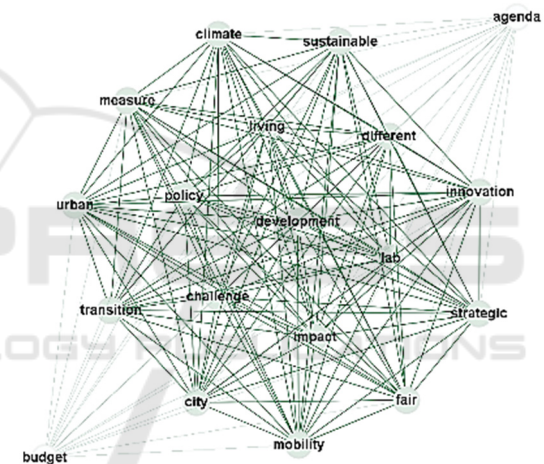


Figure 9: Compliance of MyFairShare project proposal to SRIA 2.0.

Compliance and gap of MyFairShare project proposal, DUT and SRIA 2.0., using word co-occurrence analysis, is given in Figure 8 and Figure 9.

3.4 Unique Terms Identification and Text Gap Analysis

While word frequency provides representations of content that include commonly used terms, the identification of topics specific to each text is achieved using the *TF-IDF* measure, which stands for term frequency-inverse document frequency.

For topics identification that are specific to each text the *TF-IDF* measure is used and is calculated as *TF* and *IDF* product:

$$TF-IDF = TF * IDF, \quad (1)$$

where

$$IDF = \frac{1}{\log\left(\frac{\text{No. of documents in corpus}}{\text{No. of documents in corpus contain terms}}\right)} \quad (2)$$

$$TF = \frac{\text{No. of times terms appears in document}}{\text{total No. of terms in document}} \quad (3)$$

Based on the analysis, the gap between DUT, SRIA, and 15 EN-UAC projects can be observed in the following 10 words: energy, regenerative, ecosystem, neutral, climate, transformation, economy, dilemma, transition and joint (Figure 10).

The highest *TF-IDF* (tf-idf) values for DUT & SRIA documents were PED, energy, CURE, climate and 15minC (Figure 10). Comparison with DUT Roadmap text & word co-occurrence analysis shows that energy & transition are words with a lower co-occurrence of two adjacent terms in a text corpus. The similar result is found in the identification of unique terms for DUT&SRIA: energy, ecosystem & climate. The gap between DUT&SRIA and ENUAC projects lies within the subject of energy, climate and ecosystem.

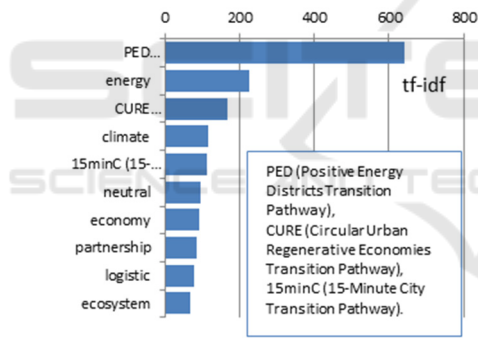


Figure 10: TF-IDF of unique terms for DUT & SRIA documents.

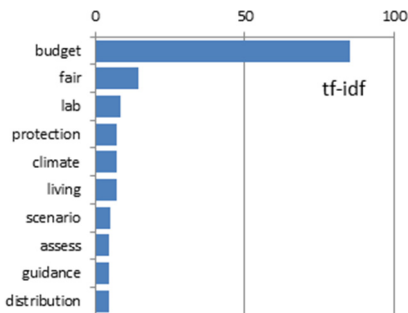


Figure 11: TF-IDF of unique terms for MyFairShare proposal.

The top 10 unique words for MyFairShare project are budget, fair, living lab, protection, climate, scenario, guidance, distribution and constraint (Figure 11).

4 CONCLUSIONS

Published research suggest that text mining methods significantly reduced manual work compared to conventional methods as well as also enables us to interpret initial information of the projects proposals, to estimate bias of the documents and to understand the connections between them.

Our analysis indicates the most common words for DUT and SRIA are urban, city, innovation, energy and transition and for EN-UAC projects are mobility and city. Based on text analysis the compliance or main subjects for EN-UAC projects are strategy and model development, implementation and network infrastructure; transport accessibility and policy; urban city mobility.

The identification of unique terms for DUT&SRIA shows that the terms energy, ecosystem, and climate are not found in 15 EN-UAC project applications. For MyFairShare project words budget, fair, living lab, protection, and climate were with higher tf-idf value and the terms budget and fair are unique and not found in the other projects.

Data analysis suggests also that the gap between DUT&SRIA and EN-UAC projects is within the subject of energy, climate and ecosystem.

As the number of evaluated projects within the EN-UAC programme rises, text analysis methods will facilitate faster and more scalable comparative assessments among them compared to traditional approaches.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Horizon 2020 ERA-NET Cofund Urban Accessibility and Connectivity (EN-UAC) projects “Individual Mobility Budgets as a Foundation for Social and Ethical Carbon Reduction” (MyFairShare) and “Accessibility and Connectivity knowledge hub for Urban Transformation in Europe” (ACUTE).

REFERENCES

- ACUTE (2023). *JPI Urban Europe's EN-UAC Knowledge Hub Accessibility and Connectivity knowledge hub for Urban Transformation in Europe*. <https://jpi-urbaneurope.eu/news/the-knowledge-hub-a-platform-for-gathering-knowledge-and-experiences-for-a-sustainable-city/>
- Arhipova, I., Bumanis, N., Paura, L., Berzins, G., Erglis, A., Rudloff, C., Vitols, G., Ansonska, E., Salajevs, V. and Binde, J. (2023). Municipal Transport Route Planning Based on Fair Mobility Budget. *Rural Sustainability Research*, 50(345), 44-58. <https://doi.org/10.2478/plua-2023-0014>
- Arhipova, I., Bumanis, N., Paura, L., Berzins, G., Erglis, A., Vitols, G., Ansonska, E., Salajevs, V., Binde, J. (2023). Optimizing transport network to reduce municipality mobility budget. *Proceedings of the 5th International Conference on Finance, Economics, Management and IT Business (FEMIB 2023)*, 38-47.
- Bartzokas-Tsiompras A. and Bakogiannis E. (2023). Quantifying and visualizing the 15-Minute walkable city concept across Europe: a multicriteria approach. *Journal of Maps*, 19(1), <https://doi.org/10.1080/17445647.2022.2141143>
- Bartzokas-Tsiompras, A., Bakogiannis, E. and Nikitas A. (2023). Global microscale walkability ratings and rankings: A novel composite indicator for 59 European city centres. *Journal of Transport Geography*, 111. <https://doi.org/10.1016/j.jtrangeo.2023.103645>.
- Basecamp (2023). *JPI Urban Europe on Basecamp - a space dedicated for exchanges with to with the first 15 EN-UAC Co-funded projects coordinators*. <https://3.basecamp.com/3660243/projects>
- Clerici Maestosi, P., Andreucci, M.B. and Civiero, P. (2021). Sustainable Urban Areas for 2030 in a Post-COVID-19 Scenario: Focus on Innovative Research and Funding Frameworks to Boost Transition towards 100 Positive Energy Districts and 100 Climate-Neutral Cities. *Energies*, 14(1), 216. <https://doi.org/10.3390/en14010216>
- DUT (2023). *JPI Urban Europe's programme Driving Urban Transitions to a Sustainable Future*. <https://dutpartnership.eu>
- EN-UAC (2023). *JPI Urban Europe's ERA-NET Urban Accessibility and Connectivity partnership*. <https://www.era-learn.eu/network-information/networks/en-uac>
- Feinerer, I., Hornik, K. (2020). tm: Text Mining Package. R package version 0.7-8. <https://CRAN.R-project.org/package=tm>
- Gephi (2023). *The Open Graph Viz Platform*. <https://gephi.org/>
- Lancia, F. (2007). Word Co-Occurrence and Similarity in Meaning. Some Methodological Issues. *Linguistics, Computer Science Corpus ID: 46954203*, <https://api.semanticscholar.org/CorpusID:46954203>
- Moreno C. The 15 minutes-city: for a new chrono-urbanism! (2019). <http://www.moreno-web.net/the-15-minutes-city-for-a-new-chrono-urbanism-pr-carlos-moreno/>.
- MyFairShare (2023). *Horizon 2020 ERA-NET Cofund Urban Accessibility and Connectivity (EN-UAC) project "Individual Mobility Budgets as a Foundation for Social and Ethical Carbon Reduction"*. <https://jpi-urbaneurope.eu/project/myfairshare/>
- Olsen, J. R., Nicholls, N., Caryl, F., Mendoza J. O., Panis, L. I., Dons, E., Laeremans, M., Standaert, A., Lee, D., Avila-Palencia, I., de Nazelle, A., Nieuwenhuijsen M. and Mitchell, R. (2022). Day-to-day intrapersonal variability in mobility patterns and association with perceived stress: A cross-sectional study using GPS from 122 individuals in three European cities. *SSM - Population Health*, 19. <https://doi.org/10.1016/j.ssmph.2022.101172>.
- R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>.
- Robinson, D., Silge, J. (2017). *Text Mining with R: A Tidy Approach*. United States: O'Reilly Media.
- SRIA 2.0 (2023). *JPI Urban Europe's Strategic Research and Innovation Agenda*. <https://jpi-urbaneurope.eu/about/sria/sria-2-0/>
- Szmrecsanyi, B. (2012). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge University Press.
- Vijayarani, S., Ilamathi, M.J. and Nithya, M. (2015). Preprocessing Techniques for Text Mining – An Overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.