sciendo

Transport and Telecommunication, 2024, volume 25, no. 3, 266-277
Transport and Telecommunication Institute, Lauvas 2, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2024-0019

HORIZON 2020 PROJECT ANALYSIS BY USING TOPIC MODELLING TECHNIQUES IN THE FIELD OF TRANSPORT

Domokos Esztergár-Kiss[1,2]

[1]Budapest University of Technology and Economics (BME), Faculty of Transportation Engineering and Vehicle Engineering (KJK), Department of Transport Technology and Economics (KTKG)
Budapest, Hungary, 1111 Budapest, Műegyetem rkp. 3.
[2]HUN-REN Institute for Computer Science and Control (SZTAKI)
Budapest, Hungary, 1111 Budapest, Kende utca 13-17.
esztergar@mail.bme.hu

Understanding the main research directions in transport is crucial to provide useful and relevant insights. The analysis of Horizon 2020, the largest research and innovation framework, has been already realized in a few publications but rarely for the field of transport. Thus, this article is devoted to fill this gap by introducing a novel application of topic modelling techniques, specifically the Latent Dirichlet Allocation (LDA), in the Horizon 2020 framework for transport projects. The method is using the Mallet software with pre-examined code optimizations. As the first step, a corpus is created by collecting 310 project abstracts; afterward, the texts of abstracts are prepared for the LDA analysis by introducing stop words, optimization criteria, the number of words per topics, and the number of topics. The study successfully uncovers the following five main underlying topics: road and traffic safety, aviation and aircraft, mobility and urban transport, maritime industry and shipping, open and real-time data in transport. Besides that, the main trends in transport are identified based on the frequency of words and their occurrence in the corpus. The applied approach maximizes the added value of the Horizon 2020 initiatives by revealing insights that may be overlooked using traditional analysis methods.

Keywords: Horizon 2020, innovation framework, Latend Diriclet Allocation, topic modelling technique

1. Introduction

The Horizon 2020 framework, the largest EU research and innovation initiative, is crucial in advancing science, promoting international cooperation, and solving relevant societal concerns. The framework aimed to encourage the realization of sustainable and innovative research projects from 2014 to 2020 contributing to the economy and society (Giarelis and Karacapilidis, 2021). Moreover, the Cordis portal is a huge database of the Horizon 2020 projects of various fields, and this vast amount of information provides the opportunity to retrieve valuable insights into specific scientific fields (Giarelis and Karacapilidis, 2021). While many scientists participated in the projects, few of them tried to identify potential development approaches that help to create an impact through the analysis of the project data.

There is a variety of Horizon 2020-related papers where different approaches are used to analyze the project data. For example, Suran *et al.* (2019) conducted an exploratory analysis of 10 projects related to crowd-oriented data science and collective intelligence. The researchers applied a genome model to identify the comprehensiveness of the model for text data analysis and to categorize the data with similar information contexts. However, solely a few projects were analyzed by using qualitative approaches.

Several projects were analyzed by Marzi *et al.* (2022) where the authors found 70 projects about electro-fuels by using the CORDIS portal. The scholars distinguished the main research areas in electro-fuel development as well as identified the key features, project aims, integration levels, and the outcomes of the projects. The authors presented some recommendations for scientists and specialists, who can use the information to make contribution in this field. However, the analysis of the projects was made based on desk research by comparing the data with different research tools without using any statistical analysis.
Furthermore, Radoselovic (2019) investigated the Horizon 2020 incentives through the analysis of the CORDIS portal where dependencies between allocated budgets to countries and their relationships with such economic factors as GDP, unemployment rate, and average salary, were presented (Radoselovics, 2019). A close correlation was found between the invested money and GDP per capita while presenting clusters with countries of similar investment and GDP rates. Nevertheless, this research did not provide a deeper analysis of the projects but focused solely on some economic factors and its correlations.

By revealing these research gaps, this paper focuses on identifying the main research areas in the transport industry to provide a wide overview of the conducted research projects in the topic called Transport and Mobility. The most important contribution of current paper is making an exploratory analysis of the Horizon 2020 projects to identify key research pathways and development directions for the stakeholders' benefit in the transport field by using statistical data and text mining techniques. In order to provide sophisticated and valuable results, an advanced text-mining method is applied. The Latent Dirichlet Allocation (LDA) topic modelling tool is used to retrieve the key research areas and identify the research gaps in the field thus improving the strategic planning in transport and enhancing the potential collaborations between governmental bodies, scientific institutes, and professional companies. The LDA technique is chosen as it provides the ability to identify topics hidden from the readers in a large database of documents and has not been applied before for an in-depth analysis of the Horizon 2020 framework. The method can reveal the primary research topics by identifying frequently occurring words and detecting the frequency of their occurrences with other words.

This study uses the LDA to find latent topics within the Horizon 2020 framework, which provides novel insights. While numerous facets of Horizon 2020 have already been subjects of prior research, the use of the LDA method offers a computational tool to uncover hidden themes and patterns in the enormous corpus of research texts. Current study provides a thorough examination of the underlying theme structures, interdisciplinary partnerships, knowledge gaps, and the temporal evolution of research directions in the Horizon 2020 framework. By revealing latent information and enabling a richer understanding of the research environment, the application of the LDA provides a valuable approach that fosters strategic decision-making and facilitates future research and innovation initiatives.

This paper is organized into six main sections. Section 1 provides a general overview of the topic to help the readers in understanding its importance and scale. Section 2 focuses on the literature review thus covering the use of the LDA method in the transport industry. Section 3 presents the detailed description of the methodology. Section 4 aims to interpret the results of the applied method. The discussion of the outcomes is demonstrated in Section 5, while the conclusion is found in Section 6.

2. Literature review

In the era of digitalization where most texts are in electronic format and relatively openly available on the Internet, text mining is important in the scientific field (Kherwa and Bansal, 2020). Text mining is used in several scientific publications and is gaining more and more popularity in all scientific fields. The main reason behind the growing popularity of this methodology is that it can analyze textual data and find information hidden from the readers at first glance (Uys *et al.*, 2008). The method is especially effective in cases where there is a large number of text documents.

Text mining is a powerful and efficient analysis method for large text databases. This technique uses artificial intelligence aiming to retrieve information from various collections of text documents by applying machine learning techniques and natural language processing. Text mining is very popular in the transport industry, especially in case of massive databases of text documents (Gopalakrishnan and Khaitan, 2017). Usually, so many texts appear on the Internet or social media where a huge amount of unsorted data, like comments and reviews, is listed. For example, Serna and Gasparovic (2018) analyze reviews from the social media website of TripAdvisor where the researchers focus on the travellers' comments and try to figure out the impact of social media on transport and tourism planning. The methodology includes such steps as collecting data manually, preparing the texts for processing, and applying various machine learning techniques to analyze the data. Moreover, the authors create a dashboard that visually reflects the TripAdvisor reviews by using several data collection, analysis, storage, and visualization tools which categorize the reviews for future improvement of transport services.

In other publications, Twitter is used where people express their opinion about different transport services. For example, Magrebi *et al.* (2015) study tweets about smart transport systems in Sydney. The authors apply text mining and clustering methods to identify the main keywords and the frequency of these keywords in tweets. This approach helps to find what users are most worried about, to better understand the content of text data, and to give recommendations on how the results can be utilized by transport operators. The research work applies some statistical and linguistic techniques to analyze the content of tweets; however, the LDA method is not applied for this purpose.

To the best of the authors' knowledge, there are not many publications applying the LDA on a huge database in the transport industry. A similar study (Gopalakrishnan and Khaitan, 2017) shows the utilization

of more than 14,000 research papers of the Transport Research Board (TRB) database. Gopalakrishnan and Khaitan collect the text of the papers to show patterns and predict the future of big data in transport. The methodology is based on data collection and processing by machine learning and natural language processing techniques. By using neural network models, the scholars can identify the most adequate funding agency for researchers and their scientific areas with 75% accuracy. The combination of machine learning, natural language, and neural networks helps authors to create a framework for the prediction of funding agencies based on the research topic. Although the paper uses a complex mixture of techniques, it does not utilize the LDA for large text data analysis.

Similarly, Liu and Yung (2022) collect reported incidents in rail transport to determine the level of safety risks. The authors apply machine learning and text mining techniques together with deep learning tools. As a result, a method is developed to improve the effectiveness of risk assessment and to reduce the probability of hazards in the railway segment. The proposed approach includes various techniques that allow the prediction and minimization of future risks based on the labeled accident report data. Nevertheless, handling non-labeled data is not possible without the initial pre-processing of information, and the authors do not focus on retrieving the hidden information of the accident reports.

The use of different text mining techniques can bring great benefits as they take the word order used in the texts into account; however, they do not consider those topics of the text that are hidden at first glance. Therefore, topic models can help to reveal these issues without relying on the word order. Topic modelling is a tool aiming to uncover the main tendencies and trends in big textual datasets where usually the textual format is considered as unstructured or unlabeled (Blei, Ng, and Jordan 2003). As an outcome, such models give an idea of how a particular industry has evolved over time while highlighting development directions as well as providing an opportunity to a wide range of analyses.

The topic modelling is based on the following methods: latent semantic analysis, probabilistic latent semantic analysis, LDA, and correlated topic modelling. The LDA is a more advanced model of the first two options as it covers the disadvantages of these techniques (Alghamdi and Alfalqi, 2015). The correlated topic modelling covers some drawbacks of the LDA when it is not possible to find a connection between the identified topics. However, it is merely required in some special cases. At the same time, the LDA is the most common and widely-used technique as it gives efficient and desirable results in topic modelling.

For example, Liu and Cheng (2020) apply the LDA method to identify travel patterns in smart card-based automated fare collection. Based on the textual data of the Oyster card system in London, the researchers extract some travel patterns to provide recommendations on transport planning while better understanding the travellers' behaviour. Another adoption of the LDA method in transport is shown in the publication of Rachman *et al.* (2021), who apply the method on Twitter text data. The research is an analysis of the transport systems in Jakarta, which helps to understand the trends and provides a ranking of the transport modes in the city by extracting some topics of traveller reviews and comments on Twitter. Even though the LDA is used on massive textual data, the authors use the technique to identify travel patterns and understand some trends important for future transport development. However, the LDA is not applied for the analysis of huge research frameworks.

Furthermore, the LDA can be used to identify primary topics and trends in transport. For example, the LDA is used by Bai *et al.* (2021) to analyze the main trends in maritime research. The authors collect more than 3,000 articles from 23 peer-reviewed journals. As a result, the researchers find some trends in the textual data, address several research gaps, and provide recommendations to maritime specialists. In another research, Suh (2021) applies the LDA method to identify the causes of incidents in the US highway system by using the national traffic accident reporting system, which contains over 21,000 documents. By applying the LDA method, it is possible to identify the 10 most common topics in the text database related to traffic incidents. The study presents main trends in different sectors of road transport and provides recommendations for further use by the government thus enhancing safety on the roads.

Another example is presented by Sun and Kirtonia (2020), who use the Transport Research International Documentation (TRID) database. The authors extend the LDA method with geoRegion modelling. The method is applied to identify the main topics of the TRID and its distribution over geographic regions, which helps to cover the knowledge gaps regarding scientific studies and regions where these studies have been performed. Even though the LDA has been applied before on big textual transport databases, it has not been applied on one of the largest European research project databases called the Horizon 2020, which drives the major research pathways including the transport sector.

In other publications various text mining techniques are used to identify some main research topics. For example, Giarelis and Karacapilidis (2021) analyze the Horizon 2020 framework by processing around

30,000 projects and 80,000 publications related to the projects. The authors use text mining and a graph-based approach to identify hidden information and to discover the latent links between the documents. Having collected a large set of textual data, the researchers can identify the key phrases and topics of documents thus creating an approach for analyzing progress in scientific fields. According to the authors, revealing the results of the text mining helps to create the right research plan thus contributing to better results in future research works. However, the proposed approach uses unsupervised learning to extract the key phrases from the text, while the LDA technique allows the formation of the documents into groups based on the topics and identifies the share of topics in particular documents thus giving a variety of analysis opportunities. In this regard, the LDA modelling tool may provide more opportunities to analyze scientific frameworks by utilizing more information.

Finally, Malloci *et al.* (2020) use the CORDIS database to analyze the Horizon 2020 framework by ranking innovation insights from research projects. The authors apply artificial text analysis collecting a large set of documents where the abstracts of each project are used to identify the primary problems and challenges. As a result, by using text mining and natural language processing, the authors rank the main problems found in the Horizon 2020 framework, which can help in future project preparations by presenting the most relevant and recent insights. Nevertheless, the LDA is not applied in the study to identify the problems and challenges, and it is not specialized in the transport and mobility field.

Based on the literature review, the following statements can be formed:

- Frequently, the Horizon 2020 framework and its projects are analyzed in general terms without specifically focusing on the transport sector.
- The Horizon 2020 framework is analyzed in a few publications by using text mining techniques.
- Usually, in these publications, topic modelling techniques are not applied even though the goals are similar to those of current study.
- In the field of transport, the LDA is not applied on big frameworks, such as the Horizon 2020.

Based on the literature review, limited analysis of the Horizon 2020 framework with text mining has been done before in the field of transport. Due to the fact that the projects are presented in text format on the CORDIS portal, the use of topic modelling can provide significant results for transport experts and researchers. The application of the LDA method in the Horizon 2020 framework promises to expose underlying topic structures across research initiatives thus providing a novel solution to uncover untapped knowledge, promote multidisciplinary cooperation, and maximize the added value of future projects.

3. Methodology

In this study, a text mining methodology called topic modelling using the LDA technique is chosen, which allows the processing of a large amount of textual information. Figure 1 shows a schematic overview of the applied method, which can be divided into three main steps. The first step is the data collection where the term corpus is used for creating the dataset. In the second step, the modelling process is shown where the topic modelling software called MALLET is used. The final step illustrates the results of the topic modelling where the main topics, the distribution of documents, and the frequency of words are identified.

There are several topic modelling techniques where the LDA is an evolution based on previous improvements overcoming the disadvantages of older techniques thus giving more significant results. The main advantages of the LDA are the followings (Alghamdi and Alfalqi, 2015; Dang and Ahmad, 2014):

- The LDA can identify topics from a large amount of text data.
- It can reveal the percentage content of the identified topic in each text document.
- It can provide the frequency of the occurring words in the corpus.
- The LDA can form meaningful topics compared to other topic modelling techniques.
- It can apply the method in an easy way in various programming environments (e.g., MALLET).

Based on the review and the advantages and disadvantages of the LDA method, this topic modelling method is chosen to analyze the Horizon 2020 framework and to identify the relevant directions in the transport sector. A topic in this regard is a cluster of words frequently used in the text. Thus, topic modelling algorithms can connect frequent words in a logical way and distinguish their use in different contexts.

As the first step, the project abstracts are searched in the Horizon 2020 framework by using the CORDIS website. In the next step, the texts are processed by introducing some stop words, the criteria of the optimization, the number of words per topic, and the number of topics. Finally, the generated topics are analyzed according to the frequency of the words and the distribution of the data corpus elements among

the generated topics. In Figure 1, the curved forms represent the main steps and actions, the oval forms are related to the optimization criteria, while the rectangular forms show the results.

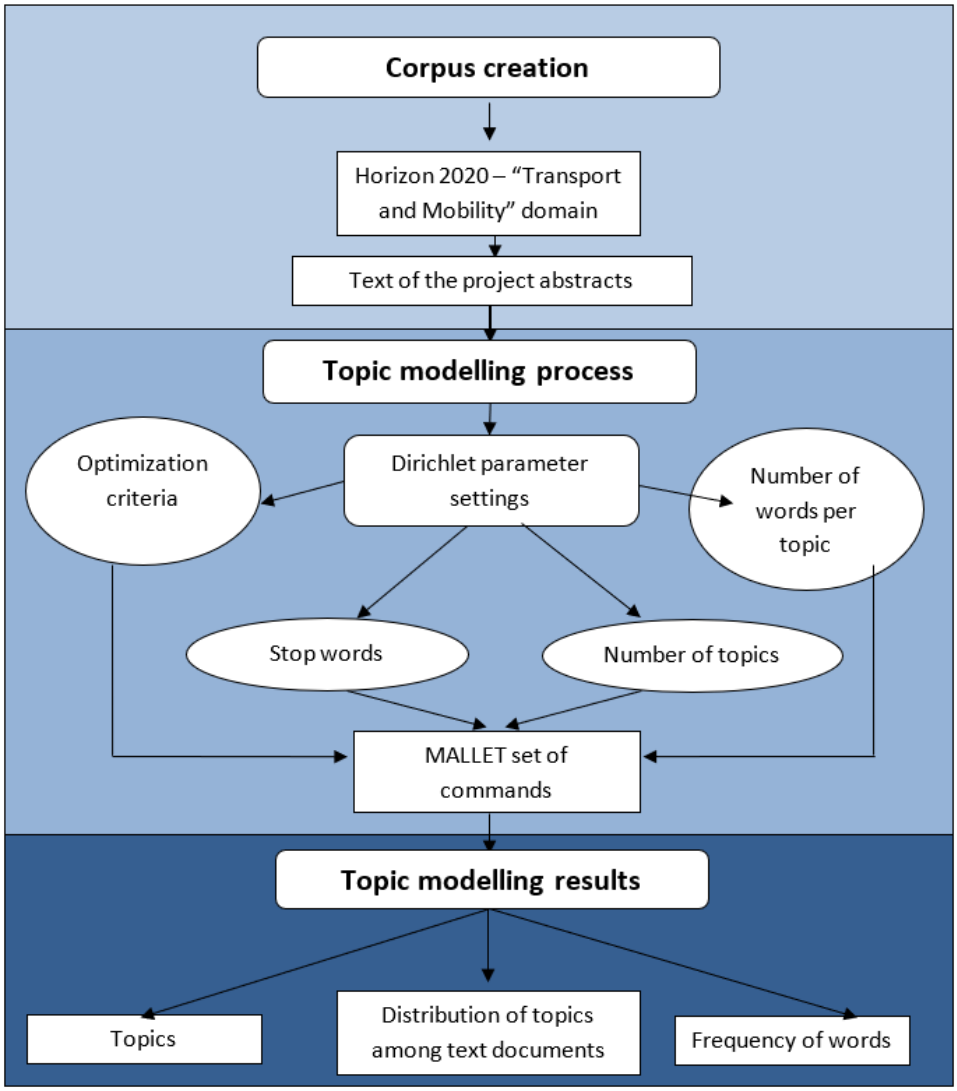

Figure 1. Topic modelling process

3.1. Corpus creation

To start the topic modelling, a database is created where the texts of the project abstracts are extracted for further analysis and modelling. In order to find the transport and mobility-related projects, the CORDIS website, which provides detailed information on every project, is used (European Commission, 2024). The project ID is the unique identification of a project, which is connected to the title of the project, the acronym, the total costs, the participants of the project, and to the project abstracts together with the project objectives and results. Among the participants of a project, there are coordinators (who are responsible for the project) and participants (who contribute to the objectives), who are from different European countries. The project abstract is a detailed explanation of the aims, goals, and key pathways.

3.2. Topic modelling process

MALLET is Java-based software applied for statistical natural language processing, for the classification of various documents, topic modelling, clustering, and extraction of information based on machine learning applications (Serna and Gasparovic, 2018). To identify the main topics of the Horizon 2020 framework, the suitable commands are chosen and tested to familiarize with the potential outputs of the topic modelling. There are two sets of commands, which provide specific functions dealing with the

preparation of the text dataset and manipulating it by setting the criteria, such as the optimization parameters, the number of desired words in topics, the number of desired topics, and number of iterations.

The MALLET software has its own list of stop words that are not counted for the analysis. Such stop words are usually prepositions (e.g., "a", "the", "of") or entering words (e.g., "moreover", "furthermore"). The possibility of including own stop words is provided, too. Such stop words should be included where the existence of these words can impact the credibility of the topic modelling. For instance, if a specific abbreviation (e.g., EU) is present, it can affect the interpretation of the identified topic. In this case, the presence of such abbreviations does not impact the results; thus, no additional stop words are included.

When running the program, two sets of commands are created to model the main topics, while the final sets of commands identify the topics of the Horizon 2020 framework.

- Text dataset processing and upload:

```
bin\mallet import-dir --input sample-data\web\en --output tutorial.mallet --keep-sequence --remove-stopwords
```

- Topic modelling and processing of words:

```
bin\mallet train-topics --input tutorial.mallet --output-state topic-state.gz --output-topic-keys tutorial_keys.txt --output-doc-topics tutorial_composition.txt --num-top-words 10 --num-topics 10 --optimize-interval 10
```

4. Results

4.1. Corpus creation results

The text dataset (i.e., corpus) is created as input data for the topic modelling, and the project objectives are stored based on the abstracts of the projects. Abstracts are chosen because they are concise and provide high-quality descriptions of the projects where the main problems, goals, and solutions are described along with the keywords. The formation of the corpus is based on the information from the CORDIS website where projects are selected from the Transport and Mobility domain, more specifically from the Transport and Mobility calls. The Transport and Mobility domain consists of 146 call topics; 52 of them were realized in 2014 and 2015, 39 in 2016 and 2017, 55 in 2018 and 2020. In each call topic, there are projects where the project partners should accomplish the call topic goals in collaboration. A total amount of 310 projects is collected.

Each project is coordinated by an institute in the European Union with other participating countries. Organizations from 22 countries act as coordinators. Figure shows the distribution of the project coordinators by the regions of the European Union. In most cases (46%), the countries of Western Europe coordinate projects, Southern and Northern European countries are coordinators in 32% and 20% of the projects, respectively. Eastern European countries are coordinators in solely 2% of the projects. Germany is the most frequently chosen coordinator country with 54 realized projects, while the second one is Spain with 45 coordinated projects, which is followed by 36 coordination from Belgium.

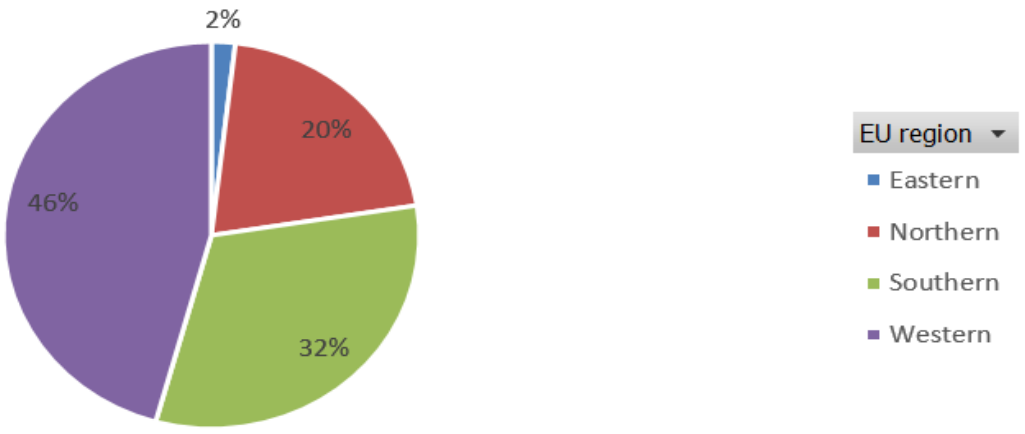

Figure 2. The coordinators' distribution based on EU sub-regions

Many more institutions and countries took part in the projects, not merely from Europe, but from all corners of the planet. A total of 4,535 participants supported the projects and contributed to their successful completion. The average number of participants per projects is 14 where the smallest number is 2, and the highest number is 53. Figure shows the distribution of the participants by regions. Besides the regions of Europe, institutes from Africa, Asia, North and South America, and Oceania (Australia) participated. However, partners from Europe contributed mostly to the fulfillment of the project objectives, i.e., 4,460 out of 4,535 (99%) participants. Institutions from Western Europe participated in 2107 cases, which is 47% of the total number of participants. Southern Europe has a share of 28%, while institutions from Northern and Eastern Europe are involved in 18% and 6% of the cases, respectively. In total, institutes from 11 European countries participated 100 or more times: Norway (100), Austria (121), Sweden (201), Greece (247), the Netherlands (387), United Kingdom (408), Belgium (426), Spain (426), France (457), Italy (466), and Germany (675). These 11 countries account for 87% of all participants in the Transport and Mobility domain. In terms of those participants outside of Europe, Asia contributed and supported the projects the most with a total of 47 participants, while North America, Africa, South America, and Oceania contributed 14, 7, 6, and 1 times, respectively.

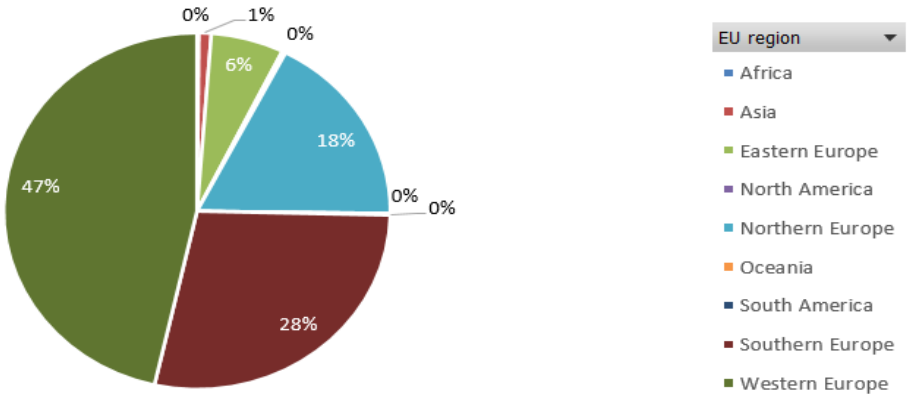

Figure 3. The participants' geo-distribution

As illustrated in Figure 2, more than half of the projects (i.e., 171 out of 310) in the Transport and Mobility domain received funding in the range of 1-5 million Euros, while 101 projects received funding between 5-10 million Euros, 15 projects received funding in a range between 10-15 million Euros, while merely 12 received budget between 15-20 million Euros. The average value of the allocated grant for a project is 5,458,781 Euros where the scale of the project costs ranges from 499,937 Euros to 29,546,161 Euros. 5 projects spent over 20 million Euros, while 20 projects received a less than 1 million Euro budget.

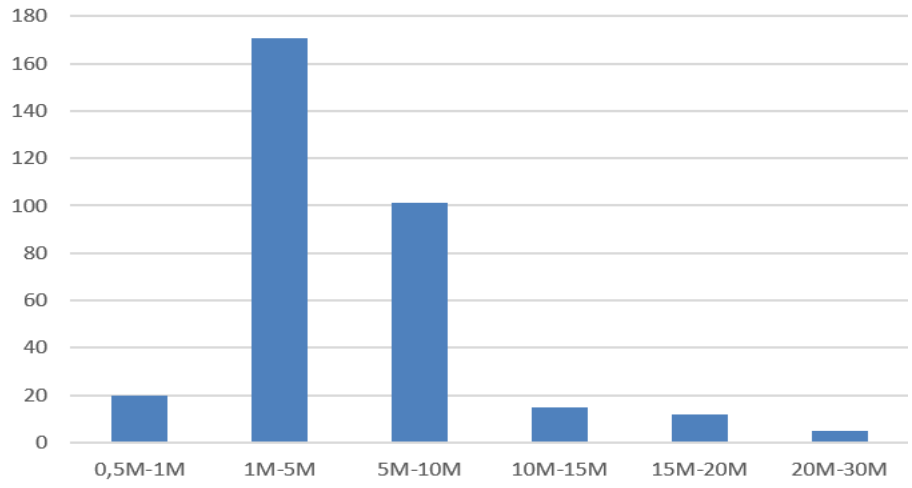

Figure 4. Funding distribution among the projects

4.2. Topic modelling results

After the creation of the corpus, where 310 project abstracts are collected in textual format, the LDA method is applied by using the MALLET software. Despite the usability of the topic identification, every time when the program is run, the words of the topics are changed moderately, which can be connected to the fact that the LDA cannot identify the exact number of topics. To reduce the inaccuracy in the topic identification and to increase robustness, the program is run 10 times with different word counts (i.e., 2, 3, 5, 7, and 10) in one topic. The program with a pre-identified set of commands is initiated in a total of 50 times. As it has been stated before, the LDA cannot determine the exact number of topics hidden in the corpus. Thus, 10 topics with different word counts are to be modelled because the analysis shows that the value for this corpus is sufficient. As a result, based on the chosen approach, in the final analysis, 5 main topics are identified as they frequently appear in the results for each word count. The final topic names are created following the logics of the set of words that identify the topic.

Using different word counts when running the program gives such results where the set of words of a particular topic are different in nature. Thus, when starting a process with two words in the topic, in most cases, these two words are nouns, such as logistics + port, emissions + aviation, and safety + roads. When running the program with three words, in most cases, the set of the words consist of nouns, as well. However, in rare cases, adjectives and verbs appear. Starting with a word count of 5, adjectives and verbs occur moderately. In case of 7 words, the words in a word set are more varied, namely deverbal nouns and deverbal adjectives (e.g., manufacturing and planning) appear. Running the LDA with 10 words gives results with more adjectives and verbs where adverbs rarely appear in a set of words that in some cases could give more information to identify a specific topic. However, the level of noise with higher word counts increases thus making it difficult to identify the topics. Running the LDA with varied word counts help to decrease the noise by identifying which topics are modelled more frequently. As a result of the runs with 2,3,5,7, and 10-word counts, the 5 topics identified with their most frequent words are the followings:

- **Road and traffic safety:** safety, road, systems, vehicles, traffic, users, driver, data.
- **Aviation and aircraft**: emissions, aviation, aircraft, air, fuel, noise, propulsion, effects.
- **Mobility and urban transport**: mobility, urban, cities, sustainable, policy, solutions, city, planning.
- **Maritime industry and shipping**: maritime, ship, port, vessel, shipping, concept, inland.
- **Open and real-time data**: data, information, services, travel, open, time, real-time, operators.

Table 1 shows the frequency of the topics for each word count when running MALLET. The identified topics are met in every word count run at least with 50%. This means that if the value in the table is five, out of 10 topic modelling iterations, it is met 5 times out of 10.

The results show that road and traffic safety is the biggest topic in the Transport and Mobility domain of the Horizon 2020 framework because this topic reaches 50 times out of 50. The topic connected to urban mobility development is created 39 times, around 70% (35 times) rate is achieved by the topics connected to aviation and aircraft, as well as to maritime industry and shipping (36 times), while the lowest frequency is shown by the topic connected to open and real-time data in transport (29 times).

Table 1. The frequency of topics per word count

Topic name	**Word count**					**Total**
	2 words	3 words	5 words	7 words	10 words	
Road and traffic safety	10	10	10	10	10	50 times
Aviation and aircraft	9	7	6	6	7	35 times
Mobility and urban transport	5	7	9	8	10	39 times
Maritime industry and shipping	6	7	7	6	10	36 times
Open and real-time data	6	6	5	5	7	29 times

Besides that, the LDA calculates not merely the frequency of the words and the frequency of their occurrence with other words but the distribution of these topics through one text document in the corpus, which means that one text document (project abstract) can have several modelled topics with different percentage of share. The share of the projects in the created topics shows that 138 projects for the topics of road and traffic safety, maritime industry and shipping, mobility and urban transport, aviation and aircraft,

and open and real-time data score more than 50 percent in the distribution of topics in one text document (Table 2). Therefore, 138 projects out of 310 have a pattern of belonging to more than one topic. A total of 262 projects out of 310 score more than 30 percent related to one modelled topic.

Table 2. Amount of projects scored with a high share in one text document

Modelled topic	2014-2015	2016-2017	2018-2020	Total per topic
Road and traffic safety	10	8	10	28
Maritime industry and shipping	6	4	16	26
Mobility and urban transport	9	11	13	33
Aviation and aircraft	14	17	12	43
Open and real-time data	8	0	0	8
Total per year	47	40	51	**138**

To mention a few projects, PANACEA (87%), ENGIMMONIA (72%), FastTrack (80%), SHEFAE 2 (89%), and ETC (64%) are the clearest representatives of each modelled topic (Table 3). The PANACEA project focuses on the development of the driving assessment system for truck and taxi drivers, courier service riders, and other commercial drivers to minimize the accidents on roads. The ENGIMMONIA project is directed to the search of alternative carbon-free fuels for vessel to prevent the emission of greenhouse gases and minimize the impact of the maritime industry on climate change. The FastTrack project aims to develop and deploy sustainable mobility innovations for urban transport thus improving the socio-economic conditions and leading to sustainable growth. The SHEFAE project focuses on the development of heat exchangers in aero turbo fan engines to decrease the amount of burnt fuels and consequently decrease the emissions of aircraft. Finally, the ETC project targets travellers' commuting by providing a traveller-centric platform where they can choose optimized travel routes by using dynamic and static data. Moreover, the data for transport suppliers are open to increase the effectiveness and utilization of the transport infrastructure. The clearest representative can be considered as evidence that the LDA approach works efficiently since the high share of one topic in the text document proves that the identified topic is correct and can be traced throughout the corpus with different percentage shares.

Table 3. Projects as the representative of a modelled topic

Modelled topic	Topic representative	Share
Road and traffic safety	PANACEA	87%
Maritime industry and shipping	ENGIMMONIA	72%
Aviation and aircraft	SHEFAE	89%
Mobility and urban transport	FastTrack	80%
Open and real-time data	ETC	64%

5. Discussion

In current study, the topic modelling methodology using the LDA technique is applied because it provides the ability to identify topics in a large database of text documents. Having collected abstracts of 310 projects, a database of text documents is obtained. As a result, the following five main topics are identified: road and traffic safety, aviation and aircraft, mobility and urban transport, maritime industry and shipping, and open and real-time data in transport.

The literature review shows that limited analysis of big research frameworks like the Horizon 2020 with the application of text mining techniques has been realized before. Despite of utilizing text mining techniques for the analysis of databases with large amount of textual data, the LDA approach has not been applied to identify main trends in transport and to retrieve hidden topics. Moreover, the LDA has not been previously used for the analysis of the Horizon 2020 framework; however, the obtained results show the effectiveness of the method.

Besides the identification of main topics, some trends are retrieved, as well. The topic of aviation is connected to noise pollution and emission as these words are the most frequent along with aviation and

aircraft. It can mean that pollution from fuels and noise are the main problems for aviation analyzed in the Horizon 2020 projects. When running the LDA, a frequently encountered word is logistics that often appears with the words freight and supply chains. However, the topic of logistics is not identified since it represents not solely one direction in the transport industry, but it is part of a larger term in transport. Logistics are moderately encountered in a set of words related to the maritime industry but less often met with the topic of road transport and aviation. In addition, rail and railways are the most frequently identified words. Although they might appear often in the LDA process, their related topics are not identified. Moreover, such words as MaaS and c-ITS are frequently present in the LDA process, as well. Usually, these abbreviations appear in a set of words together, but no topic-related logic is found. Although these are frequently used words, the logic behind the process can be explained by the fact that those words appear very frequently but solely in a few projects. Thus, if the distribution of the frequency among all projects is not homogeneous, these words are not chosen as topics by the LDA method. Furthermore, the words noise, pollution, and emission are considered as frequently appearing words in the results mostly connected to aviation, moderately to maritime industry, and infrequently to mobility and urban transport. This means that probably these words represent a main problem for aviation and maritime industry, as well as it is relevant for mobility and urban transport, too.

Moreover, by looking into the distribution of topics through one textual document, a clear trend is traced when some text documents represent almost equivalent topic shares. Open and real-time data shares the same percentage as the topics of mobility and urban transport, and road and traffic safety. An explanation could be that in case of mobility and urban development as well as in road and traffic safety, open and real-time data play a key role in the improvement of various concepts in several projects.

The main limitation of the research is that it is not possible to know the exact number of hidden topics in the corpus. Therefore, the initial set of commands in the MALLET tool with 10 topics and 10 words per topic to obtain the results of every run changes moderately. To overcome this stochastic change, process repetitions are used to estimate the highest probability of the occurring topics. The process is run 50 times, 10 times per each word count, which is enough to determine the five topics occurring more than 50% of the process runs. Afterward, the LDA process is launched with the five topics where very similar results are obtained every time. The changes in the set of words in most cases are the words with the highest repetitions, such as rail, railways, noise, emission, logistics, and MaaS, which are discussed in the previous paragraph.

In the realm of topic modelling, the LDA method stands as a robust tool, yet it is crucial to mention that it has limitations regarding the comprehensive understanding. One notable constraint appears when dealing with short documents where the method may grapple with the scarcity of contextual information, which potentially compromises its efficacy. Even though abstracts are relatively short textual documents, the main information of projects is included in the abstract together with the keywords and objectives. The sensitivity to the selection of the number of topics represents another critical facet, which emphasizes the need for a specific approach to ensure meaningful results. Moreover, the LDA operates under the assumption that the documents inherently embody mixtures of topics (the distribution of topics through one text document). However, this assumption might not universally be true, which leads to potential discrepancies in the representation of real-world document structures. Additionally, the capacity of the LDA to capture intricate nuances, such as word order and semantic relationships, is limited. This constraint implies that the model may struggle in other scenarios where the precise arrangement of words or the semantic interplay between the terms hold paramount importance. The process of parameter tuning in the LDA, though essential for optimal performance, introduces another layer of complexity. Striking the right balance in parameter selection demands careful consideration because suboptimal choices can impact the effectiveness of the model. Acknowledging these limitations is pivotal for researchers and practitioners fostering a comprehensive approach to the application of the LDA in diverse contexts.

The contribution of this paper is in the approach to identify the main issues and key pathways of the transport sector in the Horizon 2020 framework by using the LDA topic modelling. Scientists can benefit from this paper by finding insights into new research investigations by taking the identified main trends and topics into consideration. For example, the increasing trend in the mobility and urban transport sector shows that specialists could focus more on the issues of air and noise pollution in the aviation and maritime industry. In theory, bigger budget should be allocated to the more trending topics, which means more opportunities and financial benefits. In the future, topic modelling can play a crucial role in identifying the opportunities for future developments based on past and new results. The extension of this research could be realized by the analysis of all research outputs published related to each project thus collecting a bigger corpus and analyzing more texts to identify more specific issues and problems.

6. Conclusion

The Horizon 2020 framework has been used in various publications to analyze data for different sectors and industries. Previous attempts aimed to identify the key research areas and to look for gaps to assist the research community. Limited analyses using qualitative techniques to analyze the projects is found in the field of transport. Current paper covers these gaps by providing a deeper analysis of the main topics, issues, and trends in the transport sector by using a topic modelling technique. The research considers 310 projects in the Transport and Mobility domain of the Horizon 2020 framework by applying the LDA technique of topic modelling to identify the main topics in the field. As a result, the following five topics are identified in the created text corpus retrieved from the projects: road and traffic safety, aviation and aircraft, mobility and urban transport, maritime industry and shipping, and open and real-time data in transport. Mobility and urban transport show that the interest in the realization of this topic is raising year-by-year, while open and real-time data is not a focus area nowadays as it was several years ago. High interest has been traced in the maritime industry and shipping topic in recent years. Aviation and aircraft projects primarily focus on the decrease of noise and air pollution. The results of the analysis provide the main topics, and the most frequent words in the text database help to obtain the main issues, key pathways, and trends in the transport industry.

Acknowledgements

Project no. TKP2021-NVA-02 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

References

1. Alghamdi, R. and Alfalqi, K. (2015) A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 7. DOI:10.14569/IJACSA.2015.060121.
2. Bai, X., Zhang, X., Li, K. X., Zhou, Y. and Yuen, K. F. (2021) Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, 102, 11–24. DOI:10.1016/j.tranpol.2020.12.013.
3. Dang, S. and Ahmad, P. H. (2014) Text mining: Techniques and its application. *IJETI International Journal of Engineering & Technology Innovations*, 1(4), 22–25. ISSN: 2348-0866.
4. European Commission. (2024) CORDIS EU research results. https://cordis.europa.eu/search?q=contenttype%3D%27project%27%20AND%20frameworkProgramme%3D%27H2020%27%20AND%20applicationDomain%2Fcode%3D%27trans%27, Accessed 15.03.2024.
5. Giarelis, N. and Karacapilidis, N. (2021) Understanding Horizon 2020 data: A knowledge graph-based approach. *Applied Sciences (Switzerland)*, 11(23). DOI:10.3390/app112311425.
6. Gopalakrishnan, K. and Khaitan, S. K. (2017) Text mining transportation research grant big data: Knowledge extraction and predictive modeling using fast neural nets. *International Journal for Traffic and Transport Engineering*, 7(3). DOI:10.7708/ijtte.2017.7(3).06.
7. Kherwa, P. and Bansal, P. (2020) Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), 1–16. DOI:10.4108/eai.13-7-2018.159623.
8. Liu, C. and Yang, S. (2022) Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Systems with Applications*, 207, 117991. DOI:10.1016/j.eswa.2022.117991.
9. Liu, Y. and Cheng, T. (2020) Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science*, 16(1), 76–103. DOI:10.1080/23249935.2018.1493549.
10. Maghrebi, M., Abbasi, A., Rashidi, T. H. and Waller, S. T. (2015) Complementing travel diary surveys with Twitter data: Application of text mining techniques on activity location, type and time. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 208–213. DOI:10.1109/ITSC.2015.43.
11. Malloci, F. M., Penadés, L. P., Boratto, L. and Fenu, G. (2020) A text mining approach to extract and rank innovation insights from research projects. *International Conference on Web Information Systems Engineering*, 143-154. DOI:10.1007/978-3-030-62008-0_10.
12. Marzi, E., Morini, M. and Gambarotta, A. (2022) Analysis of the status of research and innovation actions on electrofuels under Horizon 2020. *Energies*, 15(2). DOI:10.3390/en15020618.

13. Rachman, F. F., Nooraeni, R. and Yuliana, L. (2021) Public opinion of transportation integrated (Jak Lingko), in DKI Jakarta, Indonesia. *Procedia Computer Science*, 179, 696–703. DOI:10.1016/j.procs.2021.01.057.
14. Radoselovics, A. (2019) *HORIZON 2020 analysis.* https://openaccess.uoc.edu/bitstream/10609/99906/6/aradoselovicsTFM0619memory.pdf, Accessed 15.03.2024.
15. Serna, A. and Gasparovic, S. (2018) Transport analysis approach based on big data and text mining analysis from social media. *Transportation Research Procedia*, 33, 291–298. DOI:10.1016/j.trpro.2018.10.105.
16. Suh, Y. (2021) Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database. *Safety Science*, 142, 105363. DOI:10.1016/j.ssci.2021.105363.
17. Sun, Y. and Kirtonia, S. (2020) Identifying regional characteristics of transportation research with transport research international documentation (TRID) data. *Transportation Research Part A: Policy and Practice*, 137, 111–130. DOI:10.1016/j.tra.2020.05.005.
18. Suran, S., Pattanaik, V., Yahia, S. Ben and Draheim, D. (2019) Exploratory analysis of collective intelligence projects developed within the EU-Horizon 2020 framework. *Lecture Notes in Computer Science, LNAI*, 285–296. DOI:10.1007/978-3-030-28374-2_25.
19. Uys, J. W., Du Preez, N. D. and Uys, E. W. (2008) Leveraging unstructured information using topic modelling. *PICMET: Portland International Center for Management of Engineering and Technology, Proceedings*, 955–961. DOI:10.1109/PICMET.2008.4599703.