24 January 2024

Three decades of transport corridor research: a literature review applying text mining

Houssain Baalla
Student number: 2637138

Utrecht University
Applied Data Science

Project supervisor: ***P. A. Witte***
Second examiner: ***S. M. Labib***
Type of document: ***master's thesis***

Contents

1. Introduction

How can the use of text mining techniques in a literature review enrich the transportation corridor debate? This study aims to answer this specific question. Before delving into this topic, the question arises why there is a need to enrich the corridor debate. The reason is because the academic debate exhibits a lack of cohesion and linkage. There is a lack of connection and integration which makes it difficult for readers interested in transportation corridors to come to an understanding of the main topics and focus points in the corridor debate. However, modern day text mining techniques applied in a corridor literature review could play a role in creating a clearer view for readers interested in transportation corridors. Before explaining why text mining techniques can make a literature review more effective and efficient, it is good to understand why the current academic debate is failing to grant readers a holistic and clear view of the topic.

Background

Starting with a definition of transport corridors: Priemus & Zonneveld (2003) state that "we can imagine corridors to be bundles of infrastructure that link two or more urban areas. These can be highways, rail links, separate bus lanes, cycle paths, canals, short-sea connections and air connections. In general, corridor development concerns connections that use different transport modes (car, train, tram, ship, aeroplane), and carry both passenger and freight transport. Furthermore, one can also adopt a broader interpretation of corridors that encompasses things like ICT-infrastructure, power lines and cables as well as pipes for drinking water, natural gas, crude oil, electricity and sewage". Upon encountering such a definition, the dynamic and multidimensional nature of the academic debate comes as no surprise, for a transport corridor is, in and of itself, inherently complex.

Corridors might consist of different transport modes such as car, train, tram, ship, aeroplane and even a combination of those, which leads to academics focusing on different kinds of corridors in the sense of transport modes. Another aspect which causes corridors to be complex is the diverse scales of different transport corridors, some being relatively small and local/regional while others are larger and (inter)national or even global in terms of geographical scope. Furthermore, the stakeholders of transport corridors are also diverse, ranging from the public sector to the private sector, as well as non-profit organizations and citizens (Öberg et al., 2016), resulting in different interests and priorities concerning transport corridors. This diversity translates to a similar diverse pursuit of interests by academics, at times focusing on policies and regulations while other times focusing on economic benefits, sustainability, infrastructure, or bottlenecks and chokepoints of corridors.

These are a few examples of why the academic debate makes it difficult for readers to gain a clear view of the topic. In order to facilitate readers to this end, this study aims to provide a literature review that connects the various topics and themes present in the corridor debate, by using text mining techniques.

Text mining

Text mining techniques can offer various benefits to a transport corridor literature review, some of them increasing the efficiency and others the effectiveness of the review, depending on the specific techniques that are employed. Due to the short period span in which this study was made, it was necessary to focus on only a select few text mining techniques, namely: keyword extraction, topic modeling and sentiment analysis. The use of each of these techniques will be briefly explained, as well as the beneficial effect they have on this literature review.

- **Keyword extraction**
  Text mining allows for frequent terms and keywords to be extracted from documents and examined. This aids in determining important words that appear frequently in the corridor literature, giving the reader an understanding of the main ideas and subjects that researchers have concentrated on (Gupta, 2017). Displaying these keywords and most frequent terms into aesthetic visualizations such as word clouds enhances the accessibility and interpretability of the corridor literature review. This allows readers to process complex information more naturally.

  Furthermore, employing this technique is less arduous than extracting keywords from documents 'manually'. Not only is it more time efficient to do this using text mining, but it is also easier to scale this particular analysis in the case that more academic reports are added to the dataset. Another reason why this text mining technique benefits a literature review, is the fact that it decreases the chance of human error that would occur when doing this analysis manually, making the results more accurate and increasing the quality of the literature review.

- **Topic modeling**
  Topic modeling is a flexible technique that may be used to extract valuable insights, find hidden patterns, and make sense of large amounts of textual data. The corridor literature's major themes and topics can be identified using topic modeling. The words in each topic could also be regarded as keywords, so topic modeling has good synergy with the previous analysis by displaying the relationship between keywords and putting them in context of their respective topics. It facilitates the organization and classification of corridor information, giving readers a deeper understanding of the structure of the literature review. Furthermore, topics and emerging trends in the literature that might not be immediately evident in a manual examination might be found using topic modeling. Topic modeling algorithms examine the word distribution in documents and automatically detect latent topics (Blei, 2011). Emerging themes and trends that may be obscure or develop over time can be discovered through this technique. These subtle trends could potentially be missed if the literature review was done manually.

- **Sentiment analysis**
  Sentiment analysis can be used to determine the overall sentiment conveyed in the transport corridor literature (Taboada, 2016). By offering a complementary viewpoint, the sentiment analysis adds depth and nuance to the literature review. Moreover, a sentiment analysis makes it possible to quickly and efficiently filter for papers containing sentiments of interest. With this text mining technique, it is also possible to determine whether corridors in various geographic locations are perceived with differing overall sentiments. The geographical aspects of transport corridors are one of its most important focus points, and it might be that certain regions take a more prominent place in the corridor debate. It will be interesting to know whether those regions also have a higher overall positive sentiment, and on the contrary whether regions that are less frequently written about have a higher negative sentiment. This adds an additional layer of depth to the literature review and provides readers with a more holistic view of the topic at hand.

A literature review, incorporating these text mining techniques, will give readers interested in transport corridors a comprehensive grasp of the subject. Using these techniques will also display their uses in general and more specifically; their uses in conducting literature reviews. Most importantly, it will hopefully serve to reduce the lack of connection and integration currently present in the corridor debate.

2. Review Methodology

This section will delve into the methodology part of the transport corridor literature review. First the data collection will be described, followed by the data preprocessing steps which were taken. This section will conclude with the description of the methodology of the data analysis techniques that were used in this study.

2.1 Data collection

The corridor reports, which where the subject of this review, were collected by Patrick Witte. The review methodology consisted of defining the scope of the review. This was set to be three decades, ranging from 1990 to the year 2021. Afterwards, a systematic search of scientific literature was performed to collect corridor reports over those three decades using Google Scholar's search engine. This resulted in a gross list of 137 academic, peer-reviewed papers published in international journals. After filtering the 137 papers and selecting only those papers that took corridors or corridor development as the main focus of research, the gross list was reduced to a selection of 79 papers for the full review procedure.

Additionally, Witte made a systematic inventory of the content of the 79 selected papers in an Excel file. The file contained information such as: year of publication, title, authors, geographical focus etc. The abstract parts of each of the 79 papers were also included in the Excel file later on. For this literature review, both the 79 papers as well as the Excel file were used to conduct the research. Due to the short timespan in which this study was conducted, it was determined that the data of the 79 papers would suffice, and no additional papers were added to this dataset.

2.2 Data preprocessing

The analyses in this study were performed using the programming language R. Within R, various packages and libraries were employed, the most notable of those packages being the *Quanteda* package which was used for the text mining analyses.

Tokenization

A part of the preprocessing phase of the analysis was the tokenization of the abstract parts of the 79 corridor papers. Tokenization is a foundational step which enables the transformation of unstructured text into structured units that can be analyzed. This facilitates a wide array of text analysis techniques and improves the efficiency and effectiveness of the text mining process. The main use of tokenization is identifying meaningful words. However, tokenization by itself is not enough to achieve this, as there are still many words which recur frequently but are not relevant for many text analysis purposes, an example of this is stop words such as 'and', 'are', 'this' etc. (Gurusamy & Kannan, 2014). These stop words, and other unnecessary aspects of the textual data (such as punctuations and symbols), are to be removed so that the tokens facilitate further meaningful analyses.

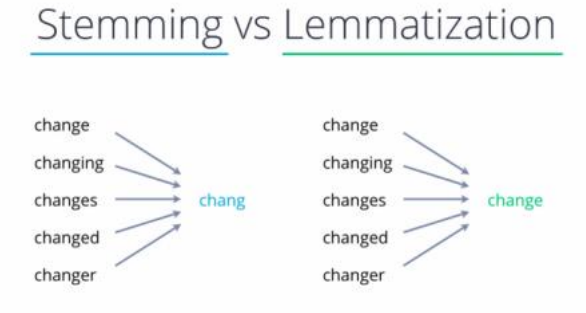
Lemmatization

Another preprocessing technique used in this study is lemmatization. Lemmatization is the reduction of words to their base form, minimizing variations. This improves consistency in analysis by standardizing the text and ensuring that various word derivations or inflections are treated equally. It furthermore provides a clearer understanding of the corridor papers by converting words to their core meaning. This helps determine the essence of the text, making later analyses more accurate and insightful (Hickman et al., 2020). Furthermore, the transformation of words to their base forms helps in reducing the dimensionality of the feature space. This reduction of feature space can improve model efficiency and performance (Kumar, 2012). Which is relevant in this study because the

lemmatized words served as input for the clustering technique topic modeling, which will be explained in the *data analysis* section of the methodology.

A similar preprocessing technique which could have been used in this study is stemming. The purpose of stemming is also to reduce words to their base or root form. The way to achieve this using stemming is often to remove the derivational affixes of words. However, stemming of words may result in an output that has no semantic meaning, while lemmatization tends to preserve the linguistic understanding of words. An example of this is illustrated in ***figure 1: Word stemming and example of reduction in lemmatization***. The benefit of using stemming lies in its undemanding application. Nonetheless, lemmatization was implemented in this study due to its more accurate results (Khyani & Siddhartha, 2021).

Figure 1: Word stemming and example of reduction in lemmatization



(Source: Khyani & Siddhartha, 2021)

2.3 Data analysis

Keywords extraction

Keyword extraction involves the identification of the lexical units that best represent the document (Firoozeh et al., 2020). In this study, they are the words that represent the main topics and concepts in the 79 corridor papers. A possible way to get an indication of these keywords is by looking at the most frequently used words in the papers after filtering out the stop words and other words that have no relevant meaning in this analytical context. For the purpose of making the word cloud, the words 'corridor', 'transport', and 'development' were also removed. The exclusion of these words is justified by the anticipation that they occur frequently in the corridor papers, thereby contributing limited additional value to the present analysis. The most frequent words were put in a word cloud, with a maximum limit of 100 words to ensure readability. The input of the word cloud was a document feature matrix containing the preprocessed data of the abstract parts of the 79 corridor papers.

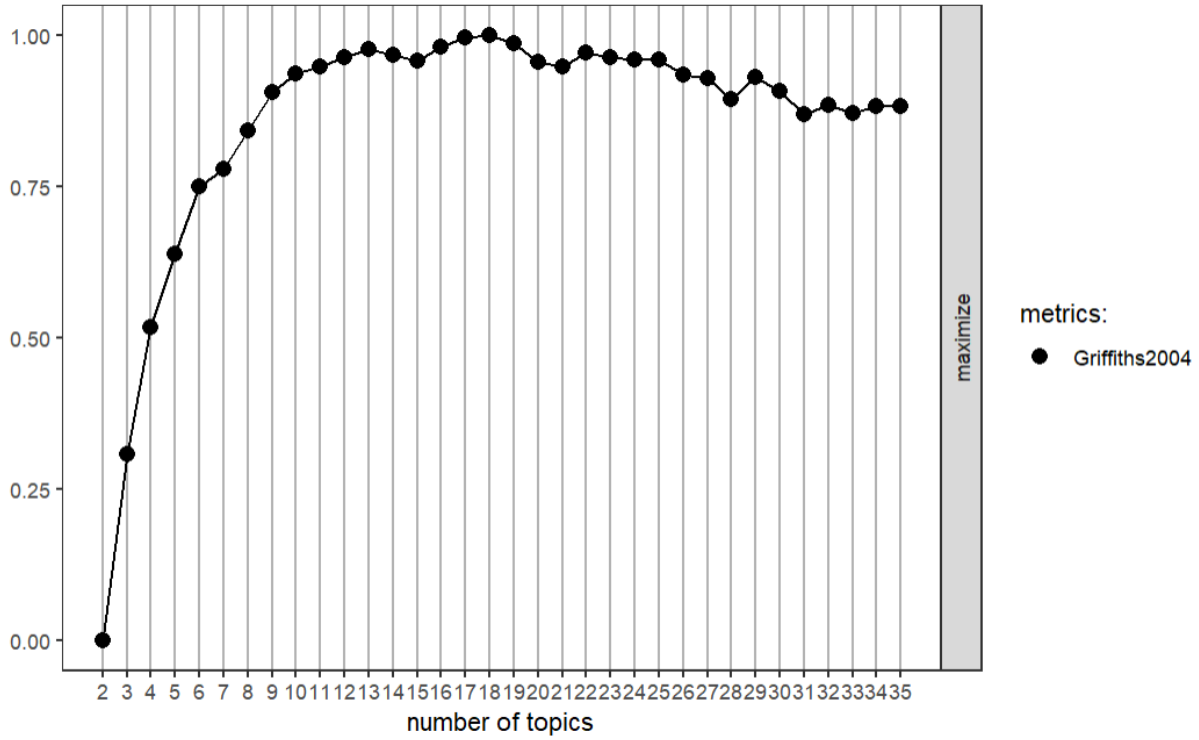
It is also possible to determine keywords by calculating the Term Frequency-Inverse Document Frequency (TF-IDF) scores (Abid et al., 2023). The IDF part of this technique helps to give more weight to relatively rare terms, by making infrequent terms have a higher impact. So, words that appear many times in one of the corridor papers, but do not appear many times in the other papers, are still treated as relevant words in the scoring process of the TF-IDF. The keywords can then be determined by looking at the words with the highest TF-IDF scores. In this study, the TF-IDF scores were calculated for the words in the abstract parts of the 79 papers, and the top fifteen scores were extracted to display relevant keywords.

Topic modeling

There are different techniques that can be used for topic modeling. A popular technique, which is used in this study, is the Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model which is relatively simple and yields interpretable results. Using LDA, topics are assigned to the words of the documents (Blei, 2011). This way, the words in the abstract parts of the corridor documents can be categorized to certain topics using the LDA topic modeling technique. The LDA uses the Dirichlet distribution to find themes for each document model and words for each topic model.

One of the assumptions of LDA however is that the number of topics is known. Therefore, it is necessary to provide the model with a k number of topics. There are different ways to find an appropriate number of topics. In this analysis, Gibbs-sampling was used to find an appropriate k (Griffiths & Steyvers, 2004). Using Griffiths' and Steyvers' method on the cleaned data, an indication of a good k is a number above 9 topics. This can be read out of the plot in ***figure 2: Griffiths maximization plot***. In this method, the aim is to maximize the metric seen in the plot. The metric improves significantly when increasing the topics from 2 to 9, after that it begins to saturate.

Figure 2: Griffiths maximization plot



(*Source: Authors' own*)

Upon analyzing the maximization plot, it was determined that employing 11 topics yielded optimal results for the topic modeling technique. Therefore, this number of topics was specified in the model parameters to build the LDA model using the *topicmodels* package in R. The LDA model was trained using 500 iterations after an initial 100 burn-in iterations. The input data for the model consisted of a document feature matrix containing the preprocessed data of the abstract parts from the 79 corridor papers.

Sentiment analysis

The sentiment analysis was performed using the Lexicoder Sentiment Dictionary (LSD). Instead of using the abstract parts of each document, the complete papers were used. This yielded more accurate results, as the abstract parts alone were not sufficient to determine the overall sentiments. One paper, however, could not be used in the sentiment analysis. The paper, "*Major European Transport Corridors and their role in the Republic of Moldova development*", written by Vladislav Machidon (2015) contained an encryption security which makes it impossible to copy the text or process it with data analytics. The rest of the 78 papers were suitable for a sentiment analysis. The sentiment analysis was performed on the tokenized papers, after the tokens were preprocessed removing punctuations, symbols, numbers and other non-relevant parts of the papers. The tokens were furthermore converted to lowercase letters and lemmatized. One change in the lemmatization list was made, transforming the lemma of the word '*number*' to '*number*' (previously '*numb*'). This yielded more accurate results. The LSD dictionary was used to match the preprocessed tokens with negative or positive sentiments, neutral words were not taken into consideration.

The Lexicoder Sentiment Dictionary also filtered out patterns wherein positive words are preceded by a negation (such as 'not good') as well as negative words that are preceded by a negation (such as 'not terrible'). This allowed for an analysis on the sentiments per paper, as well as an examination of the type of positive/negative words used in corridor literature. In order to quantify the overall sentiment of each of the 78 papers, the number of positive words was put in proportion relative to the amount of negative words. To determine the overall sentiment of the corridor debate, the mean of all the positive proportions and the mean of all the negative proportions were calculated and displayed in a pie chart.

Similarly, the overall sentiments of corridor papers which focused on specific regions were determined. For each region, a mean proportion of positive and negative words was calculated. This was also done for the conceptual papers (i.e. corridor papers that did not focus on a specific geographical region). This allowed for a comparison of overall sentiments between the different regions. The examination of geographic regions also included the number of authors who had written about corridors in specific regions. By splitting the authors in the data frame into separate columns, it was possible to determine the number of individual authors who have focused on certain regions.

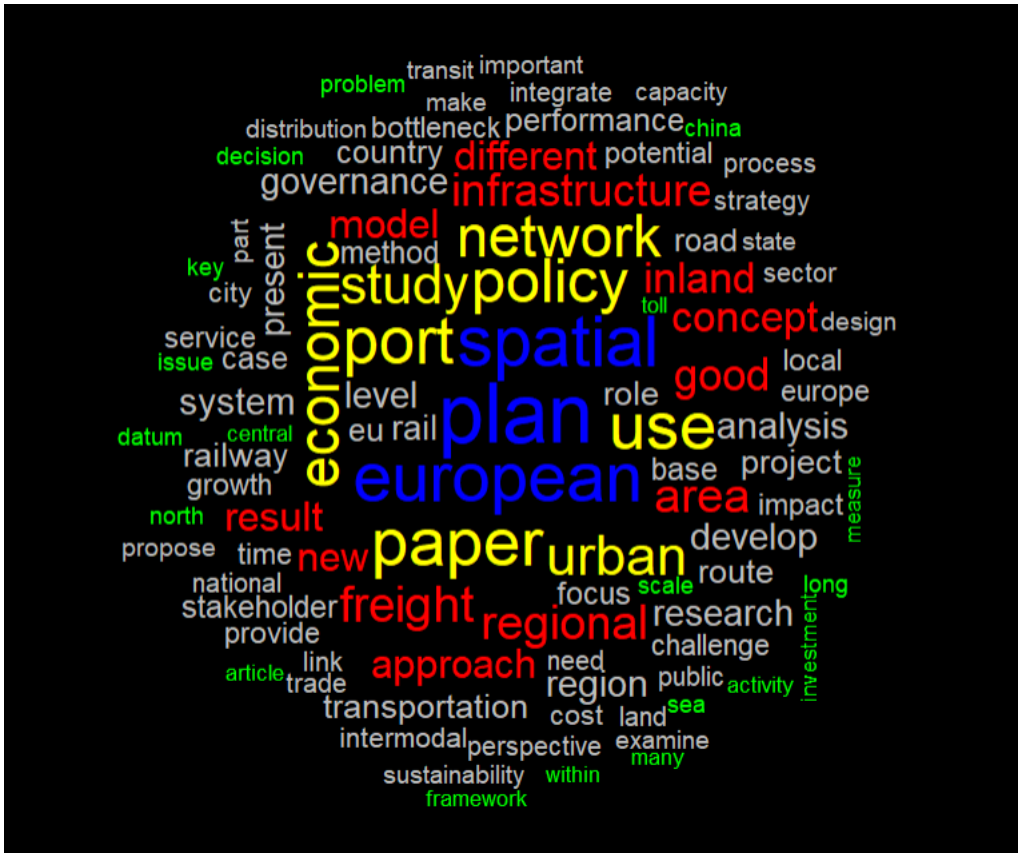
3. Results

This section of the paper will delve into the results of the analyses. First, the results of the keyword extraction will be displayed and explained, followed by the outcomes of the topic modeling analysis. Afterwards, this section will be concluded by presenting the findings of the sentiment analysis.

3.1 Keyword extraction

Looking at the most frequent words in the corridor literature gives a good indication of common themes and concepts discussed in the corridor literature. ***Figure 3: Word cloud*** displays the most frequent words, whereby the size and color of the words indicate the frequency of occurrence. A quick glance at these words will shed some light on the primary concerns, sentiments, and recurrent themes within the discourse.

Figure 3: Word cloud



(Source: Authors' own)

The words '*plan*', '*European*', and '*spatial*' are the most occurring words in the cleaned list of tokens. The words '*plan*' and '*spatial*' are within the scope of expectations and are most likely related to the inherent nature of transport corridors. Corridors are geographical points spanning through multiple cities and countries, which explains the word '*spatial*' in the word cloud. As for the word '*plan*', due to the size and complexity of corridors as well as the strategical benefits and practical usage, a fair amount of planning needs to be done beforehand to determine the optimal course to follow.

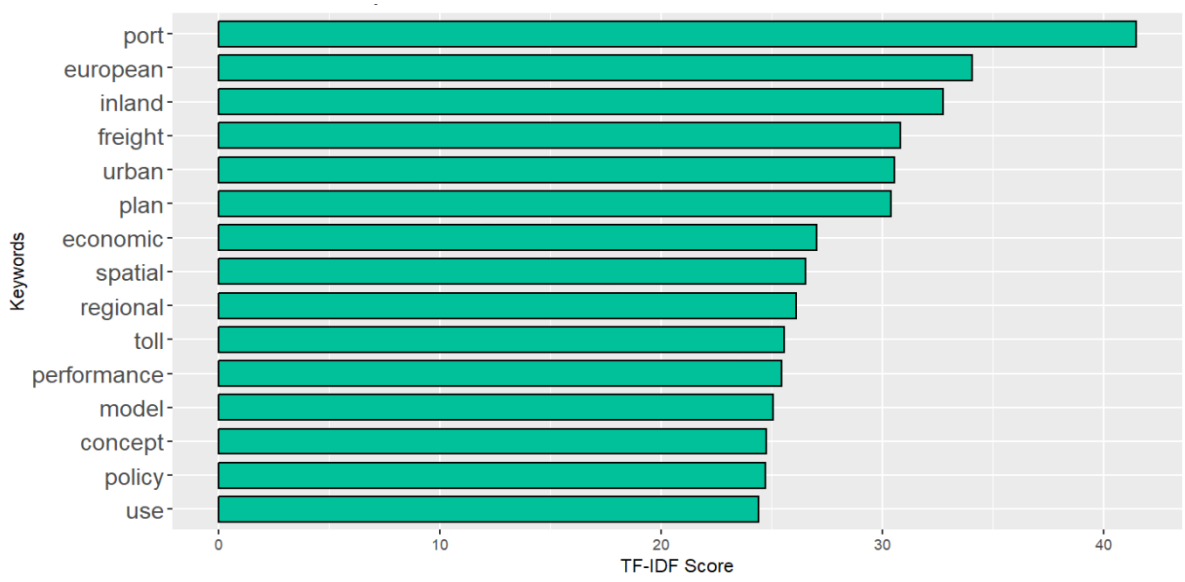
The third most frequent word in the word cloud, '*European*', was an unexpected result, outside the scope of the initial expectations. A possible explanation for this might be that Europe is the more

prominent geographical focus of transport corridors in the academic debate. *Chapter 3.3* of this report will delve into this subject to determine the validity of this assumption.

The various words displayed in the word cloud reflect the diverse and broad nature of the discourse. Words like '*economic'* indicate the monetary aspect of transport corridors, while words like '*bottleneck*' and '*analysis*' denote the complications and challenges associated with evaluating and assessing the efficiency and functionality of these corridors. Other words like '*freight'*, '*urban'*, '*time'*, '*performance'*, and '*policy'* further display the variety of topics and themes present in the corridor debate.

To further deepen the understanding of important terms used in the corridor debate, keywords were identified through the calculation of TF-IDF scores. The top fifteen keywords are displayed in ***figure 4: TF-IDF Scores of keywords***.

Figure 4: TF-IDF Scores of keywords



(*Source: Authors' own*)

The keywords show a variety of themes, similar to the words in the word cloud. Keywords like '*inland'* and '*port'* highlight the importance of hinterland connectivity and maritime gateways in corridor studies (Notteboom & Rodrigue, 2005), whereas '*regional'* and '*European'* emphasize the collaborative and geographic aspects of transport corridors (Williams & McGreal, 2002).

The keyword '*freight'* reflects the pivotal role of efficient transportation of goods within corridors (Hesse & Rodrigue, 2004). Keywords like '*urban'* and '*spatial'* highlight the influence of corridor design on urban landscapes and spatial patterns, illustrating the interaction between transportation corridors and urban areas (Rodrigue, 2004). '*Plan*' and '*economic*' underscore the strategic aspect of corridor development and the necessity of carefully considered plans to potentially spur economic growth (Ahmed, 2018).

As for the keyword '*toll'*, it emphasizes possible streams of income for governments. Moreover, it is associated with sustainability while at the same time reflecting the financial considerations linked to corridor usage (De Borger, 2006).

The keywords 'p*erformance*' and '*model*' explore the analytical facets of corridor assessment, emphasizing the significance of determining how effective transportation networks are. Closely related to these keywords is the overarching '*concept*', which represents the wide-ranging and dynamic character of transport corridor notions, encompassing a variety of approaches and methods (Priemus & Zonneveld, 2003). Policies also play a crucial role, as evidenced by the keywords '*policy*' and '*use*', which highlight the regulating and functional components of corridors. In order to shape the success of transportation corridors and guarantee that they are in line with more general economic and environmental objectives, competent policy formulation and execution are crucial (Jain & Jehling, 2020).

The word cloud and keywords plot provide a concise, yet informative, overview of the main ideas and important words pertaining to the corridor debate. This examination seeks to deepen the comprehension of the fundamental concepts affecting the discourse and establishes the framework for a more thorough analysis of the underlying dynamics. Furthermore, from an analysis of the keywords it can be determined that some words are closely related to each other, belonging to the same overall topic. The next part of this paper will therefore delve into the topic modeling analysis, in order to further analyze the exact relations between the words and provide a clearer overview of the themes in the corridor literature.

3.2 LDA topic modeling

To use the LDA topic modeling technique, a number of topics must be determined beforehand. Using Griffiths' & Steyvers' method (2004), a number of 11 topics were chosen as input for the model. The results of the LDA topic modeling technique are displayed in ***table 1: LDA Topics***.

Table 1: LDA Topics

1. Strategy	2. Difficulties	3. National aspects	4. Corridor development	5. Governmental need	6. Urban design	7. Corridor methods	8. Logistic	9. Operational application	10. Conceptual	11. Economic value
plan	present	country	development	transport	urban	port	transport	road	corridor	economic
spatial	result	route	new	european	can	paper	freight	cost	concept	growth
regional	bottleneck	use	transportation	network	city	inland	good	one	study	policy
area	perspective	time	research	policy	link	model	railway	system	performance	activity
level	challenge	also	sector	governance	local	different	rail	two	develop	economy
project	problem	trade	integrate	stakeholder	transit	intermodal	use	assess	europe	corridor
infrastructure	framework	china	land	eu	design	approach	service	analysis	aim	region

(*Source: Authors' own*)

The table displays the 11 topics that resulted from the LDA model, alongside their corresponding words. The topics are all different, but certain topics share some commonalities with each other. The topics that are more related to each other will be grouped together and further explained below.

- **Group 1**: (*1. Strategy – 5. Governmental need – 11. Economic value*)
  These topics are grouped together because they are related to the 'upper hierarchy' (government, EU, etc.) and the plans, goals and policies that they establish. The governments are concerned with the long-term value of corridors as well as the strategical impacts they make. And the economic advantages of corridors can be considered as a strategic goal.

  1. Strategy:
  The first topic – *Strategy* – has a focus on the strategical aspects and higher-level implementation of corridors. The important aspects in this topic are the long term

goals of corridors and the groundwork that needs to be present and formulated beforehand in order to achieve this. An example of strategic guidelines in the context of transport corridors is to create a polycentric/multi-centered urban system that will increase collaboration between urban and rural areas and encourage integrated transportation and communications concepts (Williams et al., 2002).

5. Governmental need:
The fifth topic – *Governmental need* – highlights the importance and need of corridor governance. Governance of corridors is not only limited to the public level, rather, the private level and non-profit organizations are also needed when it comes to governing transport corridors. After all, to achieve a well-functioning transportation system, a wide range of stakeholders must contribute (Öberg et al., 2016). Bringing these different stakeholders together is not an easy task in and of itself. A (new) tool that can facilitate this is Corridor forums. Corridor forums aim to bring key stakeholders together for consultative purposes.

11. Economic Value:
The last topic of this group – *Economic value* – delves into the economic interests of corridor applications. Economic value can be approached from different perspectives. One study, for example, focused on the impact of a transport corridor on the regional employment growth. The result of that study was that there was no significant relationship between the two (Bruinsma et al., 1997). Although it is not yet proven that corridors bring an economic advantage to the countries/regions that are part of the corridor network, it is nonetheless a topic of debate amongst researchers of corridor development.

- **Group 2**: (*3. National aspects – 6. Urban design*)
In this group, the topics are focused on the geographical aspects of corridors. The relevance of the spatial aspects of regions, countries and cities comes to light in these two topics. Corridors are after all connections between geographical key points and that is the focal point of these two topics.

  3. National aspects:
  The third topic – *National aspects* – focuses on the national level of corridors. This topic entails the geographical and national components of corridors, and the pertinent matters that ensue on a national level as well as the transnational aspects of corridors. Since corridors differ in size and scope, it is natural that the academic debate differentiates in this regard as well, with some papers focusing on relatively smaller corridors and others on the larger corridors. An example of corridors that relate to this particular topic are the relatively larger Rhine-Alpine corridor and the Chinese High-Speed Rail corridors.

  6. Urban design:
  The sixth topic – *Urban design* – is similar to the previous topic but instead focuses on the relatively smaller corridors that span through cities yet do not exceed national borders. An example of such a corridor is the Dutch A1 corridor. These relatively smaller corridors are less complex in the sense that they lack the

transnational aspects and vast sizes. This also reduces the number of stakeholders involved and makes the governance of these corridors relatively easier.

- **Group 3**: (*8. Logistic – 9. Operational application*)
  This pair of topics focuses on the practical and concrete aspects of corridors. What is highlighted in these two topics are the operational – 'in the field' – implications that exist within the functioning of corridors. The conceptual aspects of corridors are less relevant in this group but come to light more in the next group.

  8. Logistics:
  The eight topic – *Logistic* – is related to the logistic aspects of corridors. In essence, the corridors serve to enhance the network of transportation for commodities, humans and/or information. This topic highlights that transportation function and the relevant matters that entail this aspect.

  9. Operational application:
  The ninth topic – *Operational application* – is related to the operational level of corridors. The ninth topic involves the 'in the field' affairs. An example of this is how transport corridors can help to solve the uneven coal distribution problem in China. In order to tackle this problem a possible solution could be to "enhance the railway corridors connecting major coal fields by upgrading the existing railway, building dedicated lines for freight, and facilitating intermodal transfers between rail and water systems" (Wang & Ducruet, 2014). These are all examples of the operational aspects around transport corridors.

- **Group 4**: (*2. Difficulties – 4. Corridor development – 7. Corridor methods – 10. Conceptual*)
  In contrast to group 3, these topics are grouped together because they share the commonality of an abstract/theoretical nature. These topics focus more on the optimality and functioning of a corridor based on a sound concept and method. The operational aspect is less relevant in these topics.

  2. Difficulties:
  The topic – *Difficulties* – has a focus on the challenges and solutions of the implementation of corridors. What differentiates this topic from the others is the emphasis on the inherent and external challenges that result from the complex structure of corridors. An example of this can be found in the bottlenecks relating to the infrastructure in Europe. Think of the lack of rail transport capacity, competition between freight transport, a lack of noise protection near housing areas adjacent to transport infrastructure, and many more bottlenecks (Witte et al., 2012).

  4. Corridor development:
  The topic – *Corridor development* – delves into the research and new developments of corridors. The aspects that are important in this topic are the potential improvements that result from research and previous cases of corridor

implementations. This topic entails conceptual and theoretical designs that focus on the development of corridors. An example of this is the research on potential new multi-corridors in the city of Athens (Tsigdinos et al., 2021). In this research, a method for the development of multi-corridors in the main urban core of the city is formulated. These corridors would result from a new operational categorization scheme that encourages inclusive and sustainable travel and places a higher priority on people than on vehicles.

7. Corridor methods:
The seventh topic – *Corridor methods* – focuses on different methods and aspects of corridors such as sea and land corridors. This topic is also theoretical and abstract in its nature, but unlike the fourth topic that focuses on development, this topic focuses on different methods and approaches that are possible with corridors. A corridor is not a homogeneous construct. It can take shape in many ways such as rail corridors, sea corridors, river routes, aerial routes, or even a combination of those which is known as intermodal transport. An example of this is the northeast corridor of Taiwan. In this corridor, the highway transportation is characterized by frequent gravel movement carried by trucks. However, the laden trucks damage the pavements causing other methods such as a combination of truck and rail movement to be considered (Shiau & Chuang, 2012).

10. Conceptual:
Topic 10 focuses on the conceptual aspects of corridors. This topic involves the scenarios of an optimal transport corridor. An example of a report wherein this topic comes to light is the report of Chapman et al., (2003). In this report, where the corridor of UK's West Midlands to London corridor was analyzed, the desirable performance and concept of the corridor was discussed and examined from different perspectives. For instance, it was stated in the report that "functionally and economically, it is desirable to make the (West Midland to London) corridor as short as possible while still providing effective access to all of the accommodation required". The tenth topic, *Conceptual*, corresponds to such analyses on the conceptual level.

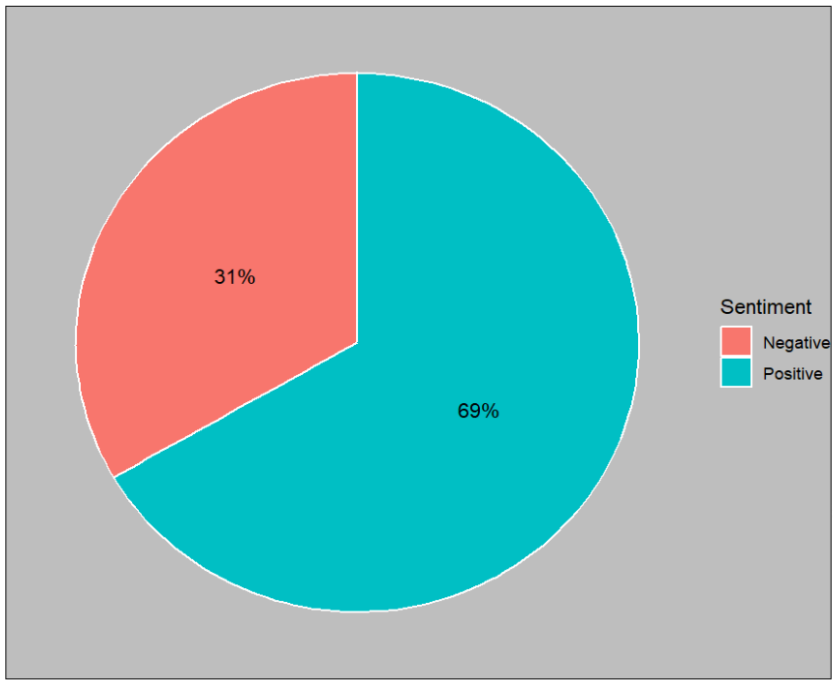
It is displayed that there are different topics within the corridor debate. Some topics seem to be more related to one another, while other topics share less commonalities. The variety of topics were not surprising after the analysis of the word cloud and keywords, but the relationships between some topics were a surprising notion which gives an interesting nuance to the topic modeling part of this research. The next part of this research will focus on the sentiment analysis in order to shed light on the overall sentiment within the corridor debate.

3.3 Sentiment analysis

In the previous part of the literature review, the topic modeling and keyword analysis revealed the various themes and concepts of the corridor debate. It did not however shed light on the sentiment of the debate, which is an important factor to consider when analyzing the corridor literature.

The overall sentiment of the corridor debate is positive. The proportion of positive words in the 78 papers is higher than that of negative words***. Figure 5: Mean sentiment proportion*** displays these proportions in a pie chart. After calculating the proportions of positive words in the 78 corridor papers, the mean positive proportion was found to be 69%. As for the proportion of negative words in the papers, the mean was 31%. This analysis indicates a significant prevalence of positive words compared to negative words.

Figure 5: Mean sentiment proportion



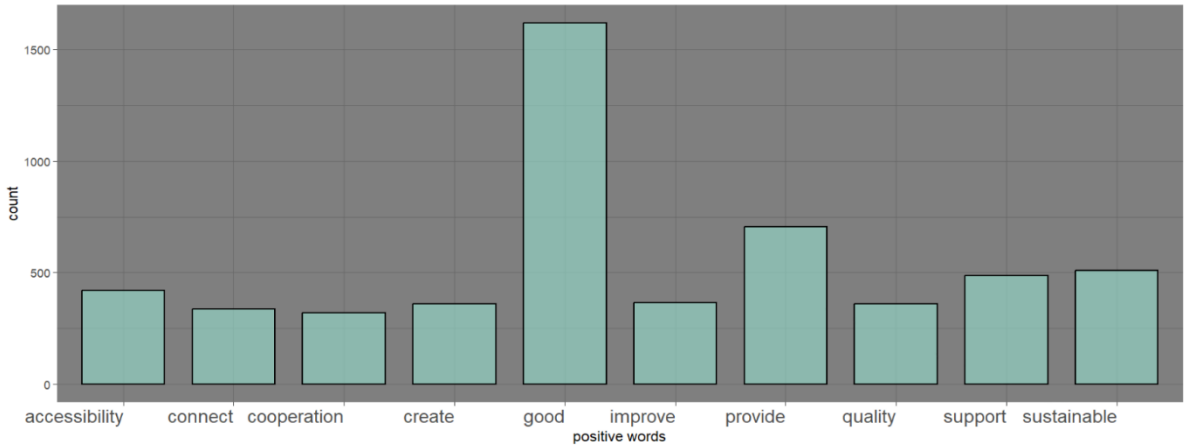
(Source: Authors' own)

The superiority of positive words indicates an overall optimistic or favorable tone in the debate. Such a positive sentiment might contribute to a more optimistic portrayal of corridor-related topics, affecting readers' views. It is important to note that the context and subtleties within the papers should be taken into account when interpreting sentiment. Nonetheless, the observed difference in mean proportions shows the significance of sentiment analysis in understanding the general tone and emphasis of the corridor papers.

An example of positive words used in the corridor literature is displayed in ***figure 6: frequent positive words***. This figure displays the most frequent positive words which can be found in the 78 papers. An analysis of these words provide readers of this literature review with an insight into the intrinsic values and aspirations ingrained in the corridor debate. Words such as '*connect*', '*improve*', '*provide*' and '*create*' express the beliefs of potential benefits that come with transport corridors. While words such as '*accessibility*', '*quality*', and '*sustainable*' highlight some of the positive attributes inherent in

a well-functioning transport corridor. The positive terms '*support*' and '*cooperation*' underscore the collaborative efforts and backing that transport corridors receive from various stakeholders, which plays a pivotal role in the establishment and success of the corridors. The word '*good*' connotes an assessment that is overall positive, indicating that the discourse typically concentrates on positive results and benefits related to transportation corridors.

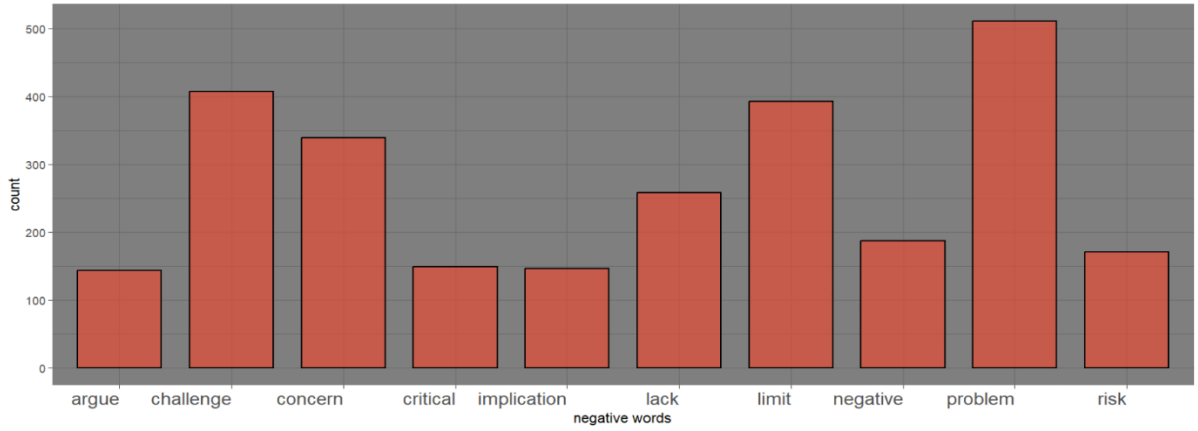
Figure 6: Frequent positive words



(*Source: Authors' own*)

Similarly, ***figure 7: frequent negative words*** displays the most frequent negative words used in the corridor papers. These negative words focus on areas of concern and challenges of transport corridors. The words '*challenge*', '*implication*', '*problem*' and '*risk*' underscore the various bottlenecks and potential chokepoints that arise in the implementation or management of transport corridors. The words '*argue*', '*concern*' and '*critical*' relate to the considerations and debates on corridors as result of their complex designs and nature. As for the terms '*lack*' and '*limit*', they draw attention to possible shortcomings or limitations that could prevent transport corridors from operating as efficiently as possible or from realizing the benefits that are intended. The word *'negative'* denotes a reserved or unfavorable stance when addressing certain aspects of transport corridors, recognizing negative aspects or difficulties in the debate. The word '*negative*' is in its explanation similar to the word '*good*' in the frequent positive words plot. It is noteworthy to emphasize that the word '*good*' occurs many more times than the word '*negative*'. This mirrors the positive/negative proportion illustrated in the pie chart introduced earlier in this section of the literature review.

Figure 7: frequent negative words



(*Source: Authors' own*)

Exploring papers with high positive or negative proportions provides deeper context to sentiments in the corridor debate, offering an additional understanding of transport corridors by illustrating the topics and themes that evoke such sentiments. ***Table 2: High negative proportion*** displays the five corridor papers with the highest negative proportion. The *doc_id* is the identification value of the individual papers in the Excel file. The topics discussed in these papers include: challenges of multilevel governance (Van Straalen & Witte, 2018), economic impacts of an interruption to corridors because of flood waters (Rolfe et al., 2013), negative aspects of freight transport corridors such as noise, traffic accidents, emission of greenhouse gasses, etc. (Janić & Vleugel, 2012), negative impacts of heavy trucks in Taiwan's North-East corridor (Shiau & Chuang, 2012), and bottlenecks in the European Corridor 24 (Witte & Spit, 2014).

An examination of the corridor papers with the highest positive proportion, displayed in ***table 3: High positive proportion***, gives a glance into some of the themes that are related to favorable sentiments. Discussed themes include: possible impact of green corridors (Panagakos et al., 2015), as well as strategies and approaches for integrating land use and transportation planning and decision making (Rooney et al., 2010). Furthermore, insights from successful corridors in Munich (Hale, 2010) and Tokyo (Chorus & Bertolini, 2016) were also associated with high positive sentiments. This pattern was similarly observed for the topic of sustainability in EU documents (Öberg et al., 2017).

The examination of documents with high positive or negative proportions offers significant insights into the various sentiments within the corridor literature. This examination sheds light on important obstacles, implications, and possible solutions, adding to a thorough comprehension of perspectives on transportation corridors.

Table 2: High negative proportion

doc_id <chr>	negative <dbl>	positive <dbl>
cd46	0.65	0.35
cd26	0.56	0.44
cd22	0.56	0.44
cd19	0.55	0.45
cd63	0.49	0.51

(*Source: Authors' own*)

Table 3: High positive proportion

doc_id <chr>	negative <dbl>	positive <dbl>
cd76	0.11	0.89
cd73	0.14	0.86
cd52	0.14	0.86
cd68	0.15	0.85
cd79	0.16	0.84

(*Source: Authors' own*)

3.3.1 Geographical sentiment analysis

After analyzing the keywords and word cloud, it became apparent that '*Europe*' is one of the most frequent words used in the corridor papers. This might be an indication that the region Europe is the more prominent geographical focus within the corridor debate. An examination of the different geographical regions will help readers interested in transport corridors gain a perspective on the variety of corridors. Moreover, a sentiment analysis per region could answer the question of whether corridors in different geographical locations are perceived with differing sentiments in the corridor debate, adding an extra layer to the comprehensive view this literature review provides.

With regard to the regions, there are roughly 10 geographical regions that authors have focused on when writing about corridors. Additionally, some authors did not focus on a geographical region and wrote about corridors at a conceptual level. The ten regions are: *Europe*, *North-America*, *Asia*, *Europe-Asia comparison*, *Europe-America comparison*, *Pacific*, *United Kingdom*, *Africa*, *Middle-East*, and *Latin-America*. These regions contain various transport corridors causing them to be an object of focus in the corridor debate.

Most authors focused on the region Europe when writing about corridors. ***Table 4: Regional focus of authors*** displays the count of authors who wrote about a specific region, be it individually or in collaboration.

Table 4: Regional focus of authors

Region	Authors
Europe	76
North-America	17
Asia	12
Europe-Asia comparison	12
Europe-America comparison	10
U.K.	8
Pacific	7
Africa	6
Latin-America	5
Middle-East	2
Unknown/conceptual	12

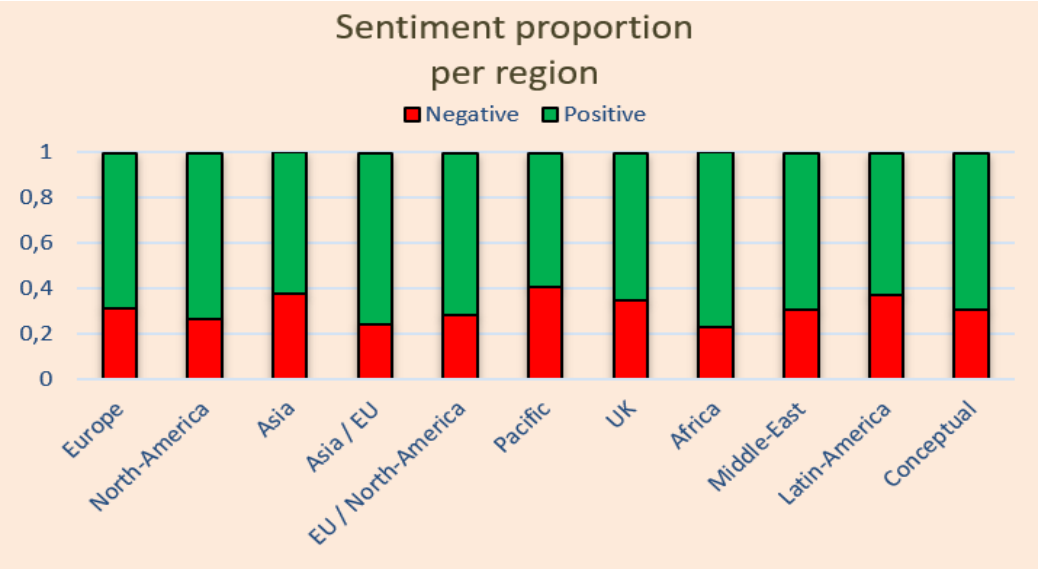
(*Source: Authors' own*)

The table above displays the number of authors that have written about corridors focusing on a specific region (or no region at all). The amount of corridor papers that were analyzed in this literature review amounted to only 79, it is therefore noteworthy to mention that most papers were written in collaboration between two or more authors. Using data transformation techniques, the authors that wrote about corridors were split up into individuals and processed in new data frames. This method enables a more in-depth analysis, making it easier to pinpoint writers and their contributions to particular areas in the context of corridor conversations.

The table shows that most authors wrote papers about corridors focusing on the region Europe. This is in line with the results of the word cloud, which suggested that '*Europe*' is one of the most occurring words. When considering that authors have also written about the *Europe/Asia* and *Europe/North-America* comparisons, it further emphasizes how the region Europe comes to the forefront in the corridor debate.

Besides the region Europe, the other 9 regions are also subjected to the corridor debate, albeit in lesser frequencies. Within the context of these regions, it is of interest to analyze the sentiments in the papers related to those specific geographic locations. The next figure, ***figure 8: sentiment proportion per region***, displays the proportion of positive versus negative words per region.

Figure 8: sentiment proportion per region



(*Source: Authors' own*)

The figure above displays the sentiment proportion per region. Similar to the general sentiment analysis conducted previously in this report, it was within expectations that the overall sentiment in most regions would be positive. The reason for this expectation is because corridors in general aim to bring certain benefits. Be it a perceived economic benefit, an aid for citizens, or a facilitation for information or merchandise circulation, a transport corridor will mostly bring benefits or solve a problem. So, when a transport corridor is an inherently positive instrument, it is expected that the overall sentiment would also be positive.

There is also a proportion of negativity observed in the figure above. From the previous analysis it became apparent that the negativity is, among other things, the result of bottlenecks and difficulties related to corridors and corridor development. The results of the topic modeling analysis also align to this trend of thought, by dedicating one of the topics (***topic 2: Difficulties***) to the problems and chokepoints of transport corridors. It is therefore no surprise that a proportion of the sentiment is negative, but in most regions this proportion lies around 20 to 30 percent, which is significantly smaller than the proportion of the positive sentiment.

As for the regions between themselves, there does not seem to be much difference with regards to the positive-negative proportion. All regions – and the conceptual papers – follow the same trend with most of the sentiment being positive. These results are somewhat surprising, because one would have expected areas with a greater concentration of developing nations, such as Africa and Latin America, to show a more pronounced negative sentiment. This presumption is based on the idea that, in comparison to more economically developed regions like Europe, these developing regions may face proportionally greater obstacles, such as inadequate infrastructures and underdeveloped logistical sectors. However, it might also be the case that despite the bottlenecks, transport corridors still have a lot of potential benefits for these underdeveloped countries (Quium, 2019), causing the overall sentiment to still be high.

4. Discussion

This section of the literature review will delve into the discussion of the limitations of this study, as well as the recommendations for further research. Before that, a short summary of the results will be provided.

Summary of results

The results of this study show that the corridor debate entails various subtopics. Some of the most frequent words occurring in the debate are '*plan*', '*European*', and '*spatial*'. These words, among others, are also considered keywords based on the TF-IDF scores. Furthermore, the corridor debate is a rich debate delving into various aspects such as: strategies, difficulties, national aspects, corridor development, governmental needs, urban designs, corridor methods, logistics, operational applications, conceptual focusses, and economic values. The overall sentiment in the debate is positive, which is related to the beneficial aspects of transport corridors as shown by the most frequent positive words. A small proportion of the debate is negative, which relates to the various difficulties that arise due to the complexity of transport corridors. The overall sentiment does not seem to differ between geographical regions.

Limitations

The objective of this study was to provide a holistic view of the corridor debate, using text mining techniques in a literature review. This study does indeed provide readers interested in transport corridors with insights and understanding of corridors, but some conclusions cannot be made due to certain limitations. The first limitation is the fact that all the collected papers were written in the English language. Papers about transport corridors written in other languages were excluded from the analysis, while they could potentially complement the literature review with valuable insights and depth. Another limitation of this study is the search engine – Google Scholar – used to find the relevant papers. All 79 papers resulted from a search query in Google Scholar. Although this search engine has a broad coverage of academic papers, it might be the case that some relevant papers are not indexed in Google Scholar yet can be found in other databases. Another limitation is that this study did not include a temporal analysis. This means that conclusions about evolving trends or changes over time cannot be made.

The secondary objective of this study was to demonstrate the use of text mining techniques in a literature review, showcasing the benefits of this combination. This practice is not new, Karami et al. (2020) used text mining techniques in a literature review on academic papers which used twitter data, while Feng et al. (2017) used text mining techniques to conduct a systematic literature review. The benefits of using text mining in a literature review are apparent, as seen by the holistic view this study provides on the context of transport corridors. However, these text mining techniques have limitations of their own. One important limitation is related to the ambiguous nature of language. This is not only the case with homonyms, where one word can have multiple meanings, but also the other way around, where multiple words convey the same meaning. This ambiguity cannot be entirely eliminated from natural language, and it might lead to noise in the extracted information (Gaikwad et al., 2014).

A concrete example is the *hash_lemmas* lemmatization list which was used in this study and is available in R. This list contains tokens in one column, and their corresponding lemma form in the other column. One of the tokens '*number*' shows its lemma to be '*numb*'. This would be correct in the context that number relates to the comparative degree of the adjective numb (numb-***number***-numbest). However, in the context that number relates to a digit, it would be wrong to lemmatize it to numb. Especially since numb has a negative meaning when performing sentiment analysis, while

number (as in digit) does not. If remained unnoticed, this would lead to inaccurate results when performing a sentiment analysis.

This problem is also related to the difficulty of text mining tools to read the context within phrases. Humans have no problem in understanding that 'blind as a bat' relates to an animal and 'swinging a bat' relates to a sport tool. Similarly, by knowing the context in a phrase it is easy for humans to understand when the word 'bow' is a noun or a verb. But these distinctions still prove to be challenges for text mining tools. Moreover, the context is often related to the specific field wherein the study is conducted. Though it is possible for a set of rules to be established and then embedded as plug-ins into text mining tools, still, it is an arduous task that requires domain knowledge and entails lots of time and effort (Talib et al., 2016).

Recommendations for further research

Avenues for future studies include the incorporation of multilingual corridor papers into the literature review. Text mining can play a large role in facilitating this endeavor. Cross-lingual text retrieval techniques can be employed to find relevant corridor papers written in different languages (Chau & Yeh, 2004). Then, associations between those multilingual corridor papers could be extracted by clustering and organizing the papers into hierarchies, using the growing hierarchical self-organizing map model proposed by Yang et al. (2009). Furthermore, corridor papers retrieved from search engines other than Google Scholar, such as Corpus, could be included in the literature review.

Another recommendation is the use of deep neural network models for the extraction of keywords within corridor papers. One such model is the tc-LSTM model (target center-based Long Short-Term Memory) proposed by Zhang et al. (2020). This model can capture sentence-level information of keywords, by modeling both the preceding and following contexts of the words at the same time. This might enhance the quality of the extracted keywords, conveying a better understanding of the corridor debate. Moreover, word vectors such as Word2Vec model can be used in future studies to calculate similarity scores for words (Jatnika et al., 2019). This may help in uncovering the semantic relationships between words, putting the context of words into perspective when conducting the literature review. Also interesting for further research is the incorporation of a time element into the topic modeling analysis. The use of dynamic topic models will allow for an analysis of the time evolution of topics in the various transport corridor papers (Blei & Lafferty, 2006). This temporal aspect of the literature review can help in portraying various trends and themes that develop over time. Incorporating these recommendations might help to deal with the limitations of this study and provide more depth to the literature review on transport corridors.

References

Abid, M., Mushtaq, M., Akram, U., Abbasi, M., Rustam, F. (2023). Comparative analysis of TF-IDF and loglikelihood method for keywords extraction of twitter data. Mehran University Research Journal of Engineering and Technology. 42. 88. 10.22581/muet1982.2301.09.

Ahmed, M. (2018). THE ECONOMICS AND POLITICS OF CHINA- PAKISTAN ECONOMIC CORRIDOR AND BALOCHISTAN. Regional Studies. 36. 70-111.

Blei, D.M. (2011). Probabilistic topic models. *Communications of the ACM, 55*, 77 - 84.

Blei, D.M., & Lafferty, J.D. (2006). Dynamic Topic Models. ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning. 2006. 113-120. 10.1145/1143844.1143859.

Bruinsma, F.R., Rienstra, S.A., & Rietveld, P. (1997). Economic Impacts of the Construction of a Transport Corridor: A Multi-level and Multiapproach Case Study for the Construction of the A1 Highway in the Netherlands. Regional studies, 31(4), 391-402.

Chapman, D., Pratt, D., Larkham, P., Dickins, I., (2003). Concepts and definitions of corridors: Evidence from England's Midlands. Journal of Transport Geography 11 (3), 179-191.

Chau, R., & Yeh, C. (2004). A multilingual text mining approach to web cross-lingual text retrieval. *Knowl. Based Syst., 17*, 219-227.

Chorus, P., & Bertolini, L. (2016). Developing transit-oriented corridors: Insights from Tokyo. *International Journal of Sustainable Transportation, 10*, 86 - 95.

De Borger, B., Dunkerley, F., & Proost, S. (2006). Strategic Investment and Pricing Decisions in a Congested Transport Corridor. *ERN: Pricing (Topic)*.

Feng, L., Chiam, Y.K., & Lo, S.K. (2017). Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review. *2017 24th Asia-Pacific Software Engineering Conference (APSEC)*, 41-50.

Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3), 259-291. doi:10.1017/S1351324919000457

Gaikwad, S.V., Chaugule, A.A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications, 85*, 42-45.

Griffiths, T.L., Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228-5235

Gupta, E.T. (2017). KEYWORD EXTRACTION : A REVIEW.

Gurusamy, Vairaprakash & Kannan, Subbu. (2014). Preprocessing Techniques for Text Mining.

Hale, C. (2010) The Mega-Project as Crux of Integrated Planning: Insights from Munich's Central Corridor, Planning Practice & Research, 25:5, 587-610, DOI: 10.1080/02697459.2010.522856

Hesse, M. & J.P. Rodrigue (2004), The transport geography of logistics and freight distribution. Journal of Transport Geography 12 (3), pp. 171-184.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods, 25*, 114 - 146.

Jain, M., & Jehling, M. (2020). Analysing transport corridor policies: An integrative approach to spatial and social disparities in India. Journal of Transport Geography, 86, 102781.

Janić, M., & Vleugel, J. (2012). Estimating potential reductions in externalities from rail–road substitution in Trans-European freight transport corridors. *Transportation Research Part D-transport and Environment, 17*, 154-160.

Jatnika, D., Bijaksana, M.A., & Suryani, A.A. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. *Procedia Computer Science*.

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y.K. (2020). Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access, 8*, 67698-67717.

Khyani, Divya & B S, Siddhartha. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology. 22. 350-357.

Kumar, S.A. (2012). Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering. *International journal of engineering research and technology, 1*.

Machidon, V. (2015). "Major European transport corridors and their role in the Republic of Moldova development," Eastern European Journal for Regional Studies (EEJRS), Center for Studies in European Integration (CSEI), Academy of Economic Studies of Moldova (ASEM), vol. 1(2), pages 1-142.

Notteboom, Theo & Rodrigue, Jean-Paul. (2005). Port Regionalization: Towards a New Phase in Port Development. Maritime Policy & Management. 32. 10.1080/03088830500139885.

Öberg, M., Nilsson, K.L., & Johansson, C. (2017). Major transport corridors: the concept of sustainability in EU documents. Transportation research procedia, 25, 3694-3702.

Öberg, M., Nilsson, K.L., Johansson, C. (2016). Governance of Major Transport Corridors Involving Stakeholders. Transportation Research Procedia, 14, 860-868.

Panagakos, G., Psaraftis, H.N., & Holte, E.A. (2015). Green corridors and their possible impact on the European supply chain. In Handbook of ocean container transport logistics (pp. 521-550). Springer, Cham.

Priemus, H. & W. Zonneveld (2003), What are corridors and what are the issues. Introduction to special issue: the governance of corridors. Journal of Transport Geography 11 (3), pp. 167-177.

Quium, A.S. (2019). Transport Corridors for Wider Socio–Economic Development. *Sustainability*.

Reggiani, A., Lampugnani, G., Nijkamp, P., & Pepping, G. (1995). Towards a typology of European inter-urban transport corridors for advanced transport telematics applications. Journal of Transport Geography, 3(1), 53-67.

Rodrigue, J.P. (2004), Freight, gateways and mega-urban regions: The logistical integration of the Bostwash corridor. Tijdschrift voor Economische en Sociale Geografie 95 (2), pp. 147-161.

Rolfe, J., Kinnear, S.L., & Gowen, R. (2013). SIMPLIFIED ASSESSMENT OF THE REGIONAL ECONOMIC IMPACTS OF INTERRUPTION TO TRANSPORT CORRIDORS WITH APPLICATION TO THE 2011 QUEENSLAND FLOODS. *The Australasian Journal of Regional Studies, 19*, 215.

Rooney, K., Savage, K., Rue, H., Toth, G., & Venner, M. (2010). Corridor Approaches to Integrating Transportation and Land Use. *Transportation Research Record, 2176*, 42 - 49.

Shiau, T.A., & Chuang, Y.R. (2012). Evaluating gravel transport sustainability: A case study of Taiwan's northeast corridor. Transportation Research Part D: Transport and Environment, 17(4), 287-292.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics.

Talib, R., Hanif, M.K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications, 7*.

Tsigdinos, S., Nikitas, A., & Bakogiannis, E. (2021). Multimodal corridor development as a way of supporting sustainable mobility in Athens. Case Studies on Transport Policy, 9(1), 137-148.

Van Straalen, F.M., & Witte, P. A. (2018) Entangled in scales: multilevel governance challenges for regional planning strategies, Regional Studies, Regional Science, 5:1, 157-163, DOI: 10.1080/21681376.2018.1455533

Wang, C., & Ducruet, C. (2014). Transport corridors and regional balance in China: the case of coal trade and logistics. Journal of Transport Geography, 40, 3-16.

Williams, B., Berry, J., & McGreal, S. (2002). The East coast corridor: spatial development strategies for the Dublin-Belfast metropolitan regions. Journal of Irish Urban Studies, 1(2), 19-31.

Witte, P., & Spit, T. (2014). Sectoral drawbacks in transport : Towards a new analytical framework on European transport corridors.

Witte, P., Wiegmans, B., Oort, F. van & T. Spit (2012), Chokepoints in corridors: Perspectives on bottlenecks in the European transport network. Research in Transportation Business & Management 5, pp. 57-66.

Witte, P., Wiegmans, B., Braun, C., & Spit, T. (2016). Weakest link or strongest node? Comparing governance strategies for inland ports in transnational European corridors. Research in transportation business & management, 19, 97-105.

Yang, H., Lee, C., & Chen, D. (2009). A method for multilingual text mining and retrieval using growing hierarchical self-organizing maps. *Journal of Information Science, 35*, 23 - 3.

Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., & Liu, T. (2020). Keywords extraction with deep neural network model. *Neurocomputing, 383*, 113-121.