


Review

# Applications of Text Mining in the Transportation Infrastructure Sector: A Review

Sudipta Chowdhury <sup>1,\*</sup> and Ammar Alzarrad <sup>2</sup> <sup>1</sup> Department of Mechanical and Industrial Engineering, Marshall University, Huntington, WV 25755, USA<sup>2</sup> Department of Civil Engineering, Marshall University, Huntington, WV 25755, USA

\* Correspondence: chowdhurys@marshall.edu

**Abstract:** Transportation infrastructure is vital to the well-functioning of economic activities in a region. Due to the digitalization of data storage, ease of access to large databases, and advancement of social media, large volumes of text data that relate to different aspects of transportation infrastructure are generated. Text mining techniques can explore any large amount of textual data within a limited time and with limited resource allocation for generating easy-to-understand knowledge. This study aims to provide a comprehensive review of the various applications of text mining techniques in transportation infrastructure research. The scope of this research ranges across all forms of transportation infrastructure-related problems or issues that were investigated by different text mining techniques. These transportation infrastructure-related problems or issues may involve issues such as crashes or accidents investigation, driving behavior analysis, and construction activities. A Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)-based structured methodology was used to identify relevant studies that implemented different text mining techniques across different transportation infrastructure-related problems or issues. A total of 59 studies from both the U.S. and other parts of the world (e.g., China, and Bangladesh) were ultimately selected for review after a rigorous quality check. The results show that apart from simple text mining techniques for data pre-processing, the majority of the studies used topic modeling techniques for a detailed evaluation of the text data. Other techniques such as classification algorithms were also later used to predict and/or project future scenarios/states based on the identified topics. The findings from this study will hopefully provide researchers and practitioners with a better understanding of the potential of text mining techniques under different circumstances to solve different types of transportation infrastructure-related problems. They will also provide a blueprint to better understand the ever-evolving area of transportation engineering and infrastructure-focused studies.

**Keywords:** natural language processing (NLP); text mining; transportation; infrastructure; review



**Citation:** Chowdhury, S.; Alzarrad, A. Applications of Text Mining in the Transportation Infrastructure Sector: A Review. *Information* **2023**, *14*, 201. <https://doi.org/10.3390/info14040201>

Academic Editor: Katsuhide Fujita

Received: 5 March 2023

Revised: 21 March 2023

Accepted: 22 March 2023

Published: 23 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Natural language processing (NLP) is a subfield of artificial intelligence in which machine learning and computational linguistics are broadly used [1]. NLP is devoted to making computers understand the statements or words written in human languages. Essentially, NLP can be used to understand the structure and meaning of human language by analyzing characteristics such as syntax, semantics, pragmatics, and morphology [2]. This can be achieved in several ways such as tokenization, stop word removal, lemmatization and stemming, co-reference resolution, sentiment analysis, topic discovery and modeling, and part-of-speech tagging. Such linguistic knowledge can be converted into rule-based and machine learning-based systems to solve specific problems for which the NLP method was designed. Rule-based systems are the traditional methods for data processing that were originally developed based on customized linguistic rules [3]. On the other hand, machine learning-based systems learn to perform tasks based on training data they are fed and update their methods as more data is processed. These systems typically employ a

combination of machine learning, deep learning and neural networks, and natural language processing algorithms. Over the years, NLP has been used in a variety of applications such as customer feedback analysis, customer service automation, stock forecasting and insights into financial trading, and the analysis and categorization of records (e.g., of crashes/accidents, medical records, and supply chain records) [4].

Text mining uses NLP to transform the unstructured text in documents and databases into normalized, structured data suitable for further analysis [5–7]. Text mining is a flexible process that can be used as a pre-processing step for further data mining or as a standalone process for specific tasks. These tasks may involve information extraction, information retrieval, document clustering, text categorization, and text visualization [8]. Text mining is widely used in various fields such as biomedical applications [9], market prediction [10], business process management [11], sentiment analysis [12], and social media analysis [13]. Frequency analysis is arguably the most popular text mining technique that is based on the frequency of occurrence of tokens [14,15]. In addition to frequency analysis, text mining can also be used to define and quantify associations of words (e.g., association rule mining) [16,17]. Such an example of word association analysis can be used to construct a knowledge network using words as nodes and associations as edges. Text mining has also been widely applied for advanced text analysis that includes text classification, clustering, topic modeling, and sentiment analysis. Machine learning algorithms such as Naïve Bayes (NB), support vector machines (SVM), the hidden Markov model (HMM), gradient boosting trees, random forests, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) are often used to build the text classification models [18–20]. Hierarchical clustering algorithms, partitioning algorithms, and hybrid methods using both hierarchical and partitioning clustering algorithms (that are distance-based) are primarily used for text clustering (i.e., grouping text based on similarity) [21,22]. Topic models such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Probabilistic Latent Semantic Analysis (PLSA) are common machine learning algorithms used in topic modeling to discover patterns of word use in documents and connect documents that share similar patterns [23,24]. NB and SVM as well as different topic models and neural networks are also commonly used in sentiment analysis tasks that study people’s opinions, emotions, and attitudes toward entities, issues, events, topics, and their attributes [25–27]. As evidenced by the multitude of benefits, text mining has become an essential tool for organizations to make data-driven decisions via extracting insights from unstructured text data. In essence, text mining can (1) help identify *the key concepts and the main stakeholders described in a large text corpus, as well as their relationships with minimum human intervention*, (2) *be applied regardless of the different formats in which the text appears in (i.e., with a lack of consistent structure)*, and (3) *help unlock hidden information that can lead to new knowledge and improved understanding*. Such benefits and the applicability of text mining techniques make them suitable for application in the transportation infrastructure domain.

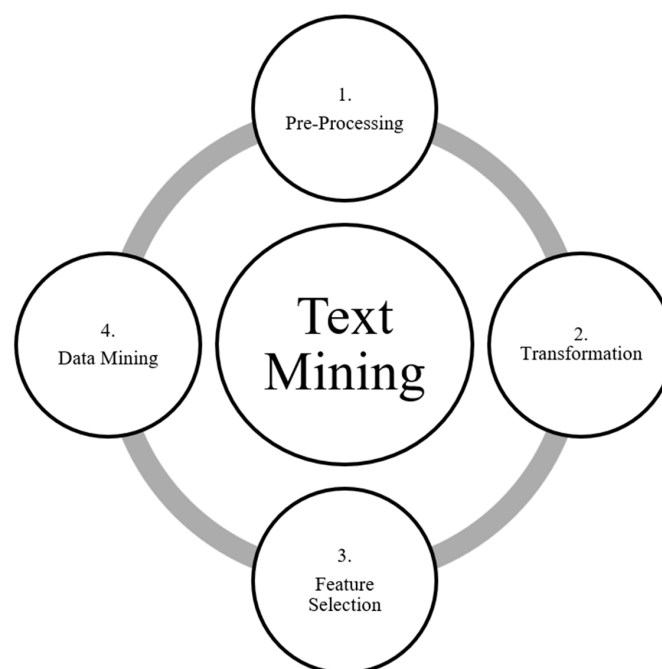
The domain of transportation infrastructure research and practice is broad, ensuring an interdisciplinary coverage of topics that include roadway traffic and mobility-related issues, infrastructure construction, crash/accidents, and supply chain and logistics. Due to the consistent evolution in technology and resultant impacts on data collection and processing, the transportation research domain has experienced an upsurge in publications/text data in recent decades [28]. These publications provide unique opportunities for researchers to conduct research via parsing these documents produced by a wide variety of entities such as the U.S. Department of Transportation (DOT) and state DOTs, government agencies, private organizations, and individuals (e.g., those behind social media posts). Integrating information from all these text documents could result in large text datasets with a high number of variables. Many of the researchers who analyzed these publications/text data adopted text mining-based approaches to identify important concepts, find hidden patterns among the texts, and identify the interrelationships for the prediction and decision-making process [29–32].

Large-scale reviews of academic research related to the application of text mining techniques in transportation infrastructure sector research are rather limited. This comprehensive review attempts to identify major research themes, types of text mining techniques employed, the prevalence of topics, and the scope and focus of different studies. To this end, this study has developed a Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)-based structured methodology to identify relevant studies that use different text mining techniques across different transportation infrastructures. The PRISMA method can conduct systematic reviews and meta-analyses effectively by summarizing aggregate data from studies, particularly the evaluations of the effects of interventions [33,34]. The PRISMA method has been proven to be useful for the critical appraisal of text data across a wide array of research domains. In this study, using the PRISMA method, the existing literature has been categorized based on how text mining techniques have been implemented in different transportation infrastructure sectors. Essentially, the types of text mining techniques, innovations in the application of these techniques, the type of data analyzed, and the scope of these applications are described in detail. The goal of this research is to identify the current state-of-the-art text mining techniques that are used for transportation infrastructure assessment and planning and the potential of new and existing text mining techniques in different transportation sectors. This could help decision-makers and planners on federal, state, and local levels to develop holistic approaches to plan, build, and manage our transportation infrastructure based on the application of different text mining techniques in existing databases.

The exposition of this study is as follows. First, the steps involved in the application of text mining techniques are explained. Second, the research steps adopted in this study are described in detail. Third, the findings of the literature review are categorized and presented. Fourth, the implications, managerial insights, contributions, and future research directions are highlighted.

## 2. Process of Text Mining

Before delving into the literature review, it is critical for readers to understand how text mining works in theory. It will be beneficial as the readers can better conceptualize how researchers have used text mining techniques in their respective research domains. The process of text mining is illustrated in Figure 1 and described below.



**Figure 1.** Process of Text Mining.

### 2.1. Text Pre-Processing

Text pre-processing transforms text into a form that is predictable and analyzable for text mining tasks. Text data in general contains noise in various forms such as punctuation marks that may be applied haphazardly across different documents. Efficient and effective text pre-processing can help to increase the accuracy of text mining results. Text pre-processing in general consists of four major steps: *tokenization*, *normalization*, *stemming*, and *lemmatization*.

*Step 1—Tokenization:* During tokenization, the text is treated as a string which is later split into smaller pieces, or “tokens”. Tokens can consist of words, phrases, and subwords such as n-grams, or characters. Traditionally, paragraphs and sentences are tokenized into sentences and words, respectively, in the majority of text pre-processing applications. The most widely used tokenization process is whitespace tokenization, where the entire text is split into words by splitting them from whitespaces.

*Step 2—Normalization:* Normalization refers to the process of converting a token into its original/base form. It aims to put all text on a level playing field. Essentially, the inflectional or enunciated form of a word is replaced by the base form. At the end of the normalization process, all the text is converted into either upper or lower case with different diacritics removed so that the text data is converted into a standard format.

*Step 3—Stemming:* A stem is a natural set of words with similar meanings. Stemming removes inflationary forms (e.g., different grammatical forms) from a given token. For example, the word “planning” can be reduced to the stem “plans”. The stem is usually a full word, but it may not depend on other pre-processing steps that precede the stemming process. Moreover, due to simple suffix rules, sometimes the final stem is not appropriate. For example, due to simple suffix rules, the tokens *universal*, *university*, and *universe* may be reduced to the stem *univers*, which is not accurate.

*Step 4—Lemmatization:* Lemmatization is essentially an improvement in the stemming process that modifies the simple suffix rules that are commonly used for stemming. The lemmatization process adopts different rules depending on a word’s lexical category as well as using information from different computational repositories to obtain the correct base forms of words. Lemmatization can only be performed if the given word has proper parts-of-speech tags. Part-of-speech tagging assigns parts of speech to each word of a given text based on context. This process typically has to cope with ambiguous word-tag mappings for complex text data, especially if the data comes from scientific or any niche disciplines where jargon is used. Depending on the task need and problem context, either stemming, lemmatization, or both can be applied. The choice may also vary depending on the availability of computational resources as lemmatization is typically more computationally expensive.

### 2.2. Transformation

This process mainly involves document representation by the text it contains and the number of occurrences. There are two methods for representations of such documents, which are the *vector space model* and *bag of words (BoW) model*. The vector space model is an algebraic model that represents text as vectors. Each vector dimension corresponds to a term that appears in their text. To determine the similarity between words, weight factors are used that record the importance of the term to the text. Cosine similarity is often used to determine the similarity between vectors. On the other hand, the BoW model involves representing a text document in terms of word frequency counts and developing an overall frequency distribution of words in the document. The underlying assumption is that the meaning of the document can be captured by the frequency distribution of the words it contains.

### 2.3. Feature Selection

Feature selection is also known as attribute/variable selection. It is the process of selecting the most relevant features from the available variables that reduce the predictive

error. The irrelevant features often increase the computational complexity and decrease the accuracy of the analysis. The feature selection process also helps with dimensionality reduction and overfitting problems, resulting in less noise during the selection process.

#### 2.4. Data Mining

Once the text data is structured, different data mining techniques can be applied to analyze the data and generate useful insights. These data mining techniques may include classification, clustering, regression, association rule mining, and many more. Depending on the context of the problem and the needs of the research, different data mining techniques may be adopted. For example, if the interrelationships between different factors or the strength of interrelationships need to be identified, association rule mining techniques may be adopted. If the emotion, mood, or feeling expressed in the document—whether it is positive, negative, or neutral—needs to be identified, sentiment analysis can be adopted, which is a type of text classification technique. Data mining facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.

Note that it is not required that a research study or practitioner applies all the above-mentioned steps together while investigating a particular problem. A research study or practitioner may only adopt a particular subset of these steps (e.g., only preprocessing and transformation). This is also true for different substeps (e.g., only *stemming and lemmatization*). In this study, while identifying articles from the existing literature, it was ensured that all studies that adopted any of the text mining steps or substeps in any combination would be included during the evaluation.

### 3. Research Methodology

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method was adopted to conduct the systematic literature review. The PRISMA method provides a structured methodology that describes how relevant data can be collected and analyzed from available studies. The underlying goal is to present an explicit and reproducible method that can identify, choose, and critically assess studies that are relevant to specific research goals. The PRISMA statement includes a 27-item checklist and a flow chart of four main phases addressing the introduction, methods, results, and discussion sections of a research study [33]. Figure 2 illustrates the flowchart of the systematic review adopted in this study according to the PRISMA method.

#### 3.1. Information Sources and Search Strategy

The literature search was conducted through bibliographic databases such as Scopus and Google Scholar. Apart from that, renowned international scientific, technical and medical publishing databases, such as Elsevier-ScienceDirect, Springer, MDPI, SAGE journals, and Taylor and Francis, were also reviewed. The following search criteria were entered during the search: “text mining and transportation”, “text mining and transportation construction”, “text mining and traffic”, “text mining and crash”, “text mining and accident”, “text mining and supply chain and logistics”, “text mining and transportation innovation”, and “text mining and transportation planning and management”. This review was conducted for a one-week period between 21 January and 28 January 2023.

#### 3.2. Screening Criteria

There was no specific time frame considered regarding when studies were published. There were also no restrictions placed on the country of origin of the publications. Regarding the screening procedure, first, the software Mendeley was used for a duplicate check. As different online databases were checked simultaneously, some studies were present across multiple of them. These duplicates were removed to ensure the uniqueness of each study. A few additional removal criteria were implemented to ensure consistency in the identified literature that included (1) book chapters or complete books, (2) grey literature (i.e., reports, working papers, newsletters, government documents, and white papers), and (3) study languages

not English. All the transportation-focused studies that matched the search criteria and used text mining techniques during the investigation were included for the study selection phase. This included a wide array of scholarly journal articles and conference proceedings.

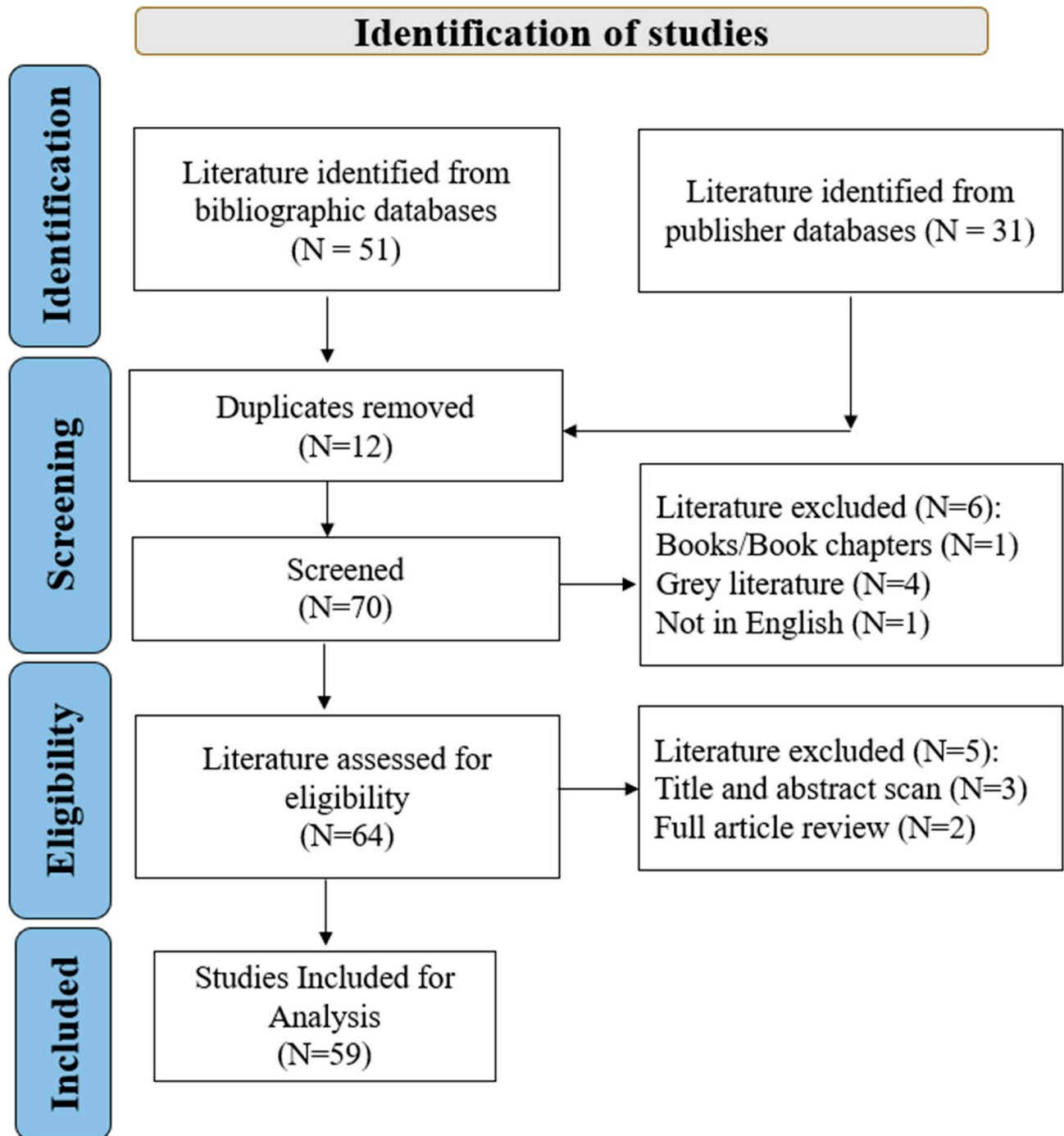


Figure 2. Literature selection criteria based on PRISMA.

### 3.3. Eligibility Criteria

A quick title, abstract, and keyword scan was done to ensure that the scope of the study was appropriate, meaning it focused on transportation-related topics and adopted one or more text mining techniques/approaches. If the scan could not provide adequate clarity, the studies were also removed from consideration. Afterward, the remaining articles

in their entirety were reviewed by the authors to ensure that the articles discussed the application of text mining techniques within the transportation context.

#### 3.4. Data Extraction, Storage, and Analysis

A detailed coding sheet in Excel was developed after finalizing the literature that included categories such as author, title, type (qualitative vs. quantitative), year of publication, keywords, focus group, sample size, analysis criteria, tools and techniques, main findings, implications, and limitations. A final check was also conducted to ensure that enough information was available in the Excel sheet for further analysis and that the contents matched the requirements for addressing the research questions.

### 4. Text Mining and Transportation Infrastructures

This section was categorized based on the different types of problems in the transportation domain that were investigated using text mining techniques. Six categories of transportation domain problems were identified that included *crashes and accidents, mobility analysis, supply chain and logistics, construction and urban infrastructure, reviews of literature, and innovation in transportation infrastructure research*. Each one of these categories will be explained in the next sections.

#### 4.1. Crashes and Accidents

Text mining techniques have been widely applied in crash/accident investigations across a wide range of sectors. The findings regarding the application of text mining techniques in crash/accident investigations have been categorized into three parts: *those in the roadways, railways, and other sectors*.

##### 4.1.1. Roadway

The majority of the studies that applied text mining techniques in this particular research area focused on roadway accidents. Since crash/accident record reports are mostly composed of unstructured data, the analysis primarily relies on a lot of expert experience and statistical analyses based on manual annotation and classification. These studies covered a broad array of focus areas such as wrong-way driving (WWD) crashes, secondary crashes, fatal crashes, traffic crash severities, and cyclist crashes. In most cases, incident reports and social media data were collected and analyzed to identify the root causes of different types of crashes and accidents. It was observed that, compared to classical methods of accident analysis considering only classified accidents, text mining techniques could provide better quality results (i.e., more crash/accident class identification) fast [35]. Studies were also conducted to enhance the accuracy of many of the text mining techniques for crash/accident analysis. For example, in [36], a text mining-based accident causal classification method was adopted based on a relational graph convolutional network (R-GCN) and the pre-trained bidirectional encoder representations from transformers (BERT) model to reduce the computational cost of text processing, thus exceeding the performance of existing methods and resulting in faster results. It was also identified that the novel BERT model had the best accuracy for identifying actual WWD crashes from potential WWD crashes in crash report narratives [37].

Many of these studies focused on identifying (1) *the key insights from the detailed crash narratives, and (2) the short-term and long-term implications* [38]. The major goal was to provide an understanding of the prevalent themes for crash/accident causation. Topic models were the primary text mining technique adopted for this purpose. For example, Structural Topic Modeling (STM) and Artificial Neural Network (ANN)-based techniques were used in [39] to gain insights from the unstructured textual descriptions of crowdsourced near-miss and collision events during cycling. In [40], STM was also used in conjunction with a network topology analysis to examine the frequency and interdependency/influence of topics from different crash narratives identified from fatal crashes. Researchers in [41] used LSA and LDA to identify the emergent themes that captured the key issues faced by the

vehicle owner. Probabilistic topic modeling was used in [42] to analyze automated vehicle crash narratives to identify safety concerns and research gaps. PLSA, LDA, sparse topical coding (STC), and fully sparse topic models (FSTM) were used in [43] to cluster visual words co-occurring in the same documents together into the same topic to automatically identify the occurrence of traffic accidents in complex scenes of traffic videos. The model of Local Interpretable Model-Agnostic Explanations (LIME) was combined with the Global Cross-Validation LIME (GCV-LIME) approach [44] to identify likely causal factors for injury severities.

Many context-specific key topics/concepts were identified from these studies. For example, it was identified that topics such as crossing the centerline, intoxication, and speeding were more prevalent for young drivers compared to topics such as turning left, failing to yield, and lane changing, which were more prevalent among older drivers [40]. Speed limit and speeding were found to be positively correlated with a crash resulting in a fatality [44]. Many text databases such as those of the Queensland Department of Transport and Main Roads, National Highway Traffic Safety Administration (NHTSA), and California Department of Motor Vehicles, and Alabama and Illinois statewide crash databases were used for this purpose. In essence, it was proven that text mining techniques, more specifically topic modeling techniques, could be used as an approach towards semi-automated encoding of qualitative data and generate useful insights.

In another stream of research, text mining techniques were primarily used to analyze the text and identify the features which were later used for the classification of different crash types. For example, using the feature extraction process, researchers in [30] transformed unstructured crash narratives into numerical features. Classification algorithms (e.g., logistic regression, random forest, NB, and SVM) were later applied to evaluate their uses in the identification of secondary crashes. Similarly, BERT models as well as five classification algorithms were used in [37] to classify crash report narratives as actual WWD and non-WWD crashes. Note that traditional text pre-processing mechanisms consisting of converting the document to a word format, removing punctuation, tokenizing, stemming, removing stop words, indexing, and vector transformation, etc., were conducted before applying the topic models. However, as these pre-processing mechanisms are common for any type of text analysis, they were not highlighted in this study.

#### 4.1.2. Rail

The use of text mining techniques in railway crashes/accidents has been limited compared to that in roadways. Topic models were also the primary text mining technique adopted by researchers in railway accident/crash narrative analysis. For example, a bi-level feature extraction-based text mining method was proposed in [45,46] that integrated features extracted on both the syntax and semantic levels with the aim of improving the fault classification performance in railway maintenance sectors. The Probabilistic Linear Discriminant Analysis (PLDA) topic model was used for feature selection on the semantic level from a railway signaling maintenance data set to reduce the data set into a low-dimensional topic space.

Partial least squares (PLS) and LDA-based text mining techniques were used in [29] to automatically discover accident characteristics/features (e.g., cross-level, milepost, worn, and gear) in the U.S. LDA was also used in [47] to identify major recurring accident topics from the text in the Federal Railroad Administration (FRA) reports. Apart from identifying key topics such as shoving accidents and hump yard accidents, these studies also identified the potential causes of the crashes. For example, it was identified that if factors related to switches and crossovers, etc., were not properly engineered for the operations being performed, it could result in a higher number of railway accidents/crashes. Although researchers in [48] did not apply topic models, they also investigated how to identify the relationship between railway crash characteristics. Using the Term Frequency (TF) and Term Frequency Inverse Document Frequency (TF-IDF) methods, frequent and characteristic words related to the highway-rail crashes in each state, such as highway, pavement, trailer,



descend, gate, and motorists, were identified. Moreover, it was identified that there are significant similarities between the reasons that cause highway-rail crashes across states. Through these studies, it was observed that text mining techniques could be successfully used to establish the associations between fault terms and fault classes, resulting in the improved precision of the fault diagnosis process.

#### 4.1.3. Other Sectors

Apart from roadway and rail crashes, there are a handful of studies that have applied text mining techniques for crash/accident investigation in other sectors. These sectors primarily include aviation, maritime, and construction. For example, STM was used by researchers in [24] to (1) *evaluate the usefulness of this method for identifying safety issues in Aviation* and (2) *find previously unreported connections or themes in incident reports*. Topics such as human factors, airspace, surface, routing, smoke, and fire were identified alongside the most significant (topic and phase of flight) relationships. The NB classifier, SVM linear kernel classifier, and SVM Gaussian kernel classifier were used in [49] for classifying the maritime crash data identified via the BoW text mining method. Researchers in [50] developed a novel framework that combined LDA and CNN algorithms to analyze hazard records on construction sites automatically. Finally, a keyword-based classifier (i.e., a unigram + bigram noisy-OR classifier) was developed in [51] that could apply text mining techniques to quickly find missed work zone (WZ) crashes from text datasets. Although the contexts are different, it was conclusively identified that the text mining techniques are a suitable approach for identifying and classifying hazard topics/subcategories. It was also emphasized that these techniques require less training time when applied in conjunction with other techniques (e.g., classification), are computationally efficient, and are easier to implement.

#### 4.2. Mobility Analysis

Another significant set of literature focused on the application of text mining techniques in investigating the mobility behavior/pattern of people/vehicles. The identified studies were broadly categorized into four research areas that included *social media-based sentiment/conception analysis, tourism, travel and traffic behavior and pattern analysis, and trend analysis*. These four areas will be discussed in detail in the following subsections. Note that one study may fall under multiple categories. For example, a study may discuss both tourism and conduct sentiment analysis. In such a scenario, a consensus was reached within the research team regarding the most suitable category of a particular study.

##### 4.2.1. Social Media-Based Sentiment/Conception Analysis

There have been numerous studies that have focused on the research domain of social media-based sentiment/conception analysis. Researchers in [52,53] discussed the potential of text mining techniques for transport data collection and to study the daily human activities/locations using social media data. In a separate study, social media data were also analyzed to develop a hierarchical approach for categorizing transport-related information. The primary goal was to demonstrate that social media data have the potential to enhance and deliver transport policy goals [54].

In [55], researchers proposed an extension of the typical text mining techniques in social media-based sentiment analysis by developing fuzzy ontology-based semantic knowledge with the Word2vec model to enhance the task of transportation feature extraction (i.e., discovering the most relevant features in social media text) and sentiment classification. Other studies focused on generating insights into how users living in a region felt about available transportation services. For example, LDA was used in [56] to obtain the most significant topics of traffic information posted by people. Topics based on regular monitoring of traffic conditions, persuasion to obey traffic rules and perform safe riding, and announcements were identified as most significant. In [55,57], researchers developed an ontology and LDA (OLDA)-based topic modeling and word embedding approach for

sentiment classification to examine traffic control and management systems using social network platforms.

Using sentiment analysis and other text mining techniques such as KoNLPy and Open Korean Text, public perceptions of bike sharing were classified in [58]. These techniques were also used to identify the different needs and desires of citizens living in the region. Although this study was not directly related to traditional mobility-focused studies, it was included due to the fact that it potentially revealed attitude changes in people's mobility. Regardless of the type of mobility analysis, in all cases, text mining techniques were observed to efficiently classify extremely ambiguous text and polarity.

#### 4.2.2. Tourism

Studies focusing on tourism primarily used text mining techniques to identify the positive and negative factors and their potential impact on tourism and transport needs. These studies highlighted the potential of text mining techniques to enhance destination tourism and hospitality services without undertaking time consuming research methods such as surveys, and interviews. For example, sentiment analysis and co-occurrence analysis were used in [59] to investigate visitors' perceptions of destination services through a hybrid analysis of travelers' review data.

Topic models using LDA and Word2Vec were used in [60] used to identify the best tour route for foreign tourists in South Korea based on the real reviews of customers and analysis. The goal was to recommend travel routes for new travelers. Finally, in [61], researchers conducted an investigation to determine insights into transportation considering geographical perspectives by identifying different modes of transportation in TripAdvisor comments. The WordNet lexical database and Suggested Upper Merged Ontology (SUMO) were used as text mining techniques to identify the positive and negative factors and their potential impact on tourism and transport needs. The WordNet lexical database relates hyponyms/hypernyms with sets of synonyms, leading to the development of semantic relationships between conceptual categories. SUMO can then be mapped onto the WordNet lexical database. As SUMO is composed of roughly 1000 terms and almost 4000 formal statements about those terms, it can help classify the concepts into categories [62]. In essence, these studies emphasized the need to consider geographical location, mode availability, etc., to develop a culture of sustainable tourism and transport.

#### 4.2.3. Travel and Traffic Behavior and Pattern Analysis

Two streams of research were identified that applied text mining techniques to travel and traffic behavior and pattern analysis problems. One stream focused on analyzing the problem on a micro scale, i.e., analyzing individual human travel behavior. For example, in [63], researchers proposed an unsupervised trajectory topic model (which was LDA-based) to identify latent driving patterns and to analyze drivers' main traveling intentions using commercial vehicle data. The Modified Latent Dirichlet Allocation (mLDA) model and modified Hierarchical Latent Dirichlet Allocation (mHLDA) model, two variations of the traditional LDA model, were used in [64] to quantitatively extract and recognize different driving styles with a hidden structure from real-world driving behavior data. The underlying structure of the driving styles was also evaluated using the models. Researchers in [23] also proposed LDA-based topic model analysis to quantify human mobility using the travel displacement and time for each trip identified via a GPS-based large-scale dataset of taxi trips. The results showed that the mobility pattern followed a power law for macroscopic characteristics (i.e., the geographic displacement of trips). The corresponding travel time followed a mixture of exponential and power law distribution. On a similar note, a LDA-based probabilistic topic model was used in [65] to discover latent activity on the individual level based on spatiotemporal data by combining continuous time variables and discrete location variables. The goal was to enhance the scope of human mobility data with representative and interpretable activity patterns that did not rely on predefined activity categories or simple rules. All these studies that analyzed the problem on a micro

scale focused primarily on the identification of (1) *specific traveling intentions* and (2) *on understanding and predicting the behavior patterns of traveling events*.

On a macro level, few studies are available in the literature that analyze the traffic patterns in a region using text mining techniques. For example, researchers in [66] used LDA-based topic modeling to find unusual traffic patterns by converting the traffic data recorded by speed cameras to words and documents. These topics were further studied geographically based on the location of the speed cameras. Researchers in [43] proposed a novel method for traffic density forecasting using low-level features and applying LDA-based topic models. The results showed that the framework could accurately estimate the density of traffic videos in both good and bad illumination conditions. Semantic transformation and LDA were employed in [67] to explore hidden ship mobility patterns from trajectories using a LDA trajectory topic probability distribution and topic-movement word probability distribution. It was observed that these results could be transformed to and visualized as pieces of evidence to make better inferences and interpretations.

#### 4.2.4. Trend Analysis

Text mining was also applied to elicit trends primarily via detecting the context of the keywords identified through different text mining techniques. In this research, we define trend as the *“clarification of concepts to identify and predict/project future potentials”*. For example, researchers in [31] applied simple text mining techniques to descriptions of smart mobility by extracting the wordings used in smart mobility descriptions and analyzing the frequency of each word. The goal was to identify the varying viewpoints in mobility and create different subtopics that would give smart mobility its true meaning and functional capabilities (e.g., individual digitization, partial integration, and full integration). A hybrid regression text mining approach was proposed in [68] to evaluate users' perceptions of several micro-mobility devices such as electric scooters, dockless bikes, and docked bikes. The goal was to identify the potential user base and the perceptions of electric scooters, dockless bikes, and docked bikes so that a future trend could be better capitalized on. By analyzing Dutch news articles and initiatives' websites, researchers in [69] used linguistic categorization-based text mining to investigate the ambiguity of the concept of smart mobility and future trends.

Researchers in [6] also attempted to identify key issues facing mobility services by using a text mining technique (i.e., the term frequency-inverse document frequency algorithm). Based on the observed trend, different short term and long term policies were also proposed such as plans and recommendations regarding self-driving public transportation systems in rural areas, interregional (capital area) and shared transportation services, data platform development, cooperation needs, the integration of public transportation, and the self-sustaining mobility ecosystem. Finally, the TF-IDF and association rule mining algorithms were proposed in [70] to identify and distil the risk factors in China's inland waterborne transportation, explore their interrelationships, and develop recommendations for improvement. This was essentially an improvement compared to the traditional quantitative risk assessment model. The integrated application of TF-IDF and association rule mining served to avoid uncertainty and subjectivity, and achieve good results proving their scientific nature as feasible methods in water transportation risk research.

#### 4.3. Supply Chain and Logistics

Text mining techniques have been extensively applied to different problems in the supply chain and logistics domain [71]. Primarily, text mining techniques were applied for information scanning, extraction, and the retrieval of global logistics-related measures. To identify the feasibility of text mining techniques in the supply chain and logistics domain, researchers in [72] demonstrated that word cloud, sentiment analysis, LDA, correspondence analysis, and multidimensional scaling could be successfully integrated to automatically analyze large chunks of textual data and to extract relevant insights into logistics and supply chain management. Studies that used different text mining techniques to solve

supply chain and logistics problems could be broadly categorized into three major themes: *sentiment/perception analysis, trend analysis, and risk and resilience analysis*.

#### 4.3.1. Sentiment/Perception Analysis

The majority of the studies in this domain primarily used social media data to generate sentiment/perception-based insights into different types of supply chain/logistics structures and problems. For example, in [73], the potential of Twitter data in supply chain contexts was analyzed. Sentiment analysis and opinion analysis were used in [13] to analyze peoples' emotions towards different smartphone brands (e.g., Apple, Samsung, and Huawei). This study also predicted how tweets could influence a smartphone company's supply chain and its management. The LDA model was first used to identify the topics that affect customer satisfaction in cold chain logistics using customer-generated reviews [32]. Later, the bi-directional long short-term memory (Bi-LSTM) model was implemented to calculate the sentiment score of the topics involved. Researchers in [74] developed an SVM-based pre-processing and text mining technique to investigate the positive and negative sentiments of tweets related to the food industry with the goal of developing a consumer-centric supply chain. Different factors such as speed, quality, error handling, and logistics information were identified, of which speed, price, transportation, and product quality were observed to significantly affect a customer's positive sentiment to a higher degree. Industry-specific factors were also identified. For example, color, food safety, smell, flavor, promotions, deals, particular combinations of food and drinks, and the presence of foreign particles were identified as key to sentiment development among consumers in a food supply chain.

E-commerce platforms, home delivery service providers, online advertisements and news, online user-generated reviews, comments, discussions, suggestions, and ideas were also used to analyze the perception of users. For example, researchers in [75] used a CNN-based text mining model to analyze online reviews of fresh e-commerce logistics services and understand consumers' shopping experiences with regard to the logistics service quality of fresh e-commerce enterprises. Researchers in [5] applied a Kansei engineering (KE)-based approach to design a cross-border logistics service (CBLS) via analyzing online contents. Text mining techniques (e.g., filtering by part-of-speech (POS) and detecting by n-gram probability analysis) were applied. This study offered a procedure for service design which is able to be applied in various service industries.

Instead of using social media or online data, in [76], word frequency analysis and sentiment analysis were used to analyze text generated from general newspapers as well as supply chain and logistics newspaper articles. The goal was to identify topics of interest and the perception in the media of supply chain management constructs. The results showed that the concepts of risk, resilience, and especially sustainability could vary in their news coverage over different time frames.

#### 4.3.2. Trend Analysis

Using the same definition of "Trend Analysis" as identified in the previous section, four studies were selected. For example, researchers in [77] reviewed 52 award-gaining green logistics initiatives/practices and used text mining techniques to explore the cooccurring links within words and to present collaborative green initiatives. Using Twitter data from 100 NASDAQ firms, researchers in [78] developed unigram-, bigram-, and trigrams-based simple text analysis to understand the direction that global supply chains may evolve in and the possible solutions that could solve many new and recurring supply chain challenges. Furthermore, in [79], Leximancer for text analysis and dictionary-based text mining programs DICTION and SPSS for rhetorical analysis were adopted to explore sustainable supply chain management (SSCM) trends, and firms' strategic positioning and execution with regard to sustainability in the textile and apparel industry based on news articles and sustainability reports. Finally, LDA was used for patent analysis in [80] to explore the technological trends in logistics. The topics identified such as shipping services

and container product information systems were further investigated regarding trends in patenting activity and major assignees for each topic. These studies reflected the expansion and convergence of trends in the supply chain and logistics field, and identified challenges and the potential opportunities to be utilized to make supply chain and logistics systems more efficient.

#### 4.3.3. Risk and Resilience Analysis

Three studies were identified that focused on the application of text mining techniques in risk management and developed data-driven approaches related to supply chain resilience analysis. Researchers in [81,82] applied TF and TF-IDF analysis to identify regional risk factors that should be taken into account while improving or maintaining global supply operations using the current supply chain literature. A total of seven regional risk factors (including political risk, logistic risk, and demand risk) and 81 generic risk factor terms were extracted and organized systematically to provide insights into supply chain policymakers. Using nonnumerical unstructured data indexing, searching, and theorization-based text mining and social media analysis, five scientific research databases were analyzed in [83] to understand future pathways that could address disruptive events such as COVID-19. This study highlighted the need for the government to offer support to expand the research capabilities that align with public concerns over supply chain resilience while bridging the gap between theory and practice.

#### 4.4. Construction and Urban Infrastructure

Construction and urban infrastructure refer to any development/rehabilitation activities associated with different civil infrastructures. Text mining techniques have been widely adopted to analyze different large text databases in this domain as well. The studies have been categorized into four parts: *new information generation, sentiment/perception analysis, trend analysis, and others.*

##### 4.4.1. New Information Generation

One of the uses of text mining in the construction and urban infrastructure sector was the identification of relevant but new information from large text databases. For example, in [84], a semantic text-pairing method (the Doc2Vec model) was proposed to identify relevant provisions from different construction specifications considering the textual properties. A total of 2527 provisions were prepared from two construction specifications of highway projects with five national standards from three countries (Australia, the United Kingdom, and the United States). The Author-Topic Model (ATM) text mining approach was adopted in [85] to identify the major aspects of bridge management (BM) from policy documents identified via all online database searches. The 12 revealed topics in the policies reflected the 12 aspects of BM, and corresponded to indirect, proactive, and remedial BM measures. Researchers in [86] used LDA to analyze the Chinese Government's replies to environment impact assessment reports of highway projects. The results showed that the government emphasized the comprehensive environmental effects more and more. Finally, using association rule mining techniques for a large-scale transportation data set, researchers in [87] analyzed the associations among frequent words in the type of work and the type of lane closure therein in expressway construction work zone areas. It was found that recurrent everyday tasks and bridge repair works tended to cause shoulder lane closure, while other works—such as tunnel repair, night work, pavement, median barrier, road surface repair, and line marking—were more associated with main lane closure.

##### 4.4.2. Sentiment/Perception Analysis

To understand user experience of and satisfaction with urban infrastructure, researchers in [88] used semantic network analysis to recognize the relationships between the extracted keywords from 2945 bridge complaint data records and 404 tunnel complaint data records. Twitter data was used in [89] to analyze the different inconveniences experienced

by users due to the significant escalation in noise and dust pollution during construction. The study also analyzed how the general public reacts to such inconveniences, and how they interact. LDA and VADER sentiment analysis were primarily used for this purpose.

#### 4.4.3. Trend Analysis

Based on the TF-IDF algorithm, researchers in [90] analyzed 291 characteristics of the competition culture of highway construction enterprises and its evolution pattern. The results showed that the competition culture of enterprises demonstrated a trend of moving away from catering to the market to the internal construction of enterprises. The internal construction of enterprises could be reflected via talent competition, technological innovation, and the optimization of management.

#### 4.4.4. Others

One study was found that could not be categorized into any of the preceding categories. The study in [91] demonstrated how text descriptions of a construction project's characteristics combined with numerical data could lead to a predictive model for a competitively bid construction project's expected cost overrun. The authors used the "Generate n-gram terms" feature method and TF-IDF formulation to analyze the text, the results of which were later used in different classification algorithms such as SVM. The results in general showed that there were words and word pairs that could be associated with different levels of cost overruns.

### 4.5. Review of Literature

This set of studies used text mining techniques to systematically analyze the literature on a particular research domain with the goal of providing a comprehensive assessment of the current state-of-the-art. Three major domains were identified that included *Industry 4.0 applications regarding the supply chain, risk management, and others*.

#### 4.5.1. Industry 4.0 Applications Regarding the Supply Chain

Researchers in [92] used a non-negative matrix factorization (NMF)-based topic model to identify the trends, advances, and gaps in Industry 4.0 applications regarding the supply chain. Furthermore, in [93], researchers used LDA to identify the existing knowledge on digital transformation due to Industry 4.0 in supply chain process management. Finally, TF-IDF and threshold analysis were used in [94] to generate concept maps of important concepts and calculate similarity coefficients regarding the analysis of IoT technologies' and services' evolution. Academic journals, market reports, and patents were collected and reviewed for this analysis. Topics such as sustainable supply chain management, circular economy, and Industry 4.0 technologies were identified as the most significant topics at the intersection of Industry 4.0 and supply chain process management. Big data analytics, Blockchain, IoT, AI, and ML were also identified as critical technologies that can modernize supply chain management. These topics and the related terms provided key insights into the integration of smart manufacturing systems and production systems, and the impacts of Industry 4.0 on industries, supply chains and logistics, and business processes.

#### 4.5.2. Risk Management

Two studies were identified that analyzed the literature to identify supply chain/logistics risks. Researchers in [95] used simple word frequency statistics to classify risk factors on different cold supply chain levels. Corresponding risk control strategies for supply chain member risk, system environment risk, coordinating role risk, and structure risk were also developed. Finally, text parsing, text transformation, text filtering, text mining, and visualization were used in [96] to review studies on the supply chain risk management and apparel industry. This study highlighted the potential of news and other available data to be transformed into valuable tools for proactively minimizing supply chain risks.

#### 4.5.3. Others

Six other studies were identified that used text mining techniques for a comprehensive literature review. As these studies fell within six distinct domains, they were all categorized into the “Others” category. Using term frequency (TF) analysis and TF-IDF analysis in a time series manner (e.g., in 2005–2009, 2010–2014, and 2015–2019), researchers in [97] identified nine topics (including preparation, recovery, emergent response, supplying, and sourcing) related to humanitarian relief logistics (HRL). This study provided a pathway to understand the research trends, areas of improvement, and a few future research directions in HRL. In [98], key journal articles in the field of maritime studies were reviewed using LDA to create a new intellectual structure for sustainability research. The underlying themes were identified, key trends and patterns were extracted, and future research development trajectories were mapped for the field of maritime studies. LDA was again used in [99] to collect freight transportation and freight system-related studies to identify trends. A total of 20 main topics were derived that included rail network, maritime transportation, freight transportation planning, vehicle size and weight, and emission and fuel consumption. It was also observed that due to the economy, politics, and the environment, these topics were highly sensitive to different exogenous and endogenous socioeconomic variables.

Researchers in [100] used relative frequency of words and correlation analysis to identify and quantify trends and innovations in supply chain and logistics and whether or not they accelerated during the COVID-19 pandemic. The results showed that innovations in supply chain and logistics were accelerating due to the emerging focus on blockchain, internet of things, data, drones, robots, and unmanned vehicles during the COVID-19 pandemic. In [101], researchers used the Voyant tool to identify trends related to non-pavement research by analyzing surveys of accelerated pavement testing (APT) applications in non-pavement research. Essentially, the number of times a word was mentioned in the text dataset and the type of assets and infrastructure used in testing were identified. Finally, IDF for concept extraction was used in [102] to summarize trends and some important points relating to airline optimization. The goal of the study was to give an idea of how the aviation sector shaped academic studies, how studies on aviation optimization could contribute in the future, and how different countries addressed important challenges in the aviation industry in the past.

#### 4.6. Innovation in Transportation Infrastructure Research

Several studies were identified that applied text mining techniques to novel areas. The criteria for being novel was fulfilled by only one study. Note that the category “Others” in the preceding subsections were not considered to be novel as they could be subcategorized as part of other categories (e.g., “Others” could be subcategorized under the “Review of literature” category).

A total of five novel problems were identified. For example, in [103], LDA was applied to determine the key topics relevant to Unmanned Aircraft Systems (UAS) sighting incidents. The goal was to help researchers understand the areas of risk and develop practices to avoid UAS-related incidents and make the airspace safer for everyone. Researchers in [104] analyzed 10,595 complaints against two major airlines using LDA to understand the main reasons for complaints being triggered during the heights of COVID-19. Although the complaint topics modeled based on the research results were generally similar, differences in complaint content were found between the two airline companies. It was also observed that during the early stages of the pandemic, passengers either ignored or did not have any negative experiences associated with issues such as cabin staff, cabin cleanliness, meals, and entertainment on both airlines. On a similar note, in [105], 103,428 Google Maps reviews of 64 major hub airports in the US were analyzed to identify representative topics of passenger concerns in airports during COVID-19 using Collapsed Gibbs Sampling-based Dirichlet Multinomial Mixture (GSDMM)-based clustering.

LSA was used in [106] to yield multiple groups of contextually similar terms from a wide range of players, including both potential users and experts from various fields, regarding future drone technologies. Essentially, this study generated future scenarios which indicated emerging technologies’ early warning signs of potential social impacts and their specific consequences to society. Researchers in [107] provided structured evidence concerning sustainability disclosure content in the container shipping industry using LDA and by computing a coherence score. The latent information of major listed container shipping companies’ sustainability reports was used for analysis. This research unveiled the specific latent sustainability disclosure framework of container shipping companies, providing them with future references for drafting sustainability reports and helping them to explore sustainability disclosure more comprehensively to pursue fruitful stakeholder engagement in the container shipping industry. Finally, in [108], a text mining technique called edit distance (ED) was employed for matching imperfectly read large truck plates. Essentially, the similarities or dissimilarities between two strings in a plate were measured and matched against one another. The results showed high-performance accuracy in reading plates using the ED approach. Table 1 summarizes the major findings of this study across different research areas.

**Table 1.** Summary of key literature reviewed in this study.

Research Domain	Focus Area	Authors	Major Text Mining Techniques	Major Objective
Crashes and accidents	Roadways	[37]	BERT	Classify crash report narratives
		[30]	GCV-LIME	Identify likely causal factors for injury severities
		[39]	STM	Estimate the cyclist’s tendency to collide
		[40]	STM	Examine the influence of different crash-causing topics on each other
		[38]	LDA	Identify a potential of a dataset in crash analysis and generating insights into the dataset’s capabilities
		[35]	LDA	Compare classical methods of accident analysis
		[41]	LSA and LDA	Identify the emergent themes that captured the key issues faced by a vehicle owner
		[42]	Probabilistic topic modeling	Identify safety concerns regarding automated vehicle crashes
	[36]	BERT	Reduce the computational cost of text processing in crash investigation	
	[43]	PLSA, LDA, STC, and FSTM	Identify the occurrence of traffic accidents in traffic videos	
	[45,46]	PLDA	Improve the fault classification performance in railway maintenance sectors	
	Rail	[48]	TF, TF-IDF	Identify the relationship between crash characteristics
	[29]	PLS, LDA	Discover accident characteristics/features and effects	
	[47]	LDA	Identify major recurring accident topics	
	Others	[24]	STM	Find previously unreported connections or themes
[50]	LDA	To analyze hazard records of construction sites automatically		



Table 1. Cont.

Research Domain	Focus Area	Authors	Major Text Mining Techniques	Major Objective
Mobility Analysis	Social media-based sentiment/conception analysis	[57]	OLDA	Sentiment classification to examine traffic control and management systems
		[58]	KoNLPy	Classify public perceptions of bike sharing
	Tourism	[59]	Sentiment, co-occurrence analysis	Investigate visitors' perceptions of destination services
		[60]	LDA, Word2Vec	Identify the best tour route for foreign tourists
		[61]	SUMO ontology	Identify the positive and negative factors and their potential impact on tourism and transport needs
		[63]	LDA	Identify latent driving patterns
		[64]	MLDA, MHLDA	Quantitatively extract and recognize different driving styles
	Travel and traffic behavior and pattern analysis	[23]	LDA	Quantify human mobility using the travel displacement and time taken for each trip
		[65]	LDA	Discover latent activity on the individual level based on spatiotemporal data
		[66]	LDA	Find unusual traffic patterns
		[43]	LDA	Traffic density forecasting
		[67]	LDA	Explore hidden ship mobility patterns
		[68]	Hybrid, regression-text mining	Evaluate users' perceptions of several micro-mobility devices
		Trend analysis	[69]	Linguistic categorization-based text mining
[6]	TF, TF-IDF		Identify key issues facing mobility services	
[70]	TF, TF-IDF, association rule		Identify current and future risk factors	
Supply chain and logistics	Trend analysis	[79]	Leximancer, dictionary-based text mining program DICTION	Explore sustainable supply chain management trends, and firms' strategic positioning and execution
		[80]	LDA	Explore technological trends in logistics
	Sentiment/perception analysis	[76]	Sentiment Analysis	Identify topics of interest and the point of view of the media on supply chain management constructs
		[13]	LDA	Identify the topics that affect customer satisfaction in cold chain logistics
		[74]	SVM-based pre-processing and text mining	Investigate the positive and negative sentiments of tweets related to the food industry
	Risk and resilience analysis	[81,82]	TF, TF-IDF	Identify regional and generic risk factors

Table 1. Cont.

Construction and Urban Infrastructure		[84]	Doc2Vec model	Identify relevant provisions from different construction specifications
	New information generation	[85]	ATM	Identify the major aspects of bridge management (BM)
		[86]	LDA	Realize governmental concerns over environmental effects of highway construction
		[87]	Association rule	Analyze the associations among the type of work and lane closure in expressway construction work zone areas
	Sentiment/perception analysis	[89]	LDA and VADER	Analyze the users' inconveniences due to noise and dust pollution during construction
	Trend analysis	[90]	TF-IDF	Analyze the characteristics of the competition culture of highway construction enterprises and its evolution pattern
	Others	[91]	TF-IDF	Design a predictive model for a competitively bid construction projects' expected cost overrun.
Review of Literature	Industry 4.0 applications regarding the supply chain	[92]	NMF	Identify the trends, advances, and gaps in the Industry 4.0 applications
		[93]	LDA	Identify the existing knowledge on digital transformation due to Industry 4.0
		[94]	TF-IDF and threshold analysis	Generate concept maps of important concepts and calculate similarity coefficients
	Others	[97]	TF-IDF	Identify key topics in HRL
		[98]	LDA	Create a new intellectual structure for marine sustainability research
		[101]	Voyant	Identify trends related to non-pavement research
		[99]	LDA	Identify trends related to freight transportation and freight systems
Innovation in Transportation Infrastructure Research	Unmanned Aircraft Systems (UAS)	[103]	LDA	Determine the key topics relevant to UAS sighting incidents
	Airline complaints	[104]	LDA	Understand the main reasons for complaints being triggered
	Airline passenger concerns	[105]	GSDMM	Identify representative topics of passenger concerns in airports
	Drone technology	[106]	LSA	Identify warning signs of potential social impacts and their specific consequences
	Container shipping	[107]	LDA	Provide structured evidence of sustainability disclosure content in the container shipping industry
	Truck plate identification	[108]	ED	Read large truck plates

## 5. Conclusions and Future Research Directions

In this study, the current state-of-the-art research available on the application of text mining techniques in the transportation domain was reviewed and analyzed. A Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA)-based

structured methodology was used to identify relevant studies that used different text mining techniques across different transportation infrastructure-related problems or issues. A total of 82 studies were initially collected. A total of 59 studies from both the US and other parts of the world (e.g., China and Bangladesh) were ultimately selected for review after rigorous quality checks. As a result of examining the studies, it was confirmed that text mining techniques had been widely applied across different types of transportation infrastructures and the related problems. Different types of transportation infrastructures such as construction, supply chain and logistics, and traffic infrastructure were observed to be significantly benefitting from the capability of text mining techniques. It was also observed that various types of problems, such as crash and accident analysis, traffic and travel pattern analysis, and perception/sentiment analysis, could be investigated using the capabilities of text mining techniques.

Another key observation was that different types of data sources can be successfully pre-processed and used for text mining to get useful insights. Many databases have been successfully analyzed using text mining techniques, such as those of the Queensland Department of Transport and Main Roads and National Highway Traffic Safety Administration (NHTSA), the California Department of Motor Vehicles database, railway signaling maintenance dataset, social media data, online reviews, and the Alabama and Illinois statewide crash database. Although these datasets had varying levels of granularity, through careful pre-processing and method development, text mining techniques were successful in identifying key insights and providing significant policy recommendations.

Traditional text pre-processing mechanisms such as converting the document into a word format, removing punctuations, tokenizing, stemming, removing stop words, indexing, and vector transformation, etc., were implemented. All of these mechanisms were not applied together in a single study. Instead, depending on the problem context and research need, one of them was applied or a few of them were applied concurrently or sequentially. Topic models and sentiment analysis were the two major text mining techniques adopted by researchers in the transportation infrastructure domain. This makes intuitive sense as topic models are arguably the most convenient to apply due to their ability to identify significant topics and trends with widely available computational capabilities. Similarly, to identify peoples' perceptions, especially using text data from social media, sentiment analysis is comparatively the most feasible alternative and a less computationally challenging alternative. Note that variations of traditional topic models (e.g., PLDA) and sentiment analysis were also developed and used by researchers depending on the research goal.

The contributions of this study are twofold. From a theoretical perspective, this study provides a structured methodology for identifying key research/studies that link text mining with the transportation infrastructure sector. The methodology can be easily extended to identify other similar studies linking the transportation infrastructure sector with other mechanisms. Regarding the practical contributions, this study provides a comprehensive overview of how researchers have used text mining techniques in analyzing different problems within the transportation infrastructure sector and the implications of these research endeavors. Moreover, readers can also identify the reasons behind the applications of different text mining techniques or their modifications in particular research contexts. This research can also act as a springboard for future research that can explore largely untapped areas of transportation infrastructure sectors (e.g., construction safety and drone routing) and investigate how text mining techniques can help generate new insights.

This study has a few limitations. First, although studies from outside the US were sought to be included, in the end, only a small sample of international studies were included. Increasing the sample size will enhance the potential for a more comprehensive evaluation of the current state-of-the-art research in this field. Second, it may be possible to explore the reasonings behind different researchers' choices of different text mining techniques/models. The current study did not go into a detailed analysis of "*why the technique/model was chosen*", instead focusing on "*what was chosen*". A comparative assessment of different models

will be conducted in future studies where the authors will investigate in detail (1) the characteristics of the data, (2) the model setup, (3) and the context of the research to identify the applicability and efficiency of different models under different circumstances. A cross-sector analysis, i.e., of whether or not models developed/adopted to analyze one type of problem in a sector may be applicable to solve problems in other sectors will also be conducted in the future.

**Author Contributions:** Conceptualization, methodology, data collection, data analysis, and writing: S.C.; writing: A.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Alexakis, G.; Panagiotakis, S.; Fragkakis, A.; Markakis, E.; Vassilakis, K. Control of Smart Home Operations Using Natural Language Processing, Voice Recognition and IoT Technologies in a Multi-Tier Architecture. *Designs* **2019**, *3*, 32. [\[CrossRef\]](#)
- Chopra, A.; Prashar, A.; Sain, C. Natural Language Processing. *Int. J. Technol. Enhanc. Emerg. Eng. Res.* **2013**, *1*, 131–134.
- Goldberg, Y. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420. [\[CrossRef\]](#)
- Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science* **2015**, *349*, 261–266. [\[CrossRef\]](#)
- Hsiao, Y.-H.; Chen, M.-C.; Liao, W.-C. Logistics service design for cross-border E-commerce using Kansei engineering with text-mining-based online content analysis. *Telemat. Inform.* **2017**, *34*, 284–302. [\[CrossRef\]](#)
- Seo, Y.; Lim, D.; Son, W.; Kwon, Y.; Kim, J.; Kim, H. Deriving Mobility Service Policy Issues Based on Text Mining: A Case Study of Gyeonggi Province in South Korea. *Sustainability* **2020**, *12*, 10482. [\[CrossRef\]](#)
- Kamerkar, N.; Patil, K.; Kale, A. Text Mining Applied to Rail Accidents. *Int. J. Future Revolut. Comput. Sci. Commun. Eng.* **2018**, *4*, 383–386.
- VijayGaikwad, S.; Chaugule, A.; Patil, P. Text Mining Methods and Techniques. *Int. J. Comput. Appl.* **2014**, *85*, 42–45. [\[CrossRef\]](#)
- Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* **2013**, *46*, 200–211. [\[CrossRef\]](#)
- Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y.; Ngo, D.C.L. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* **2014**, *41*, 7653–7670. [\[CrossRef\]](#)
- Rojas, C.V.; Reyes, E.R.; Hernández, F.A.Y.; Robles, G.C. Integration of a text mining approach in the strategic planning process of small and medium-sized enterprises. *Ind. Manag. Data Syst.* **2018**, *118*, 745–764. [\[CrossRef\]](#)
- Kim, Y.; Dwivedi, R.; Zhang, J.; Jeong, S. Competitive Intelligence in Social Media Twitter: iPhone 6 Vs. Galaxy S5. *Online Inf. Rev.* **2016**, *40*, 42–61. [\[CrossRef\]](#)
- Akundi, A.; Tseng, B.; Wu, J.; Smith, E.; Subbalakshmi, M.; Aguirre, F. Text Mining to Understand the Influence of Social Media Applications on Smartphone Supply Chain. *Procedia Comput. Sci.* **2018**, *140*, 87–94. [\[CrossRef\]](#)
- Leung, X.Y.; Sun, J.; Bai, B. Bibliometrics of social media research: A co-citation and co-word analysis. *Int. J. Hosp. Manag.* **2017**, *66*, 35–45. [\[CrossRef\]](#)
- Petrova, M.; Sutcliffe, P.; Fulford, K.W.M.; Dale, J. Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: A word frequency analysis study. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 479–488. [\[CrossRef\]](#)
- Mahgoub, H.; Rosner, D.; Ismail, N.; Torkey, F. A Text Mining Technique Using Association Rules Extraction. *Int. J. Comput. Inf. Eng.* **2008**, *2*, 2044–2051.
- Janetzko, D.; Cherfi, H.; Kennke, R.; Napoli, A.; Toussaint, Y. Knowledge-based Selection of Association Rules for Text Mining. In Proceedings of the 16h European Conference on Artificial Intelligence—ECAI'04, ECCAI, Valencia, Spain, 22–27 August 2004; pp. 485–489.
- Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent Convolutional Neural Networks for Text Classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2267–2273.
- Wang, Z.; Qu, Z. Research on Web text classification algorithm based on improved CNN and SVM. In Proceedings of the 2017 IEEE 17th International Conference on Communication Technology (ICCT), Chengdu, China, 27–30 October 2017; pp. 1958–1961. [\[CrossRef\]](#)
- Xu, S. Bayesian Naïve Bayes classifiers to text classification. *J. Inf. Sci.* **2016**, *44*, 48–59. [\[CrossRef\]](#)
- Ahmed, M.H.; Tiun, S.; Omar, N.; Sani, N.S. Short Text Clustering Algorithms, Application and Challenges: A Survey. *Appl. Sci.* **2022**, *13*, 342. [\[CrossRef\]](#)

22. Rong, Y.; Liu, Y. Staged Text Clustering Algorithm Based on K-means and Hierarchical Agglomeration Clustering. In Proceedings of the IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 124–127.
23. Xiong, H.; Xie, K.; Ma, L.; Yuan, F.; Shen, R. Exploring the Citywide Human Mobility Patterns of Taxi Trips through a Topic-Modeling Analysis. *J. Adv. Transp.* **2021**, *2021*, 6697827. [[CrossRef](#)]
24. Kuhn, K.D. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 105–122. [[CrossRef](#)]
25. Sachin, S.; Tripathi, A.; Mahajan, N.; Aggarwal, S.; Nagrath, P. Sentiment Analysis Using Gated Recurrent Neural Networks. *SN Comput. Sci.* **2020**, *1*, 74. [[CrossRef](#)]
26. Institute of Electrical and Electronics Engineers. Real Time Sentiment Analysis of Tweets Using Naive Bayes. In Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 14–16 October 2016; pp. 257–261.
27. Zainuddin, N.; Selamat, A. Sentiment analysis using Support Vector Machine. In Proceedings of the 2014 International Conference on Computer, Communications, and Control Technology (I4CT), Langkawi, Malaysia, 2–4 September 2014; pp. 333–337. [[CrossRef](#)]
28. Chowdhury, S.; Zhu, J. Investigation of Critical Factors for Future-Proofed Transportation Infrastructure Planning Using Topic Modeling and Association Rule Mining. *J. Comput. Civ. Eng.* **2023**, *37*, 04022044. [[CrossRef](#)]
29. Brown, D.E. Text Mining the Contributors to Rail Accidents. *IEEE Trans. Intell. Transp. Syst.* **2015**, *17*, 346–355. [[CrossRef](#)]
30. Zhang, X.; Green, E.; Chen, M.; Souleyrette, R.R. Identifying secondary crashes using text mining techniques. *J. Transp. Saf. Secur.* **2019**, *12*, 1338–1358. [[CrossRef](#)]
31. So, J.; An, H.; Lee, C. Defining Smart Mobility Service Levels via Text Mining. *Sustainability* **2020**, *12*, 9293. [[CrossRef](#)]
32. Lim, M.K.; Li, Y.; Song, X. Exploring customer satisfaction in cold chain logistics using a text mining approach. *Ind. Manag. Data Syst.* **2021**, *121*, 2426–2449. [[CrossRef](#)]
33. Peixoto, B.; Pinto, R.; Melo, M.; Cabral, L.; Bessa, M. Immersive Virtual Reality for Foreign Language Education: A PRISMA Systematic Review. *IEEE Access* **2021**, *9*, 48952–48962. [[CrossRef](#)]
34. Yusop, S.R.M.; Rasul, M.S.; Yasin, R.M.; Hashim, H.U.; Jalaludin, N.A. An Assessment Approaches and Learning Outcomes in Technical and Vocational Education: A Systematic Review Using PRISMA. *Sustainability* **2022**, *14*, 5225. [[CrossRef](#)]
35. Krause, S.; Busch, F. New Insights into Road Accident Analysis through the Use of Text Mining Methods. In Proceedings of the 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Cracow, Poland, 5–7 June 2019.
36. Chen, Z.; Huang, K.; Wu, L.; Zhong, Z.; Jiao, Z. Relational Graph Convolutional Network for Text-Mining-Based Accident Causal Classification. *Appl. Sci.* **2022**, *12*, 2482. [[CrossRef](#)]
37. Hosseini, P.; Khoshshirat, S.; Jalayer, M.; Das, S.; Zhou, H. Application of Text Mining Techniques to Identify Actual Wrong-Way Driving (WWD) Crashes in Police Reports. *Int. J. Transp. Sci. Technol.* **2022**. [[CrossRef](#)]
38. Das, S.; Dutta, A.; Tsapakis, I. Topic Models from Crash Narrative Reports of Motorcycle Crash Causation Study. *Transp. Res. Rec. J. Transp. Res. Board* **2021**, *2675*, 449–462. [[CrossRef](#)]
39. Kwayu, K.M.; Kwigizile, V.; Lee, K.; Oh, J.-S.; Nelson, T. Automatic topics extraction from crowdsourced cyclists near-miss and collision reports using text mining and Artificial Neural Networks. *Int. J. Transp. Sci. Technol.* **2022**, *11*, 767–779. [[CrossRef](#)]
40. Kwayu, K.M.; Kwigizile, V.; Lee, K.; Oh, J.-S. Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. *Accid. Anal. Prev.* **2020**, *150*, 105899. [[CrossRef](#)] [[PubMed](#)]
41. Mehrotra, S.; Roberts, S.; Identification and Validation of Themes from Vehicle Owner Complaints and Fatality Reports Using Text Analysis. In Transportation Research Board; 2018. Available online: <https://trid.trb.org/view/1496773> (accessed on 13 February 2023).
42. Alambeigi, H.; McDonald, A.D.; Author, C.; Tankasala, S.R. Crash Themes in Automated Vehicles: A Topic Modeling Analysis of the California Department of Motor Vehicles Automated Vehicle Crash Database. *arXiv* **2020**, arXiv:2001.11087.
43. Kaviani, R.; Ahmadi, P.; Gholampour, I. A new method for traffic density estimation based on topic model. In Proceedings of the 2015 Signal Processing and Intelligent Systems Conference (SPIS), Tehran, Iran, 16–17 December 2015; pp. 114–118. [[CrossRef](#)]
44. Arteaga, C.; Paz, A.; Park, J. Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach. *Saf. Sci.* **2020**, *132*, 104988. [[CrossRef](#)]
45. Wang, F.; Xu, T.; Tang, T.; Zhou, M.; Wang, H. Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 49–58. [[CrossRef](#)]
46. Zhao, Y.; Xu, T.H.; Hai-feng, W. Text Mining Based Fault Diagnosis of Vehicle On-board Equipment for High Speed Railway. In Proceedings of the IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 900–905.
47. Williams, T.P.; Betak, J.F. Identifying Themes in Railroad Equipment Accidents Using Text Mining and Text Visualization. In Proceedings of the 2016 International Conference on Transportation and Development, Houston, TX, USA, 26–29 June 2016.
48. Soleimani, S.; Mohammadi, A.; Chen, J.; Leitner, M. Mining the Highway-Rail Grade Crossing Crash Data: A Text Mining Approach. In Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1063–1068. [[CrossRef](#)]

49. Tirunagari, S. Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports. *arXiv* **2015**, arXiv:1507.02447.
50. Zhong, B.; Pan, X.; Love, P.E.; Sun, J.; Tao, C. Hazard analysis: A deep learning and text mining framework for accident prevention. *Adv. Eng. Inform.* **2020**, *46*, 101152. [[CrossRef](#)]
51. Sayed, A.; Qin, X.; Kate, R.J.; Anisuzzaman, D.; Yu, Z. Identification and analysis of misclassified work-zone crashes using text mining techniques. *Accid. Anal. Prev.* **2021**, *159*, 106211. [[CrossRef](#)]
52. Grant-Muller, S.M.; Gal-Tzur, A.; Minkov, E.; Nocera, S.; Kuflik, T.; Shoor, I. Enhancing transport data collection through social media sources: Methods, challenges and opportunities for textual data. *IET Intell. Transp. Syst.* **2015**, *9*, 407–417. [[CrossRef](#)]
53. Maghrebi, M.; Abbasi, A.; Rashidi, T.H.; Waller, S.T. Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 208–213. [[CrossRef](#)]
54. Gal-Tzur, A.; Grant-Muller, S.M.; Kuflik, T.; Minkov, E.; Nocera, S.; Shoor, I. The potential of social media in delivering transport policy goals. *Transp. Policy* **2014**, *32*, 115–123. [[CrossRef](#)]
55. Ali, F.; El-Sappagh, S.; Kwak, D. Fuzzy Ontology and LSTM-Based Text Mining: A Transportation Network Monitoring System for Assisting Travel. *Sensors* **2019**, *19*, 234. [[CrossRef](#)] [[PubMed](#)]
56. Hidayatullah, A.F.; Ma'Arif, M.R. Road traffic topic modeling on Twitter using latent dirichlet allocation. In Proceedings of the 2017 International Conference on Sustainable Information Engineering and Technology (SIET), Malang, Indonesia, 24–25 November 2017; pp. 47–52. [[CrossRef](#)]
57. Ali, F.; Kwak, D.; Khan, P.; El-Sappagh, S.; Ali, A.; Ullah, S.; Kim, K.H.; Kwak, K.-S. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl. Based Syst.* **2019**, *174*, 27–42. [[CrossRef](#)]
58. Kim, N.R.; Hong, S.G. Text mining for the evaluation of public services: The case of a public bike-sharing system. *Serv. Bus.* **2020**, *14*, 315–331. [[CrossRef](#)]
59. Kim, K.; Park, O.-J.; Yun, S.; Yun, H. What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technol. Forecast. Soc. Chang.* **2017**, *123*, 362–369. [[CrossRef](#)]
60. Park, S.-T.; Liu, C. A study on topic models using LDA and Word2Vec in travel route recommendation: Focus on convergence travel and tours reviews. *Pers. Ubiquitous Comput.* **2020**, *26*, 429–445. [[CrossRef](#)]
61. Serna, A.; Gasparovic, S. Transport analysis approach based on big data and text mining analysis from social media. *Transp. Res. Procedia* **2018**, *33*, 291–298. [[CrossRef](#)]
62. Niles, I.; Pease, A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Ike*; CSREA Press: Las Vegas, NV, USA, 2003; pp. 412–416.
63. Liao, L.; Wu, J.; Zou, F.; Pan, J.; Li, T. Trajectory Topic Modelling to Characterize Driving Behaviors With GPS-based Trajectory Data. *J. Internet Technol.* **2018**, *19*, 815–824. [[CrossRef](#)]
64. Qi, G.; Wu, J.; Zhou, Y.; Du, Y.; Jia, Y.; Hounsell, N.; Stanton, N.A. Recognizing driving styles based on topic models. *Transp. Res. Part D Transp. Environ.* **2018**, *66*, 13–22. [[CrossRef](#)]
65. Zhao, Z.; Koutsopoulos, H.N.; Zhao, J. Discovering latent activity patterns from transit smart card data: A spatiotemporal topic model. *Transp. Res. Part C Emerg. Technol.* **2020**, *116*, 102627. [[CrossRef](#)]
66. Gholampour, I.; Mirzahosseini, H.; Chiu, Y.-C. Traffic pattern detection using topic modeling for speed cameras based on big data abstraction. *Transp. Lett.* **2020**, *14*, 339–346. [[CrossRef](#)]
67. Huang, L.; Wen, Y.; Guo, W.; Zhu, X.; Zhou, C.; Zhang, F.; Zhu, M. Mobility pattern analysis of ship trajectories based on semantic transformation and topic model. *Ocean Eng.* **2020**, *201*, 107092. [[CrossRef](#)]
68. Kutela, B.; Novat, N.; Adanu, E.K.; Kidando, E.; Langa, N. Analysis of residents' stated preferences of shared micro-mobility devices using regression-text mining approach. *Transp. Plan. Technol.* **2022**, *45*, 159–178. [[CrossRef](#)]
69. Manders, T.; Klaassen, E. Unpacking the Smart Mobility Concept in the Dutch Context Based on a Text Mining Approach. *Sustainability* **2019**, *11*, 6583. [[CrossRef](#)]
70. Wang, Z.; Yin, J. Risk assessment of inland waterborne transportation using data mining. *Marit. Policy Manag.* **2020**, *47*, 633–648. [[CrossRef](#)]
71. Kinra, A.; Mukkamala, R.R.; Vatrappu, R. Methodological Demonstration of a Text Analytics Approach to Country Logistics System Assessments. In Proceedings of the 5th International Conference LDIC, Bremen, Germany, 22–25 February 2016; pp. 119–129. [[CrossRef](#)]
72. Treiblmaier, H.; Mair, P. Textual Data Science for Logistics and Supply Chain Management. *Logistics* **2021**, *5*, 56. [[CrossRef](#)]
73. Chae, B. Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *Int. J. Prod. Econ.* **2015**, *165*, 247–259. [[CrossRef](#)]
74. Singh, A.; Shukla, N.; Mishra, N. Social media data analytics to improve supply chain management in food industries. *Transp. Res. Part E Logist. Transp. Rev.* **2018**, *114*, 398–415. [[CrossRef](#)]
75. Hong, W.; Zheng, C.; Wu, L.; Pu, X. Analyzing the Relationship between Consumer Satisfaction and Fresh E-Commerce Logistics Service Using Text Mining Techniques. *Sustainability* **2019**, *11*, 3570. [[CrossRef](#)]
76. Meyer, A.; Walter, W.; Seuring, S. The Impact of the Coronavirus Pandemic on Supply Chains and Their Sustainability: A Text Mining Approach. *Front. Sustain.* **2021**, *2*, 631182. [[CrossRef](#)]

77. Sai, F. Study on Green Logistics Initiatives through Text Mining. In Proceedings of the 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taichung, Taiwan, 30 November–2 December 2018; pp. 110–115. [\[CrossRef\]](#)
78. Sharma, A.; Adhikary, A.; Borah, S.B. Covid-19's impact on supply chain decisions: Strategic insights from NASDAQ 100 firms using Twitter data. *J. Bus. Res.* **2020**, *117*, 443–449. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Kim, D.; Kim, S. Sustainable Supply Chain Based on News Articles and Sustainability Reports: Text Mining with Leximancer and DICTION. *Sustainability* **2017**, *9*, 1008. [\[CrossRef\]](#)
80. Choi, D.; Song, B. Exploring Technological Trends in Logistics: Topic Modeling-Based Patent Analysis. *Sustainability* **2018**, *10*, 2810. [\[CrossRef\]](#)
81. Chu, C.-Y.; Park, K.; Kremer, G.E. Applying Text-mining Techniques to Global Supply Chain Region Selection: Considering Regional Differences. *Procedia Manuf.* **2019**, *39*, 1691–1698. [\[CrossRef\]](#)
82. Chu, C.-Y.; Park, K.; Kremer, G.E. A global supply chain risk management framework: An application of text-mining to identify region-specific supply chain risks. *Adv. Eng. Inform.* **2020**, *45*, 101053. [\[CrossRef\]](#)
83. Wu, K.J.; Bin, Y.; Ren, M.; Tseng, M.-L.; Wang, Q.; Chiu, A.S. Reconfiguring a hierarchical supply chain model under pandemic using text mining and social media analysis. *Ind. Manag. Data Syst.* **2022**, *122*, 622–644. [\[CrossRef\]](#)
84. Moon, S.; Lee, G.; Chi, S. Semantic text-pairing for relevant provision identification in construction specification reviews. *Autom. Constr.* **2021**, *128*, 103780. [\[CrossRef\]](#)
85. Wen, Q.; Qiang, M.; Xia, B.; An, N. Discovering regulatory concerns on bridge management: An author-topic model based approach. *Transp. Policy* **2019**, *75*, 161–170. [\[CrossRef\]](#)
86. Wu, L.; Ye, K.; Yan, H.; Yang, T. Identifying Chinese Government's Concerns about Environmental Effects of Highway Construction: A Text Mining Approach. In *ICCREM 2018: Sustainable Construction and Prefabrication*; American Society of Civil Engineers: Reston, VA, USA, 2018; pp. 232–238.
87. Park, S.H.; Synn, J.; Kwon, O.H.; Sung, Y. Apriori-based text mining method for the advancement of the transportation management plan in expressway work zones. *J. Supercomput.* **2017**, *74*, 1283–1298. [\[CrossRef\]](#)
88. Chang, T.; Chi, S.; Im, S.-B. Understanding User Experience and Satisfaction with Urban Infrastructure through Text Mining of Civil Complaint Data. *J. Constr. Eng. Manag.* **2022**, *148*, 04022061. [\[CrossRef\]](#)
89. Das, S.; Devkar, G. Harnessing social media data for analyzing public inconvenience in construction of Indian metro rail projects. *CSI Trans. ICT* **2022**, *10*, 107–120. [\[CrossRef\]](#)
90. Chen, Y.; Lei, Z.; Ma, C. Research on the Evolution of the Competition Culture of Highway Construction Companies Based on Text Mining. *Sustainability* **2022**, *14*, 12351. [\[CrossRef\]](#)
91. Williams, T.P.; Gong, J. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Autom. Constr.* **2014**, *43*, 23–29. [\[CrossRef\]](#)
92. Abdirad, M.; Krishnan, K. Industry 4.0 in Logistics and Supply Chain Management Using Topic Modeling Method. In Proceedings of the 8th Annual World Conference of the Society for Industrial and Systems Engineering, Baltimore, MD, USA, 17–18 October 2019; pp. 001–006.
93. Tavana, M.; Shaabani, A.; Vanani, I.R.; Gangadhari, R.K. A Review of Digital Transformation on Supply Chain Process Management Using Text Mining. *Processes* **2022**, *10*, 842. [\[CrossRef\]](#)
94. Chen, M.-C.; Ho, P.H. Exploring technology opportunities and evolution of IoT-related logistics services with text mining. *Complex Intell. Syst.* **2021**, *7*, 2577–2595. [\[CrossRef\]](#)
95. Bo, J. Research on Cold Chain Logistics Risk in E-commerce Using Text Mining Technology. In Proceedings of the ICCMB 2020: Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business, Tokyo, Japan, 31 January–2 February 2020. [\[CrossRef\]](#)
96. Shah, S.; Lütjen, M.; Freitag, M. Text Mining for Supply Chain Risk Management in the Apparel Industry. *Appl. Sci.* **2021**, *11*, 2323. [\[CrossRef\]](#)
97. Kim, J.J.; Jang, H.; Roh, S. A systematic literature review on humanitarian logistics using network analysis and topic modeling. *Asian J. Shipp. Logist.* **2022**, *38*, 263–278. [\[CrossRef\]](#)
98. Shin, S.-H.; Kwon, O.K.; Ruan, X.; Chhetri, P.; Lee, P.T.-W.; Shahparvari, S. Analyzing Sustainability Literature in Maritime Studies with Text Mining. *Sustainability* **2018**, *10*, 3522. [\[CrossRef\]](#)
99. Hong, J.; Tamakloe, R.; Lee, G.; Park, D. Insight from Scientific Study in Logistics using Text Mining. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, *2673*, 97–107. [\[CrossRef\]](#)
100. Zondervan, N.A.; Tolentino-Zondervan, F.; Moeke, D. Logistics Trends and Innovations in Response to COVID-19 Pandemic: An Analysis Using Text Mining. *Processes* **2022**, *10*, 2667. [\[CrossRef\]](#)
101. Fosu-Saah, B.; Hafez, M.; Ksaibati, K. A Review of Accelerated Pavement Testing Applications in Non-Pavement Research. *Civileng* **2021**, *2*, 612–631. [\[CrossRef\]](#)
102. Atay, M.; Eroğlu, Y.; Seçkiner, S.U. Investigation of Breaking Points in the Airline Industry with Airline Optimization Studies Through Text Mining before the COVID-19 Pandemic. *Transp. Res. Rec. J. Transp. Res. Board* **2021**, *2675*, 301–313. [\[CrossRef\]](#)
103. Das, S. Exploratory Analysis of Unmanned Aircraft Sightings using Text Mining. *Transp. Res. Rec. J. Transp. Res. Board* **2021**, *2675*, 291–300. [\[CrossRef\]](#)
104. Çallı, L.; Çallı, F. Understanding Airline Passengers during Covid-19 Outbreak to Improve Service Quality: Topic Modeling Approach to Complaints with Latent Dirichlet Allocation Algorithm. *Transp. Res. Rec. J. Transp. Res. Board* **2022**. [\[CrossRef\]](#)

105. Park, J.Y.; Mistur, E.; Kim, D.; Mo, Y.; Hofer, R. Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs. *Sustain. Cities Soc.* **2021**, *76*, 103524. [[CrossRef](#)]
106. Kwon, H.; Kim, J.; Park, Y. Applying LSA text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation* **2017**, *60–61*, 15–28. [[CrossRef](#)]
107. Zhou, Y.; Wang, X.; Yuen, K.F. Sustainability disclosure for container shipping: A text-mining approach. *Transp. Policy* **2021**, *110*, 465–477. [[CrossRef](#)]
108. Oliveira-Neto, F.M.; Han, L.D.; Jeong, M.K. Tracking Large Trucks in Real Time with License Plate Recognition and Text-Mining Techniques. *Transp. Res. Rec. J. Transp. Res. Board* **2009**, *2121*, 121–127. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.