IADIS International Journal on Computer Science and Information Systems
Vol. 17, No. 1, pp. 65-80
ISSN: 1646-3692

THE APPLICATION OF MACHINE LEARNING IN LITERATURE REVIEWS: A FRAMEWORK

Yusuf Bozkurt, Reiner Braun and Alexander Rossmann
Department of Computer Science, Reutlingen University, 72762 Reutlingen, Germany

ABSTRACT

Literature reviews are essential for any scientific work, both as part of a dissertation or as a stand-alone work. Scientists benefit from the fact that more and more literature is available in electronic form, and finding and accessing relevant literature has become more accessible through scientific databases. However, a traditional literature review method is characterized by a highly manual process, while technologies and methods in big data, machine learning, and text mining have advanced. Especially in areas where research streams are rapidly evolving, and topics are becoming more comprehensive, complex, and heterogeneous, it is challenging to provide a holistic overview and identify research gaps manually. Therefore, we have developed a framework that supports the traditional approach of conducting a literature review using machine learning and text mining methods. The framework is particularly suitable in cases where a large amount of literature is available, and a holistic understanding of the research area is needed. The framework consists of several steps in which the critical mind of the scientist is supported by machine learning. The unstructured text data is transformed into a structured form through data preparation realized with text mining, making it applicable for various machine learning techniques. A concrete example in the field of smart cities makes the framework tangible.

KEYWORDS

Machine Learning, Text Mining, Research Method, Literature Review, Clustering, Framework

1. INTRODUCTION

The information systems (IS) research field is characterized by diversity, heterogeneity, and interdisciplinarity (Okoli, 2015; Tate et al., 2015). The IS research community originated in the 1960s and has proved its scientific orientation through solid research traditions (Paré et al., 2015). The rapid development and growth of existing knowledge have resulted in a rich stream of literature on various topics. IS research is becoming increasingly extensive, complex, and heterogeneous. Therefore, a proper understanding and timely analysis of the existing body of knowledge are important to identify emerging topics and research gaps (Bandara et al., 2011;

Paré et al., 2015). IS researchers are often faced with the task of conducting a systematic literature review (SLR) of a domain because the analysis of related works is the foundation of any scientific research process (Webster & Watson, 2002). High-quality literature should be systematically analyzed and synthesized to (1) provide a solid foundation for the research process, (2) contribute to IS knowledge, and (3) provide an agenda for the research domain. However, conducting a suitable SLR is not trivial, given the number of pitfalls to manage, such as the rapid growth of knowledge in the IS area (Bandara et al., 2011) and the heterogeneity and lack of consistency in metadata (Tate et al., 2015).

The interest in research methods with respect to SLRs in IS cannot be overlooked (Kitchenham & Charters, 2007; Levy & Ellis, 2006; Okoli, 2015; Webster & Watson, 2002). Leading journals often encourage authors to contribute to IS theory and methods (e.g., MISQ Theory and Review) (Paré et al., 2015). Furthermore, special issues such as the *CAIS Special Issue: The Literature Review in Information Systems* foster the development of corresponding methods. The intent of the current work is not to replicate the valuable prior work (Kitchenham & Charters, 2007; Okoli, 2015; Webster & Watson, 2002), and a list of all the different methods and approaches to SLRs is beyond its scope. On a meta-level, a general process for SLR includes six steps: (1) formulating the problem, (2) conducting a literature search, (3) screening to include the relevant literature, (4) ensuring quality control, (5) extracting information, and (6) carrying out analysis and synthesis (Tate et al., 2015). Data analysis plays an essential role in those steps. Therefore, Tate et al. (2015) noted a relevant research gap in the area of SLRs: “We do not believe that discussion of the literature review in information systems reached saturation from this ‘renaissance’ of literature analysis methods publications. For example, the potential of ‘big data’-type analytics using text-mining approaches in literature analysis remains relatively unexplored.”(p. 108). Despite the advances in information technology in the context of big data, machine learning, and text mining, the implementation of SLRs is, in most cases, still a purely manual task. This might lead to serious shortcomings in SLRs in terms of quality and time. Motivated by this challenge, we addressed the following research question:

> *How can machine learning approaches be applied in SLRs to explore a broad research topic?*

To answer the research question, a framework was developed to support the SLR process with machine learning techniques. The framework contains steps for text mining, cluster analysis, and network analysis to analyze and structure a large amount of text data. In this way, in an interdisciplinary field such as IS, it is possible to create insights that are not immediately visible. The process of analyzing literature is accelerated, and human bias in selecting and evaluating literature is counteracted.

A typical SLR is still characterized by a strong manual process (Bandara et al., 2011), but given the rapid development and growth in IS, this is becoming increasingly difficult. It is also essential to conduct an effective literature review within a short time (Bandara et al., 2011; Tate et al., 2015). Therefore, the motivation of and justification for this study are due to four aspects: (1) the rapid growth of the literature volume, (2) the increasingly easy access to relevant articles, (3) the availability of work in electronic form, and (4) the transparent presentation of the developed method to support the IS community in the adoption and further development of machine learning and data analytics methods in SLRs. We contribute to the body of knowledge by addressing an important research method and extending it with machine learning techniques. Furthermore, we intend to provide a detailed description of the developed framework of a machine learning–supported literature review (ML-SLR) and an exemplary implementation. Although the approach originates from IS research as an example, it is not limited to the IS field but can also be applied to other research areas.

We begin by presenting a collection of related work to investigate existing research using machine learning in SLRs. The selection of related work involves conventional methods and manual exclusion. The reason is that the scope of related work is quite narrow and the focus is on a limited area with a manageable amount of literature. Then, in section 3, we provide a short overview of machine learning and text mining to familiarize the reader with some concepts. Subsequently, we describe the developed framework in its overall structure and explain each phase. The framework is described in section 3 without referring to specific tools, as an exemplary implementation for the domain of smart city research with concrete tool and software suggestions are presented in the fourth section. Finally, the research concludes with a discussion and suggestions for further research.

2. RELATED WORK

The application of machine learning approaches in a literature review is not a completely novel approach. Current articles predominantly include the application of text mining and natural language processing in SLR. Therefore, we explored the existing literature to gain an understanding of the research structure and to clarify the position of the contribution of this work. For this purpose, we conducted an analysis of related works in the databases IEEE, ACM, EBSCOhost, and Web of Science. To obtain consistent results, each database was queried using the same search criteria. The starting point of the database search was the search string "text mining" OR "NLP" OR "natural language processing" AND "literature review" OR “literature analysis” queried on article titles. To ensure high quality of the sources, we considered only peer-reviewed articles by selecting article (journal)/proceedings as the document type. The results uncovered 82 articles from Web of Science, 62 articles from ACM, 23 articles from IEEE, and 10 articles from EBSCOhost. In total, the search delivered 177 articles. After excluding duplicates and articles that did not match the search criteria, the total number of valid articles was 69 (Table 1).

Table 1. Search criteria related work

In-/Exclusion criteria		Search databases
Language:	English	Web of Science: 82 ACM: 62 IEEE: 23 EBSCOhost: 10
Search string:	(“text mining” OR “NLP” OR “natural language processing”) AND ("literature review" OR “literature analysis”)	
Considered field:	Title	Duplicates and not matching articles: 108 **TOTAL: 69**
Document type:	Article (journal); proceedings	

A review of the titles and abstracts of these 69 articles indicated that 58 of 69 studies applied text mining as a technique. In general, text mining is common in the biomedical field and is often used in PubMed publications to identify specific diseases, genes, drugs, and their relationships (Alshuwaier et al., 2017; Libbus & Rindflesch, 2002; Martin et al., 2004; Quan et al., 2014; Singhal et al., 2016). Figure 1 shows the distribution of the literature.

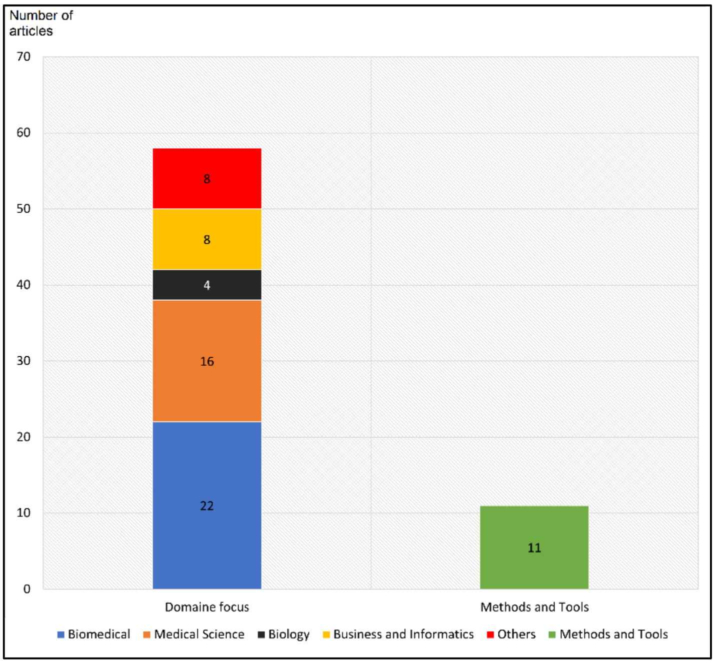


Figure 1. Distribution of literature

Eleven of the 69 papers are related to text mining systems and tools. For example, Phongwattana and Chan (2018) presented a framework that uses text mining techniques to search for similar articles and to show the most relevant sentences of the identified works. This framework facilitates the search for suitable articles by using the full abstract of an input article instead of searching with keywords. Therefore, Phongwattana and Chan (2018) created a database by extracting the entire corpus of Semantic Scholar in JSON format with the fields id, title, and abstract. Afterward, these data must be prepared with text mining, so that the search can be undertaken based on word vectors. For the actual search, the abstract of the input article must be processed with text mining. Furthermore, the TextRank algorithm identifies important sentences and lists them as extractive summaries. The main focus of this framework is on searching for articles based on other articles instead of using search strings. Nuzzo et al. (2010) described an automated method for analyzing scientific literature and extracting knowledge that consists of several tools and techniques. However, the approach is specifically adapted to biomedicine, in which special tools and databases support the analysis (e.g., Unified Medical Language System, Gene Extractor). The goals are to derive literature-based gene annotations and to identify correlations between diseases and genes. The articles exported from PubMed are prepared and enriched with data from Unified Medical Language System and Gene Extractor. Mergel et al. (2015) address the methodology of performing a systematic literature analysis by using text mining but focus only on designing proper search strings by means of text mining. They presented a method to optimize the search criteria by identifying additional relevant search terms in a research topic. According to Mergel et al. (2015), one of the most challenging tasks in conducting a literature review is the proper selection of search terms of the search string. The effect of the search string on the output of a literature review is considerable, as it is one of the major input parameters and the starting point of a literature review. Using a weak search string can lead to irrelevant results or may even exclude relevant articles. To define proper search strings of a research topic, an iterative approach is developed and implemented. The implemented tool extracts relevant search terms from selected articles using text mining and visualizes them on a graphical user interface. The database IEEE Xplore provides meta-information and abstracts of the selected articles by the user. By selecting one article, further search terms are suggested based on article abstracts, and the researcher can build the search string in an iterative way.

A gap remains with regard to the question of how machine learning can be embedded in the overall process of SLR. Especially there is little methodological work from this perspective within the IS community; most of the presented articles either use text mining procedures for their specific domain to identify relationships in the literature or deal with tools and systems for text mining. For example, Phongwattana and Chan (2018) stated that their approach is well suited as a supplementary research tool; however, they do not discuss a methodological way or focus on text mining techniques to identify suitable articles based on the article abstracts. Nuzzo et al. (2010) described their text mining system as well as the method of knowledge generation using text mining, but they refer to the special field of biomedicine. Mergel et al. (2015) contribute to IS research by optimizing SLRs with text mining. They described a method of iterative search string design; however, they do not discuss the details of implementation and only cover the task of search string building.

The paper at hand contributes to IS research by supporting the whole process of conducting an SLR with machine learning approaches such as text mining and cluster analysis. It describes how machine learning and text mining can support an SLR, especially in fast-growing domains. An exemplary implementation in the field of smart cities provides details for a simple reconstruction of the process in other research domains. By maintaining the qualitative part of a traditional SLR, as described by several IS articles (e.g., (Kitchenham & Charters, 2007; Okoli, 2015; Webster & Watson, 2002)), our method can help in broad research fields such as smart cities ((Albino et al., 2015; Batty et al., 2012; Chourabi et al., 2012; Meijer & Bolívar, 2016)), and artificial intelligence ((Das et al., 2015; Nilsson, 1982; Peek et al., 2015)) to establish a general understanding of the research topic without neglecting the qualitative and in-depth analysis of the most important articles.

3. APPROACH

3.1 Concepts

Machine learning deals with the extraction and analysis of patterns in data using concepts from statistics, text mining, and data mining (Das et al., 2015; Peek et al., 2015). Text mining refers to a collection of several computer-aided techniques, algorithms, and methods to extract implicit knowledge from unstructured text data. Text mining does not search for specific information, but rather aims to uncover patterns and relationships that human beings often cannot recognize because of the large amount of content. In contrast with data mining, the input data in text mining are unstructured in the form of documents, web pages, and so on (Gupta & Lehal, 2009; Hippner & Rentzmann, 2006; Hotho et al., 2005). A human-readable text is not usable for statistical procedures such as classification or clustering without computer-aided pre-processing

(Hotho et al., 2005). For this reason, we define the process of analysis in four stages, which form the basis of the framework:

1. Data gathering: The building of source data set (in our case, text documents).
2. Text encoding: This step transforms text documents into a computable form. All words in the text document are extracted and represented as word vectors with a weighted value. The encoded text can be considered structured data.
3. Data mining: The structured data (text) is suitable for different data mining approaches.
4. Visualization: This step is not trivial; through interactive and proper visualization, humans can recognize patterns and relationships.

3.2 ML-SLR Framework

In general, the ML-SLR framework consists of two types of tasks, the human task, and the machine task. In addition, artifacts are illustrated throughout the framework, which are either explicitly or implicitly produced after each task or serve as input for other tasks. The framework shows that the machine learning approach to conduct literature reviews is an iterative process and the critical reflection by the researcher is crucial. The qualitative and critical mind of the researcher is supported by machines, but not replaced by them. The core of the framework is highlighted by the Machine Learning–Supported Phase (see Figure *2*).

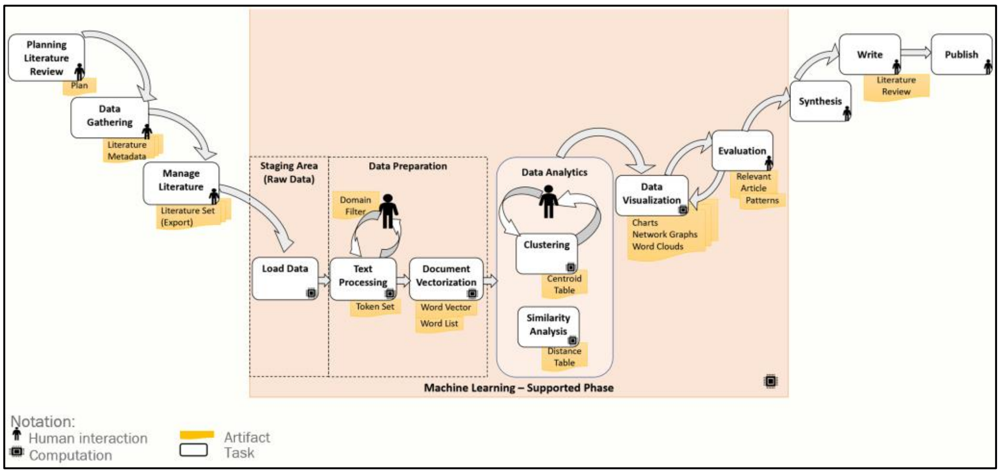


Figure 2. ML-SLR framework

The approach begins with the planning of the literature review – *Planning Literature Review*. The planning phase can be time-consuming, and we consider it one of the most important phases in the framework. Fundamental considerations must be made about the purpose and aim of the literature review. Answering some key questions can be helpful. For example, what is the purpose of the literature review? Who is the audience? Which research questions will the literature review answer? What is the search strategy (e.g., filters, databases), and how will the process be recorded? Here, the results form the artifact of this phase.

The *Data Gathering* task implements the defined search strategy. For this purpose, academic databases and search engines are queried. These databases allow a metadata export of the resulting articles as RIS, BibTex, or CSV files. The exported data includes title, author, journal, year of publication, abstract, and other information. In general, the goal is to extract as much information as possible, for potential use in data analysis. However, at least the abstract and the title of the article should be extracted, as the abstract is used for analyzing and the title for assigning. Analyzing abstracts is advantageous because these are usually written by the authors and precisely summarize the core aspects of the article. Furthermore, as the word count is much lower than that of the full text, there is less "noise", which could falsify the analysis. The data gathering task results in several database exports in different formats.

The gathered data/literature is managed with a reference management software in the task *Manage Literature*. The use of such software is essential because it allows easy integration of the different database exports. Furthermore, duplicates in the data set can be easily identified and removed. After all database exports are integrated into the reference management software, and all duplicates are removed, the entire library can be extracted as a CSV or XML file.

The first step in the *Machine Learning–Supported* phase is to import the literature set. The database should contain at least the abstracts and titles of the articles. The imported raw data are loaded into the *Staging Area* of the analytics software – *Load Data*. The raw data must be transformed and prepared to apply analytical methods. These methods have their origin in the related discipline of data mining, in which structured data is analyzed. In text mining, however, the information is encoded in texts (in our case, article abstracts). Although humans are able to interpret and contextualize what they read, analytical methods cannot process this information, which is why the term unstructured data is used (Kaiser, 2009). Therefore, a data preparation phase is required to transform the abstract of each article into a computable structured form.

The *Data Preparation* phase consists of two tasks: *Text Processing* and *Document Vectorization* (text encoding). The *Text Processing* task involves splitting each document's text into single indexable elements. The ML-SLR framework segments the text on the word level. This operation is called "tokenization". With the extraction of words in single tokens, the complexity of subsequent analysis operations is reduced. On the one hand, this increases the quality of the analysis, and on the other hand, the whole process becomes more performant (Aggarwal & Zhai, 2012a; Weiss et al., 2010). For this purpose, the *Text Processing* task combines a set of operations for data preparation:

- Removal of stop words: Stop words in English are, for example, "a," "the," "this," "me," and "you" and are not necessary for the analysis. By filtering out these words, "background noise" in the data set is eliminated.
- Uniform spelling: Uniform spelling simplifies the analysis, which is why all tokens are converted to either lower- or uppercase letters.
- Filter tokens by length: An article abstract sometimes has few tokens with only one letter. These either were tokenized incorrectly or are no longer meaningful. Therefore, a filter by token length (min. 2 letters) is recommended.
- Word stems: A word-stemming method can be used to reduce the complexity of the data set by returning the individual tokens to their basic form or to a uniform word stem (e.g., with the snowball-stemming algorithm). However, it has the disadvantage that the interpretability of the analysis results can be affected. Therefore, the necessity of word stemming should be examined for the respective application and research domain (Kaiser, 2009).

- Domain filters: Good data quality is the basis for any analysis. Therefore, a domain filter that removes certain irrelevant words is highly recommended. Especially in research articles, the abstracts often contain words such as "article" and "paper". To create such filters, the text mining process is first executed to determine the filter words from the resulting word list.
- N-grams: The complexity of the data set should be reduced, but in some cases, terms consist of two or more words and only make sense in combination. An example is "big data", which is a bigram (n = 2). N-gram algorithms are used to form such word combinations from single tokens. For example, a bigram algorithm combines two consecutive tokens and creates a new one containing both words. The use of n-grams should be considered individually, because it also generates irrelevant and meaningless combinations.

The *Text Processing* task creates a set of tokens for each article based on its abstract. The *Document Vectorization* task then transforms the data into an interpretable form and provides two artifacts, a word list, and a document vector. The word list contains all tokens after the text processing and how often they occur. The document vector is created from the tokens of the word list and the loaded documents (articles). The tokens are represented in columns, and the documents (e.g., with the title) in rows. Thus, each document becomes a vector through the use of a weighting method (e.g., term frequency–inverse document frequency) for the occurrence of a single token. Thus, the documents are presented in a structured form and can be analyzed with different data mining methods.

The analysis of texts can be divided into two basic procedures. The first is classification methods, in which a model is created by assigning (labeling) sample data sets (documents) to individual categories. This procedure is also called “supervised machine learning” (Aggarwal & Zhai, 2012b). The second procedure is clustering methods, in which similar data are clustered without a model or training data set. These methods do not require any pre-labeling by the researcher and are considered unsupervised machine learning (Feldman & Sanger, 2007). For our approach, we use unsupervised methods, because unknown structures in the data are to be determined without a target variable. The *Data Analytics* phase is structured as follows: Cluster algorithms group documents by their attributes and similarity values. This information is available in the vector space model, which has previously been built. For the determination of this similarity value, the cosine similarity has proved particularly suitable in the field of text mining (Li & Han, 2013). A good cluster result occurs if there is a high thematic similarity of the documents within a cluster but dissimilarity to the other clusters. An important variable in clustering is the number of clusters to be generated (Hotho et al., 2005). Note that the number of clusters strongly depends on the preferred level of observation. For a high level of observation, in which the main topics of a research field should be highlighted, we recommend a low number of clusters. The number of clusters can always be increased to gain more detailed insights into the structure and sub-themes. In cluster analysis, the researcher should iteratively define several cluster sizes and evaluate the results, for which methods such as the elbow method can be helpful to determine the number of clusters. For each iteration, the centroid table of the cluster is stored. In addition to the assignment of single articles to clusters, the linking of single articles to each other (*Similarity Analysis*) is an important insight for a literature review. The output of the *Similarity Analysis* is a distance table, which shows the distance of each article to the remaining articles and can be used to create a network graph.

The artifacts created so far can be used for *Data Visualization*. For example, the word list can be used to create word clouds. The artifacts centroid table and distance table can be used to create network graphs.

The *Evaluation* phase goes hand in hand with the visualization. Various visualization options support the researcher in gleaning insights and recognizing patterns: Support points of each document and linkages to other documents help identify important articles for full text reading. Furthermore, visualization of the clusters highlight the research structure and key topics within the research domain. By adjusting the number of clusters, a drill-down and roll-up of the research structure can be realized, allowing a detailed or abstract view of the research domain.

In the *Synthesis* task, the researcher first evaluates and synthesizes all insights from the analyzes and full text reading of the identified articles, and then aggregates, organizes, and compares the extracted findings (Kitchenham & Charters, 2007; Okoli, 2015; Webster & Watson, 2002). The results of the synthesis flow seamlessly into the writing process.

The penultimate step of the ML-SLR framework is *Write*. As already described in the introduction section, writing the literature review in an accessible way is important, so that readers can follow all steps exactly. Also relevant is describing the steps and decisions taken and providing the protocol of the literature search and extraction. SLRs contribute significantly to the body of knowledge. For example, they can evolve into frequently cited works of literature because they provide a sound basis for other researchers to conduct their own research (Okoli, 2015; Paré et al., 2015; Webster & Watson, 2002). Therefore, the ML-SLR framework ends with the task *Publish*.

4. EXEMPLARY IMPLEMENTATION

In this paper, we demonstrate the ML-SLR framework using an exemplary implementation. The underlying research domain is smart cities, which are particularly suitable for the implementation because of its complex and heterogeneous structure and the large number of publications. Given space limitations, we emphasize only the core of the ML-SLR framework in this example. A comprehensive description of the implementation, the source code and data is available on the following GitHub repository (https://github.com/yusufbzk/SLR-Text-Mining) to allow easy replication of the example. The framework has already been fully adopted in some published literature reviews. As for more and deeper implementation examples, we would like to refer to these works in the fields of data governance (Bozkurt et al., 2022), marketing (Rossmann et al., 2020), and smart cities (Bozkurt et al., 2020).

4.1 Setup

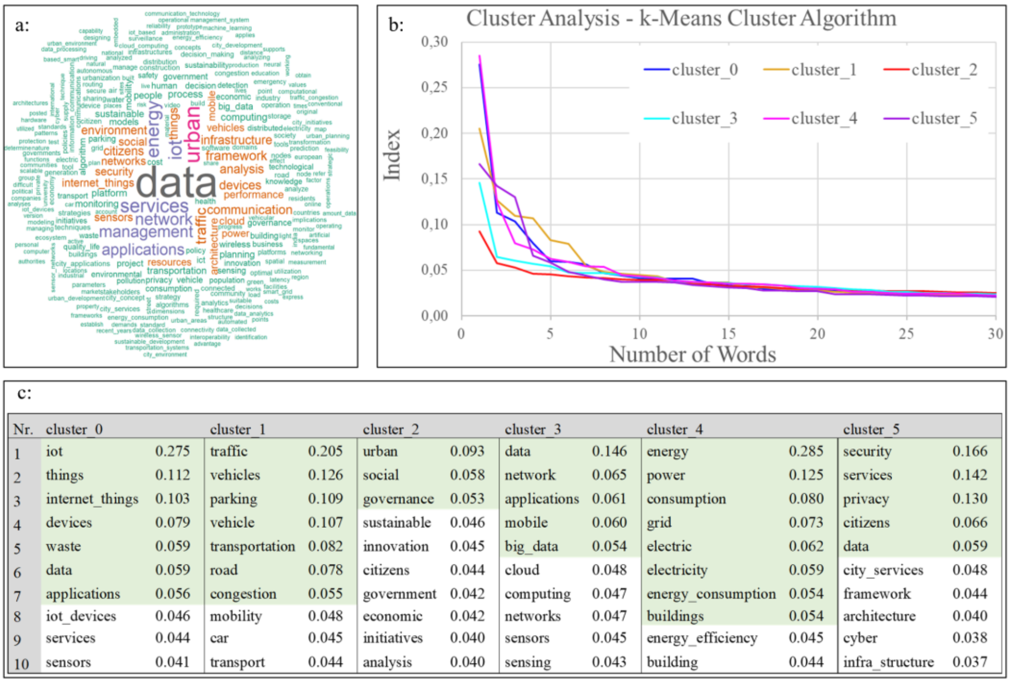
We collected the data set for the example from four scientific databases: IEEE, EBSCO, ACM, and Web of Science. Only peer-reviewed articles with the terms "smart city" OR "smart cities" in the title and published in the 2009–2019 period were included in the search. Of the articles, 2264 came from IEEE, 453 from EBSCO, 325 from ACM, and 1828 from Web of Science. For the integration and management of the individual data exports, we used the reference management software Citavi 6 (Swiss Academic Software, 2021). After removing duplicates, we exported the entire library with 4219 articles in total as an Excel file (Table 2). The exported data set was processed with RapidMiner according to the ML-SLR framework and visualized with Gephi. RapidMiner is a machine learning and data mining environment that is a suitable tool for the ML-SLR framework because of its text processing extensions (RapidMiner, 2021). For the visualization of interactive and complex graphs, we used the open-source software Gephi (Gephi Consortium, 2021), with a focus on network analysis. The following subsections show the most important artifacts of the method. We explain how these can be interpreted and how they can support the SLR. Again, this is only an example, and derivations can differ from research domain to research domain.

Table 2. Search strategy of the exemplary implementation

In-/Exclusion criteria		Search databases
Language:	English	IEEE: 2264
Search string:	“smart city” OR “smart cities”	EBSCO: 453
		ACM: 325
		Web of science: 1828
Considered field:	Title	
Time period:	2009 – 2019	Duplicates: 651
Document type:	Article (journal); proceedings	**TOTAL: 4219**

4.2 Results of the Data Preparation and Text Mining Process

As input for the analysis in RapidMiner, we exported an Excel list from Citavi 6, which contains information about the article’s author, title, year of publication, publishing journal or conference, the DOI number, and the abstract. As this information is to be combined with the other artifacts at a later time, we assigned each publication an ID number. A first indication of the subject areas appearing in the record comes from the frequently appearing words. Figure *3a),* shows the created word cloud.



Nr.	cluster_0		cluster_1		cluster_2		cluster_3		cluster_4		cluster_5	
1	iot	0.275	traffic	0.205	urban	0.093	data	0.146	energy	0.285	security	0.166
2	things	0.112	vehicles	0.126	social	0.058	network	0.065	power	0.125	services	0.142
3	internet_things	0.103	parking	0.109	governance	0.053	applications	0.061	consumption	0.080	privacy	0.130
4	devices	0.079	vehicle	0.107	sustainable	0.046	mobile	0.060	grid	0.073	citizens	0.066
5	waste	0.059	transportation	0.082	innovation	0.045	big_data	0.054	electric	0.062	data	0.059
6	data	0.059	road	0.078	citizens	0.044	cloud	0.048	electricity	0.059	city_services	0.048
7	applications	0.056	congestion	0.055	government	0.042	computing	0.047	energy_consumption	0.054	framework	0.044
8	iot_devices	0.046	mobility	0.048	economic	0.042	networks	0.047	buildings	0.054	architecture	0.040
9	services	0.044	car	0.045	initiatives	0.040	sensors	0.045	energy_efficiency	0.045	cyber	0.038
10	sensors	0.041	transport	0.044	analysis	0.040	sensing	0.043	building	0.044	infra_structure	0.037

Figure 3. a) Word cloud, b) centroid chart, c) centroid table with corresponding top 10 words

A word cloud reflects the frequency of words and can serve as a first indication. That is, a word cloud gives an indication of the importance of the corresponding words and also how the topics can be divided into clusters. Therefore, the next step considers which words and topics may result when clustering is performed with the k-means algorithm. *Figure 3c)*, shows six clusters and the corresponding words in the clusters with their importance index. To evaluate which words and topics are still relevant for a cluster, displaying them graphically is helpful (see panel b). The index value of the words shows an exponential curve. As soon as the index value of the individual words in the clusters resemble each other (in panel b at point 10), the individual topics and words can no longer be clearly assigned to a specific cluster. In the case described here, words with an index below 0.05 are no longer considered. Whether the classification of the clusters and the words is reasonable is now judged by the researcher. In our example, the cluster interpretation shown in Table 3 can be applied based on the graph (*Figure 3b))* .

Table 3. Cluster interpretation

Cluster	Topic
Cluster 0	Internet of Things
Cluster 1	Mobility
Cluster 2	Urban Governance
Cluster 3	Data and computing
Cluster 4	Energy
Cluster 5	Privacy and Security

4.3 Data Visualization and Evaluation

The last step of the implementation is to display the results in a network graph, which aids in visualizing the complex relationships. *Figure 4* shows the network graph for the smart city area.

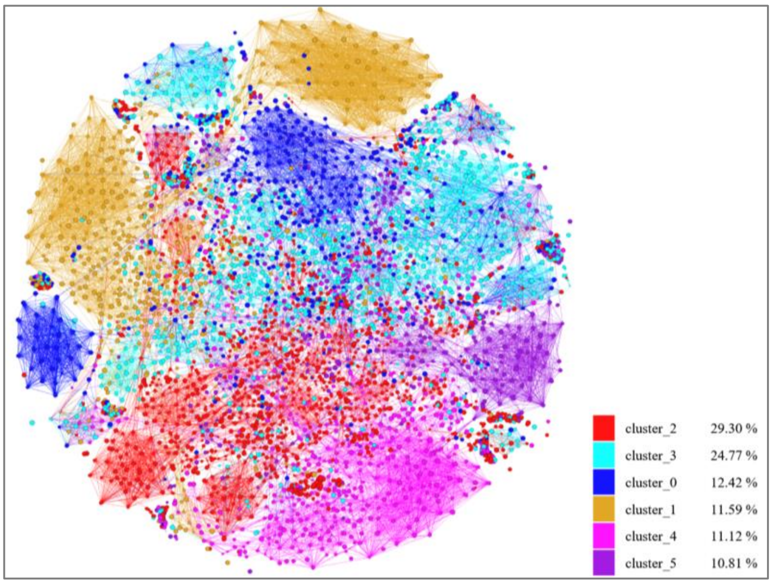


Figure 4. Network graph

Cluster 1 has two centers that are spatially separated from each other, cluster 4 has only one strong center that is close to cluster 2, and cluster 3 is strongly connected with all the other clusters. With regard to the individual topics in the clusters in *Figure 3c)*, the following can be derived: the area mobility (cluster 1) has strong connections with the areas internet of things (cluster 0), and the area data and computing (cluster 3) is connected with almost all the clusters.

5. DISCUSSION AND FURTHER RESEARCH

The main goal of this paper was to show a method for how machine learning can support SLR research. Motivated by the heterogeneity of IS research and the increasing number of accessible literature streams in electronic form, we developed the ML-SLR framework. The framework is based on the insights of the established SLR methods in IS research and combines machine learning and text mining to facilitate analyses of literature, especially in extensive research domains. The framework begins with the planning phase of a literature review. This is an important phase because it is the fundamental basis of a literature review, in which the scope is defined, research databases are identified, and search criteria are determined. The results of a suitable analysis are based on its input, so the research team should pay special attention to this. Implementing search criteria and the resultant collection of data from different research databases are not trivial steps. Dealing with research databases can be challenging, because different databases have different types of query syntax and generate different output data. The collected data can be created and exported relatively easy into a central library by using a reference management software. We recommend extracting as much metadata about documents as possible, so that analysis possibilities are not limited by missing data. However, the minimum information to be collected is the abstract and title. This information is accessible to humans as text but cannot be processed in its raw form by machine learning techniques. Therefore, the first phase in the core of the framework is the transformation of the unstructured data (text) into structured data (vector space). This transformation requires the active involvement of the researcher, which means filtering out the domain-specific irrelevant words. In the context of data analytics, the clustering method and similarity analysis are used. The evaluation represents the last step in the core. During the evaluation phase, the researcher evaluates the results of the data visualization. Here, the structure of the research domain can be explored, networks and patterns can be revealed, and important articles can be identified. Furthermore, different visualizations based on different numbers of clusters allow different depths of analysis. It is important that all articles are clearly retrievable (e.g., with a unique ID and title assignment). This allows identifying important articles for full text reading. After this stage, the findings from the full text reading and insights of the data analysis of the entire library are integrated into a synthesis to write the final literature analysis. The framework is based on the combination of human knowledge and machine learning techniques. Thus, the critical view and creativity in writing are combined with the speed and ability of machine learning techniques to process large amounts of data.

The presented framework provides a solid basis to support the researcher in the preparation of an SLR. Individual researchers need to consider when this framework is suitable. For example, in the related work section, we carried out the literature review manually using conventional approaches. The reason is that the scope for the related work is narrowly defined, and the focus is on a limited area with a manageable number of articles. Therefore, this framework is particularly suitable for SLRs with the following characteristics: (1) the research domain is interdisciplinary, and its research streams are heterogeneous, (2) a large number of publications can be considered for analysis, and (3) the research streams cannot be estimated in advance. Finally, the outcome of such a literature analysis is the identification of the research structure, research streams, theories and concepts, trends, and the creation of a research agenda that highlights research gaps. Unsupervised methods can be used to perform data-driven analyses in unknown research domains with little manual effort. Here, the data can speak for themselves, and completely new aspects can emerge that would not have been known in advance. The framework is also relevant in rapidly growing areas, particularly in the IS domain, to help react quickly to trends. A large amount of data cannot be analyzed in a suitable time frame using conventional methods. As another contribution, the framework helps prevent human bias in the early phase of literature identification, thus avoiding manual exclusion based on title and abstract, which is inherent in conventional methods (Kitchenham & Charters, 2007; Okoli, 2015; Paré et al., 2015). Thus, no important information in the article or the article itself is filtered out as a result of human misinterpretation. Furthermore, titles do not fully reflect the precise content of articles, and as such, important information can be lost unless the researcher invests a great deal of time to read all of the abstracts in the selected literature. This time-consuming, and what can be an erroneous, process can be avoided by using our framework, which adds all the results of database queries into data analysis without human filtering. In this way, a large data set can be made accessible to the researcher.

As in any data analysis, data preparation plays an important role, as the analysis procedures are only as good as the data that go into the methods. Therefore, the data preparation should be individually adapted to each application domain. A limitation of this framework is that the researcher must interpret the formed clusters by him- or herself, whereas in conventional literature reviews, such clusters are derived from the theoretical knowledge. However, this limitation can be overcome by using a supervised method in the analysis that employs classification instead of clustering. This method helps build the model on the basis of established theory. A drawback, however, is that human bias can become a handicap when creating the training model. Thus, we recommend the validation of the tagging by a team of experts. In supervised machine learning, a given set of text documents is tagged by the user with classification examples that the algorithm should search for. The dataset tagged by the user constitutes the trained model. Subsequently, the trained model is used to evaluate new text documents according to the model (Carrizosa & Romero Morales, 2013; Cunningham et al., 2008; Liu et al., 2002). A supervised method can be applied in the presented framework in the data analytics phase. For this, the researcher needs to manually tag a given amount of articles into topics based on the abstract. Thereafter, the trained model is used to classify new articles.

Further research could extend the data analytics part of the framework by using a supervised technique. Analysis of full text instead of abstracts would also be a fruitful path for research. In this case, however, the data preparation phase of the documents needs to be more complex, because many factors can falsify the analysis, such as filler words and references, as can the related work section of an article. In the future, the planning phase could also be optimized by machine learning and text mining, with the search string creation approach by Mergel et al. (2015). According to them, one of the most challenging tasks in conducting a literature review is the proper selection of search terms of the search string. The affect of the search string on the output of a literature review is considerable, as it is one of the major input parameters and the starting point of a literature review. Using a weak search string can lead to irrelevant results or may even exclude relevant articles (Babar & Zhang, 2009; Dybå & Dingsøyr, 2008; Mergel et al., 2015).

REFERENCES

Aggarwal, C. C., & Zhai, C. (Eds.). (2012a). *Mining Text Data*. Springer US. https://doi.org/10.1007/978-1-4614-3223-4

Aggarwal, C. C., & Zhai, C. (2012b). A Survey of Text Clustering Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 77–128). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_4

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, *22*(1), 3–21. https://doi.org/10.1080/10630732.2014.942092

Alshuwaier, F., Areshey, A., & Poon, J. (2017). A comparative study of the current technologies and approaches of relation extraction in biomedical literature using text mining. In *4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*.

Babar, M. A., & Zhang, H. (2009). Systematic literature reviews in software engineering: Preliminary results from interviews with researchers. In *3rd International Symposium on Empirical Software Engineering and Measurement*.

Bandara, W., Miskon, S., & Fielt, E. (2011). A systematic, tool-supported method for conducting literature reviews in information systems. *Proceedings of the 19th European Conference on Information Systems (ECIS 2011)*.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., & Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, *214*(1), 481–518. https://doi.org/10.1140/epjst/e2012-01703-3

Bozkurt, Y., Braun, R., Rossmann, A., & Hertweck, D. (2020). Smart Cities in Research: Status-Quo and Future Research Directions. *IADIS INTERNATIONAL JOURNAL on WWW/INTERNET*. Advance online publication. https://doi.org/10.33965/ijwi_202018108

Bozkurt, Y., Rossmann, A., & Pervez, Z. (2022). A Literature Review of Data Governance and Its Applicability to Smart Cities. In *Proceedings of the Annual Hawaii International Conference on System Sciences, Proceedings of the 55th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2022.333

Carrizosa, E., & Romero Morales, D. (2013). Supervised classification and mathematical optimization. *Computers & Operations Research, 40*(1), 150–165. https://doi.org/10.1016/j.cor.2012.05.015

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., Pardo, T. A., & Scholl, H. J. (2012). Understanding Smart Cities: An Integrative Framework. In *45th Hawaii International Conference on System Sciences* (pp. 2289–2297). IEEE. https://doi.org/10.1109/HICSS.2012.615

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised Learning. In M. Cord & P. Cunningham (Eds.), *Cognitive Technologies. Machine Learning Techniques for Multimedia* (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications*, *115*(9), 31–41. https://doi.org/10.5120/20182-2402

Dybå, T., & Dingsøyr, T. (2008). Strength of evidence in systematic reviews in software engineering. In D. Rombach, S. Elbaum, & J. Münch (Eds.), *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement* (p. 178). ACM Press. https://doi.org/10.1145/1414004.1414034

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Gephi Consortium. (2021). *Gephi*. https://gephi.org/

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1). https://doi.org/10.4304/jetwi.1.1.60-76

Hippner, H., & Rentzmann, R. (2006). Text Mining. *Informatik-Spektrum*, *29*(4), 287–290. https://doi.org/10.1007/s00287-006-0091-y

Hotho, A., Nürnberger, A., & Paass, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20*, 19–62.

Kaiser, C. (2009). Opinion Mining im Web 2.0 — Konzept und Fallbeispiel. *HMD Praxis Der Wirtschaftsinformatik*, *46*(4), 90–99. https://doi.org/10.1007/BF03340384

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Technical Report No. EBSE-2007-01*.

Levy, Y., & Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science: The International Journal of an Emerging Transdiscipline*, *9*, 181–212. https://doi.org/10.28945/479

Li, B [Baoli], & Han, L. (2013). Distance Weighted Cosine Similarity Measure for Text Classification. In H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li, & X. Yao (Eds.), *Lecture Notes in Computer Science. Intelligent Data Engineering and Automated Learning – IDEAL 2013* (Vol. 8206, pp. 611–618). https://doi.org/10.1007/978-3-642-41278-3_74

Libbus, B., & Rindflesch, T. (2002). NLP-based information extraction for managing the molecular biology literature. *Amia 2002 Symposium, Proceedings: Biomedical Informatics: One Discipline*, 445–449.

Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002). Partially supervised classification of text documents. In *ICML*.

Martin, E., Bremer, E., Guerin, M.-C., DeSesa, C., & Jouve, O. (2004). Analysis of protein/protein interactions through biomedical literature: Text mining of abstracts vs. text mining of full text articles. *Knowledge Discovery in Life Science Literature, Proceedings*, *3303*, 96–108.

Meijer, A., & Bolívar, M. P. R. (2016). Governing the smart city: a review of the literature on smart urban governance. *International Review of Administrative Sciences*, *82*(2), 392–408. https://doi.org/10.1177/0020852314564308

Mergel, G. D., Silveira, M. S., & da Silva, T. S. (2015). A method to support search string building in systematic literature reviews through visual text mining. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing - SAC '15*.

Nilsson, N. J. (1982). *Principles of Artificial Intelligence* (1. Aufl.). Springer Berlin Heidelberg. http://gbv.eblib.com/patron/FullRecord.aspx?p=1877166

Nuzzo, A., Mulas, F., Gabetta, M., Arbustini, E., Zupan, B., Larizza, C., & Bellazzi, R. (2010). Text Mining approaches for Automated Literature Knowledge Extraction and Representation. *E-Health - for Continuity of Care*, *160*, 954–958. https://doi.org/10.3233/978-1-60750-588-4-954

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, *37*(1), Article 43, 879–910. https://doi.org/10.17705/1CAIS.03743

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial Intelligence in Medicine*, *65*(1), 61–73. https://doi.org/10.1016/j.artmed.2015.07.003

Phongwattana, T., & Chan, J. H. (2018). A Combination of Text Mining Techniques for Relevant Literature Search and Extractive Summarization. In *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval - NLPIR 2018,* Bangkok, Thailand.

Quan, C., Wang, M., & Ren, F. (2014). An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature. *PloS One*, *9*(7). https://doi.org/10.1371/journal.pone.0102039

RapidMiner. (2021). *RapidMiner Studio*. https://rapidminer.com/products/studio/

Rossmann, A., Bozkurt, Y., & Heinz, A. Machine Learning in Marketing: A Systematic Literature and Text Mining Research. In *2021 AMA Winter Academic Conference.* (Original work published 2020)

Singhal, A., Simmons, M., & Lu, Z. (2016). Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. *Journal of the American Medical Informatics Association*, *23*(4), 766–772. https://doi.org/10.1093/jamia/ocw041

Swiss Academic Software. (2021). *Citavi*. https://www.citavi.com/en

Tate, M., Furtmueller, E., Evermann, J., & Bandara, W. (2015). Introduction to the Special Issue: The Literature Review in Information Systems. *Communications of the Association for Information Systems*, *37*(1), 103–111. https://doi.org/10.17705/1CAIS.03705

Webster, R., & Watson, R. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), 13–23.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of predictive text mining*. *Texts in computer science*. Springer.