Highlights

Machine Learning for Semi-Automated Scoping Reviews

Sharon Mozgai,Cari Kaurloto,Jade Winn,Andrew Leeds,Dirk Heylen,Arno Hartholt,Stefan Scherer

- The number of published scoping reviews conducted per year is increasing steadily while the number of published papers is also growing rapidly across many fields, yet traditional scoping review methodology cannot keep up with this deluge of data.
- We propose a semi-automated approach leveraging state-of-the-art representation learning and established clustering techniques to rapidly accelerate the process of scoping reviews.
- We present results of our methods on two separate scoping review datasets of research papers, one large (N>1000) and one small (N<500), and describe how our method successfully accomplishes our desiderata of replicability, objectivity, automation & scalability, and discovery.

Machine Learning for Semi-Automated Scoping Reviews

Sharon Mozgai[a,*], Cari Kaurloto[b], Jade Winn[b], Andrew Leeds[a], Dirk Heylen[c], Arno Hartholt[a] and Stefan Scherer[a]

[a] University of Southern California Institute for Creative Technologies, 12015 Waterfront Dr., Playa Vista, 90096, CA, USA
[b] University of Southern California Libraries, 3550 Trousdale Parkway Los Angeles, Los Angeles, 90089, CA, USA
[b] University of Twente, Drienerlolaan 5, Enschede, 7522 NB, The Netherlands

ARTICLE INFO

Keywords:
Representation Learning
Clustering
Scoping Reviews
Semi-automated Review Process
Large Document Datasets

ABSTRACT

Scoping reviews are a type of research synthesis that aims to map the literature on a particular topic or research area. Though originally intended to provide a quick overview of a field of research, scoping review teams have been overwhelmed in recent years by a deluge of available research literature. This work presents the interdisciplinary development of a semi-automated scoping review methodology aimed at increasing the objectivity and speed of discovery in scoping reviews as well as the scalability of the scoping review process to datasets with tens of thousands of publications. To this end we leverage modern representation learning algorithms based on transformer models and established clustering methods to discover evidence maps, key themes within the data, knowledge gaps within the literature, and assess the feasibility of follow-on systematic reviews within a certain topic. To demonstrate the wide applicability of this methodology, we apply the here proposed semi-automated method to two separate datasets, a Virtual Human dataset with more than 30,000 peer-reviewed academic articles and a smaller Self-Avatar dataset with less than 500 peer-reviewed articles. To enable collaboration, we provide full access to analyzed datasets, keyword and author word clouds, as well as interactive evidence maps.

1. Introduction

Scoping reviews, also called “mapping” reviews, are a type of literature review commonly used for reconnaissance to clarify working definitions and map conceptual boundaries of a topic or field (Peters, Godfrey, Khalil, McInerney, Parker and Soares, 2015). Though originally intended to provide a quick overview of a field of research (Arksey and O’Malley, 2005), scoping review teams have been stymied in recent years by the swift growth of available research literature (Thomas, McNaught and Ananiadou, 2011). The impact of this aptly named "data deluge" has been further compounded by the explosion of new information technologies that enable the discovery of vast amounts of information and provide immediate access to primary research across multiple data collections (Hey and Trefethen, 2003; Bell, Hey and Szalay, 2009).

Though increased access to research resources might be considered an asset in the comprehensive mapping of a field, traditional scoping review methodology struggles to meet the expanding number of relevant and available publications. Current attempts to rapidly execute scoping reviews are often heavily involved with manual and bespoke processes that are not easily translatable across domains. While there have been attempts to leverage state-of-the-art machine learning and automated text analysis techniques in systematic reviews (Yamada, Yoneoka, Hiraike, Hino, Toyoshiba, Shishido, Noma, Shojima, Yamauchi et al., 2020; Thomas et al., 2011; Ananiadou, Rea, Okazaki, Procter and Thomas, 2009; Tsafnat, Glasziou, Choong, Dunn, Galgani and Coiera, 2014), research related to artificial intelligence and automated text mining with an application for scoping reviews remains scarce. Hence, we seek to employ replicable, programmatic, and automated steps that can be leveraged across any scoping review to gain insights into the extant literature.

Specifically, this approach follows a number of desiderata: (1) replicability, (2) objectivity, (3) automation & scalability, and (4) discovery reporting. First, **replicability** requires the approach to be applicable irrespective of domain (e.g., computer science or medicine). For our approach to be replicable we employ accessible and well-validated methods for analysis and limit subjective methods wherever possible. Second, **objectivity** requires us to employ well-defined automatic steps to evaluate and verify characteristics of the scoping review without manual intervention. Third, scoping reviews can vary significantly in size and hence **automation & scalability** can be a crucial factor. We therefore require the methods to be robust to both large and small scoping reviews. Fourth, **discovery & reporting** of novel insights are key for scoping reviews. We employ statistical methods to describe the *health* of a specific field of research as well as provide practical and accessible methods for researchers to share their results. Specifically, we share our results as an interactive evidence map of the dataset, as well as open-source research code to improve reproducibility.

The aims of this work are the development of a document-mining approach to support scoping reviews and to provide new insights into defining a semi-automated process to facilitate scoping reviews. This paper will make the following contributions:

- Introduce current manual scoping review methodology and aims as well as an exploration of its limitations.

*Corresponding author
mozgai@ict.usc.edu (S. Mozgai)
ORCID(s):

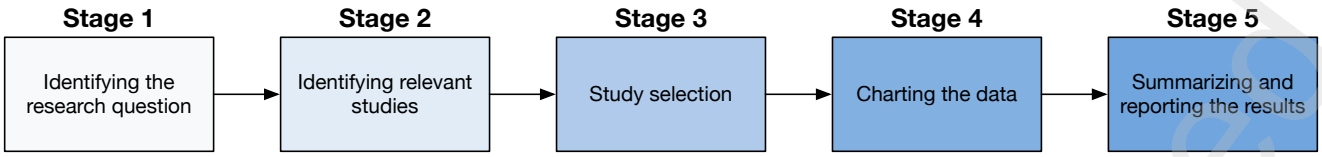


Figure 1: Scoping review stages.

- Present our semi-automated approach that systematizes and accelerates scoping review methodology through novel representation learning approaches and established clustering techniques.
- Validate the broad applicability and scalability of our approach on two datasets that significantly differ in size.
- Discuss how the semi-automated approach enables a more rapid, objective, and scaleable process of discovery within scoping reviews. Limitations of this methodology are also explored.

The remainder of this paper is organized as follows. Section 2 introduces the related work and existing research in the field. Section 3 describes the scoping review datasets included in the analysis and details the leveraged methodology to collate the documents. In Section 4 we detail the semi-automated approach employed to facilitate the scoping review. Section 5 provides specific results for the scoping review datasets and Section 6 discusses the results of our work and implications to scoping reviews writ large. Lastly, Section 7 concludes the paper. To enable reproducibility, we provide full access to the datasets, interactive evidence maps, and source code.

2. Related work

2.1. Manual scoping review framework and limitations

Scoping reviews are a relatively new approach to evidence synthesis with a general purpose of identifying and mapping the available evidence on a topic or field (Munn, Peters, Stern, Tufanaru, McArthur and Aromataris, 2018). Rather than being guided by a highly focused research question that lends itself to a particular study design (common to systematic reviews), scoping reviews are guided by the requirement to identify all relevant literature regardless of study design (Levac, Colquhoun and O'Brien, 2010; Tricco, Lillie, Zarin, O'Brien, Colquhoun, Kastner, Levac, Ng, Sharpe, Wilson et al., 2016) and can include grey literature to address questions beyond those related to intervention effectiveness (Arksey and O'Malley, 2005). The ability to synthesize findings from a variety of studies, including both quantitative and qualitative approaches, has contributed to the increased popularity of scoping reviews (Logan, Webb, Singh, Walsh, Tanner, Wall and Ayala, 2021), making this type of review particularly relevant to disciplines with emerging evidence, bodies of literature that exhibit a large, complex, or heterogeneous nature, cross-discipline investigations, or fields within the social sciences that typically do not conduct randomized clinical trials making it difficult to follow the methodology of traditional systematic reviews (Peters et al., 2015; Logan et al., 2021). The broad applicability of scoping reviews has led to their exploding popularity. In fact, the number of published scoping reviews conducted per year has increased steadily from a single report published in 2000 to over 3,093 published in 2019 (Tricco et al., 2016; Raitskaya and Tikhonova, 2019; Peters et al., 2015).

Though typically conducted with broader inclusion criteria than systematic reviews, scoping reviews still require rigorous and transparent methods to ensure that the results are valid and trustworthy (Munn et al., 2018). A formalized scoping review framework was initially proposed by Arksey and O'Malley in 2005. This framework was advanced by Levac et al. in 2010, Daudt et al., in 2013 and most recently by the Joanna Briggs institute in 2020 (Peterson, Pearce, Ferguson and Langford, 2017)). Despite this guidance, the conduct and reporting of scoping reviews is often inconsistent in the literature (Tricco et al., 2016), perhaps due to the complexity inherent in manually managing and reviewing the commonly large and diverse body of literature that is aggregated. However, a recent scoping review of scoping reviews published across multiple disciplines found that a majority of researchers do in fact attempt to follow the five-stage process originally outlined in the seminal work of (Arksey and O'Malley, 2005; Tricco et al., 2016). Broadly, this framework is as follows: Stage 1 includes refining research questions to be investigated, in Stage 2 relevant studies are identified, Stage 3 focuses on study assessment and selection, Stage 4 categorizes the data and finally, Stage 5 collates and summarizes the data (Raitskaya and Tikhonova, 2019) (cf. Fig. 1). Moreover, this scoping review revealed that the major purposes Arksey and O'Malley first outlined for conducting a scoping review remain among the most common goals pursued by published reviews. Research teams commonly undertake scoping reviews to (1) create an evidence map, (2) identify key themes and the breadth of the research, (3) detect gaps in the existing literature, and, (4) to determine the feasibility of conducting a full systematic review (Arksey and O'Malley, 2005; Peterson et al., 2017).

Navigating this five-stage framework in a thorough and thoughtful manner to achieve any of these aims takes a significant amount of time (Daudt, van Mossel and Scott, 2013). The rapid expansion of available evidence to be synthesised

is paradoxically making it more difficult to make evidence-informed decisions (Thomas et al., 2011). In fact, authors have reported that scoping reviews can take up to 20 months to complete (Peterson et al., 2017), evidence that researchers can no longer keep up with the workload using traditional manual methods of reviewing (Gusenbauer and Haddaway, 2020). Moreover, building a team that incorporates diverse expertise in library sciences, review methods, as well as searching and synthesis has been shown to significantly improve the quality of the review, but the resources and time needed to assemble these teams and execute a review may be a daunting prospect to many researchers (Gusenbauer and Haddaway, 2020). Thus, this paper builds on scoping review methodology by applying novel representation learning approaches to support review teams in successfully meeting this modern day data deluge.

2.2. Advances in knowledge synthesis

Currently, there is focus on innovation in evidence synthesis techniques that include the introduction of improved tools to make it easier to conduct this work. Evidence synthesis technologies such as reference management software and web-based software platforms allow for the more effective and efficient identification, analysis, synthesis, and reporting of research (Gusenbauer and Haddaway, 2020). Additionally, there is an extensive body of research focused on automating or assisting the tedious process of systematic reviews (Beller, Clark, Tsafnat, Adams, Diehl, Lund, Ouzzani, Thayer, Thomas, Turner et al., 2018; Tsafnat et al., 2014; Jonnalagadda, Goyal and Huffman, 2015). Available tools and task automation algorithms range from assisted meta-search tools (Tsafnat et al., 2014), machine learning based abstract screening (Wallace, Trikalinos, Lau, Brodley and Schmid, 2010; Wallace, Small, Brodley, Lau and Trikalinos, 2012), and automated result synthesis tools, like RevMan-HAL (Torres and Adams, 2017) and PRISMA (Page, McKenzie, Bossuyt, Boutron, Hoffmann, Mulrow, Shamseer, Tetzlaff, Akl, Brennan et al., 2021). For example, the tool *abstrackr* discussed in (Wallace et al., 2012) relies on a semi-automated active learning approach during which the human reviewer iteratively trains a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) to categorize titles and abstracts as either "relevant" or "irrelevant" for the systematic review. The features leveraged for the process rely on traditional representations of documents named bag-of-words (Wallach, 2006). A bag-of-words is a vector representation $x = x_1, ..., x_V$ of a document with a given vocabulary size V, where $x_i \; \forall i \in \{1, \dots, V\}$ is either 1 if the word i is present in the document or 0 if it is not present respectively. While these representations have the ability to capture topics with standardized vocabulary - which is not trivially extensible - they are not able to represent complex relationships *between* words (e.g., word order in bag-of-words is typically ignored) (Wallach, 2006). While extensions to bag-of-words, such as n-grams (Damashek, 1995) are possible, they are fundamentally limited to small $n \ll 10$, due to the combinatorial explosion of the vocabulary size. In the recent past, document representation algorithms relying on extremely large datasets and neural network models (e.g., transformer models) have been developed that are able to learn latent concepts and representations of language that go beyond any engineered approach (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017; Cohan, Feldman, Beltagy, Downey and Weld, 2020). Such neural models are specifically capable to represent complex relationships across sentences as well as entire documents and hence hold great promise for the automated analysis of scientific research documents (Cer, Yang, Kong, Hua, Limtiaco, John, Constant, Guajardo-Céspedes, Yuan, Tar et al., 2018; Beltagy, Peters and Cohan, 2020).

We leverage such document representation algorithms for the present work to identify relationships between documents present in the scoping review datasets.

3. Methods

3.1. Context

The here presented work is part of the Virtual Human Fidelity Coalition (VHFC). The VHFC is a collaboration between the University of Southern California Institute for Creative Technologies (USC ICT) and USC Libraries, sponsored by the US Army DEVCOM Soldier Center with specific guidance from its Simulation and Training Technology Center (STTC). Virtual humans are growing in cross-domain application, both as self-avatars to represent a specific person and as standalone artificial intelligence-controlled agents (Cassell, Sullivan, Prevost and Churchill, 2000; Hartholt, Traum, Marsella, Shapiro, Stratou, Leuski, Morency and Gratch, 2013; Hartholt, Mozgai and Rizzo, 2019; Hartholt, Fast, Reilly, Whitcup, Liewer and Mozgai, 2020). This cross-disciplinary work aims to investigate virtual human fidelity, defined as the degree to which a virtual human reproduces the sensory experience of interacting with a real person. While the technology to make a digital human look, sound and feel more realistic is improving, we know little about the differing levels of realism that may be required for a virtual human to be deemed acceptable and effective across different contexts and end users.

3.2. Datasets

The overarching goal of the VHFC is to explore the optimal fidelity of virtual humans across domains and end user demographics to maximize the efficacy of training outcomes. This research began with a scoping review of the literature related to virtual human fidelity to explore (1) Fidelity Domains (e.g., rendering style, voice quality, facial expressions, etc.), (2) Intervention Contexts (e.g., pedagogical, mental health, physical health, etc.), (3) Fidelity Evaluation Strategies (e.g., direct measures such as subjective surveys and indirect measures such as physiological responses), and (4) End Users (e.g., students, soldier, the elderly, etc.).

The date range for this review spans 1990 to 2021. Initial inclusion criteria for the scoping review are: journal articles, conference proceedings, dissertations and theses, and review

Table 1
Scoping review datasets.

Name	Search Terms	Date Range	Total Found	Included
VH dataset	virtual human(s), embodied conversation agent(s), virtual agent(s), digital human(s)	1990-2021	60640	32934
SA dataset	self-avatar(s)	2000-2021	1182	447

articles published in English. To be included articles also had to include all three of the following criteria: be a human subject research topic with real humans interacting with at least one virtual human, one aspect of the virtual human's fidelity needed to be varied, and both direct and indirect measures had to be collected. Newspapers, magazines, press releases, books, book chapters, conference reviews, editorials, notes, letters, short surveys, retracted, erratum, and undefined articles were excluded. Furthermore, articles that discussed robots and conversational agents that were not embodied were not included.

Given the breadth of the topic, the initial scoping review resulted in a sizable body of research to be synthesized. The total number of resources advancing to data screening was 34,153 after removing duplicates. This incredibly large dataset inspired us to investigate document-mining algorithms to accelerate and facilitate a comprehensive scoping review allowing us to explore the last thirty years of research on the digital representation of humans.

To further show generalizability of the here presented document-mining approach to a broad set of scoping reviews we further added a much smaller dataset to the investigations. We apply the here proposed semi-automated method to two separate datasets, the above described Virtual Human dataset (*VH dataset*) as well as a smaller Self-Avatar dataset (*SA dataset*). *VH dataset* is a large general scoping review with more than 30,000 documents and *SA dataset* is a small specific scoping review with around 450 documents (cf. Table 1). Both datasets were collected following the same protocols with different keyword terms as specified by the research team. A summary of search terms, date ranges, total number of works found (i.e., before removing duplicates or illegal entries), and total number of included publications is provided in Table 1.

Before processing the data, we clean the data of any missing datapoints. Due to the semi-automatic collection of the dataset using tools such as IEEEXplore or Web of Science, it is expected that a number of extracted metadata is missing, e.g., titles, abstracts, authors, and years. We therefore remove all entries with missing data from our analysis. After this pre-processing step, 32,934 papers remain in the *VH dataset* and 447 in the *SA dataset* respectively.

4. Semi-automatic document-mining approach

In the following, we present the semi-automatic document-mining approach ensuring the aforementioned desiderata of (1) replicability, (2) objectivity, (3) automation & scalability, and (4) discovery & reporting of results for both the *VH dataset* and *SA dataset* (cf. Section 3). When not explicitly discussed, methods, parameters, and approaches exactly matched between the document-mining of either dataset. To visualize complex relationships between papers, to discover emerging trends and topic clusters, as well as evidence gaps within a large unstructured document dataset, we devised a multi-step process visualized in Figure 2.

First, we leverage the state-of-the-art document-level representation learning method SPECTER trained directly on academic paper titles and abstracts as well as their citation-relationships (Cohan et al., 2020) to derive dense high-dimensional numeric representations for each document. Second, we employ t-SNE, a dimensionality reduction algorithm, to render the high-dimensional embeddings on a two-dimensional interactive mapping (Van der Maaten and Hinton, 2008), enabling the visual inspection of the relationships between papers. Third, to identify the number of research topics and their cluster entries within the vast field of research we employ the elbow method (Kodinariya and Makwana, 2013) to optimally identify k for the k-means clustering (Ahmed, Seraj and Islam, 2020). Fourth, we identify the topic of each cluster leveraging word cloud analysis (Cui, Wu, Liu, Wei, Zhou and Qu, 2010). While the naming of each cluster topic and its key terms is still a manual process, the visualization of word clouds greatly improves the comparability and accelerates the process considerably. Fifth, we leverage descriptive statistics to assess the clusters' overall characteristics and trends. Sixth, to identify possible evidence gaps within the scoping reviews data, we employ keyword matching to visualize how certain keywords appear across the entire space of the scoping review data. This process is further enabled by leveraging an interactive map and clear visualizations of the dataset.

4.1. Document embedding and visualization

As shown in Figure 2, we embed all documents using the document-level transformer model SPECTER (Cohan et al., 2020)[1] in a 768-dimensional representation space. For the embeddings we feed the model paper titles and abstracts in order for the model not to be biased by author names and publication years. We do not conduct any fine-tuning on the SPECTER model as we seek to make it as broadly applicable as possible for any dataset.

As it is difficult for a human to grasp the meaning and relationships of a high-dimensional representation of the document embeddings, we employ t-SNE (Van der Maaten and Hinton, 2008) to reduce the dimensionality of the document embeddings to only two. t-SNE is a non-linear dimensionality reduction technique that is widely used in

[1] https://github.com/allenai/specter

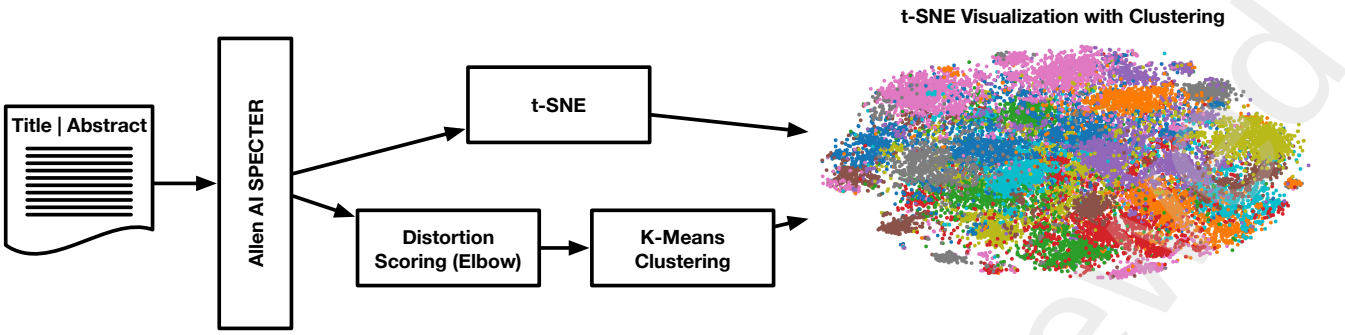


Figure 2: Approach overview. A collection of all papers' titles and abstracts are processed through SPECTER to derive high-dimensional semantic embeddings of each paper. To visualize the data in a relationship preserving two-dimensional representation we leverage t-SNE, while the high-dimensional embeddings are evaluated and clustered on a separate path. Lastly, the cluster assignments are used to color each datapoint in the two-dimensional representation.

the machine learning (Cohan et al., 2020; Ghosh, Chollet, Laksana, Morency and Scherer, 2017) and other fields of academic research. Using t-SNE we are able to visualize the data as well as maintain complex non-linear relationships between the datapoints. The goal of the t-SNE algorithm is to render similar datapoints close to each other and dissimilar datapoints further apart from each other on a low-dimensional space (i.e., two dimensions for the present work). For the purpose of this work we leverage the SciKit Learn implementation of t-SNE with the default parameter setting and a random seed of 0 for reproducibility.[2]

4.2. Topic clustering

For the clustering of the documents into topics we leverage the 768-dimensional representations. We use the high-dimensional document embeddings as input to the well known k-means clustering algorithm (Hamerly and Elkan, 2003). While k-means is a popular approach to cluster data and reveal groupings in the data, the right choice of k is not trivial and can severely bias the outcome. To identify the optimal number of clusters k, we employ the elbow metric (Satopaa, Albrecht, Irwin and Raghavan, 2011). The principal idea of the elbow method is to iteratively run k-means clustering on the dataset for a range of values of k. In our analysis, we ran the clustering for $k \in [2, 100]$ for the *VH dataset* and $k \in [2, 20]$ for the *SA dataset* respectively. For each value of k we calculate the sum of squared errors (SSE) as the distortion score. Then we choose the elbow as the trade-off value between an optimal SSE and a small k. Unfortunately, as seen in Fig. 3 no clear elbow is visible as the data is likely not very well clustered and significant overlap is expected. This finding may be an artifact of the dataset or signal pointing toward the homogeneity of the topic space within the dataset.

To further evaluate the quality of clusters created using k-means, we calculate the Silhouette score. This score is assessed by comparing for each cluster the intra-cluster distance (i.e., how similar data points are within the same cluster) and the mean nearest-cluster distance (i.e., how different a data point within a cluster is to the closest one in a different cluster). The Silhouette score is calculated for each sample of different clusters. The value of the Silhouette score ranges from -1 to 1. We observe a score of <0.1 for both datasets, which represents that the clusters are overlapping.

4.3. Topic identification and topic trends

To understand what each of the clusters represents we employed a word-cloud algorithm to distill each cluster's main topic in a human readable format. For the purpose of this work we use a common Python word cloud package[3]. Before running the algorithm we removed common words known as *stopwords* (e.g., *a*, *do*, *get*, *she*, or *I*) to render the word-cloud plots more meaningful and focused on the actual topic rather than just common English words. Specifically, we use the standard stopword dictionary that accompanies the Python implementation of the word cloud library. Once the word clouds (cf. Figure 4) were rendered, two reviewers reviewed each plot carefully and identified keywords (see Tables 3 and 4) present in each plot to provide a meaningful label to each cluster. While the process of naming the clusters may be somewhat subjective, the access to a reproducible, digestible, and quantitative algorithm such as the word cloud algorithm renders this process transparent and efficient.

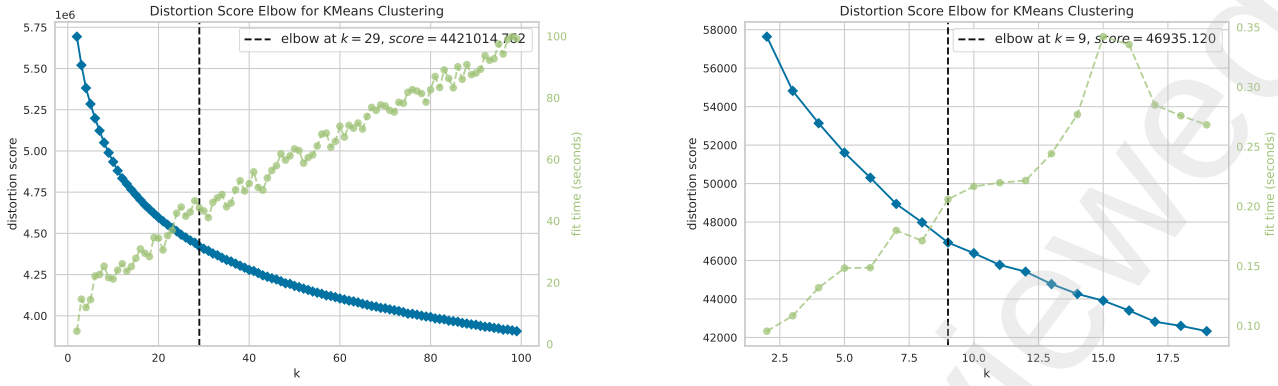
While it is possible to run any descriptive statistic across the datasets, for the purpose of this work we were specifically interested in understanding the *size* of each identified cluster (i.e., the number of publications assigned to the cluster) and the year over year (YoY) growth of each cluster. These statistics provide us a rough understanding on the trends, nascence, and possible decay of the topic within the scoping review.

4.4. Keyword matching

To allow a researcher to understand how a specific topic of their interest maps into the identified clusters and embedding space of the scoping review documents, we also provide access to regular expression based keyword matching. For

[2] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

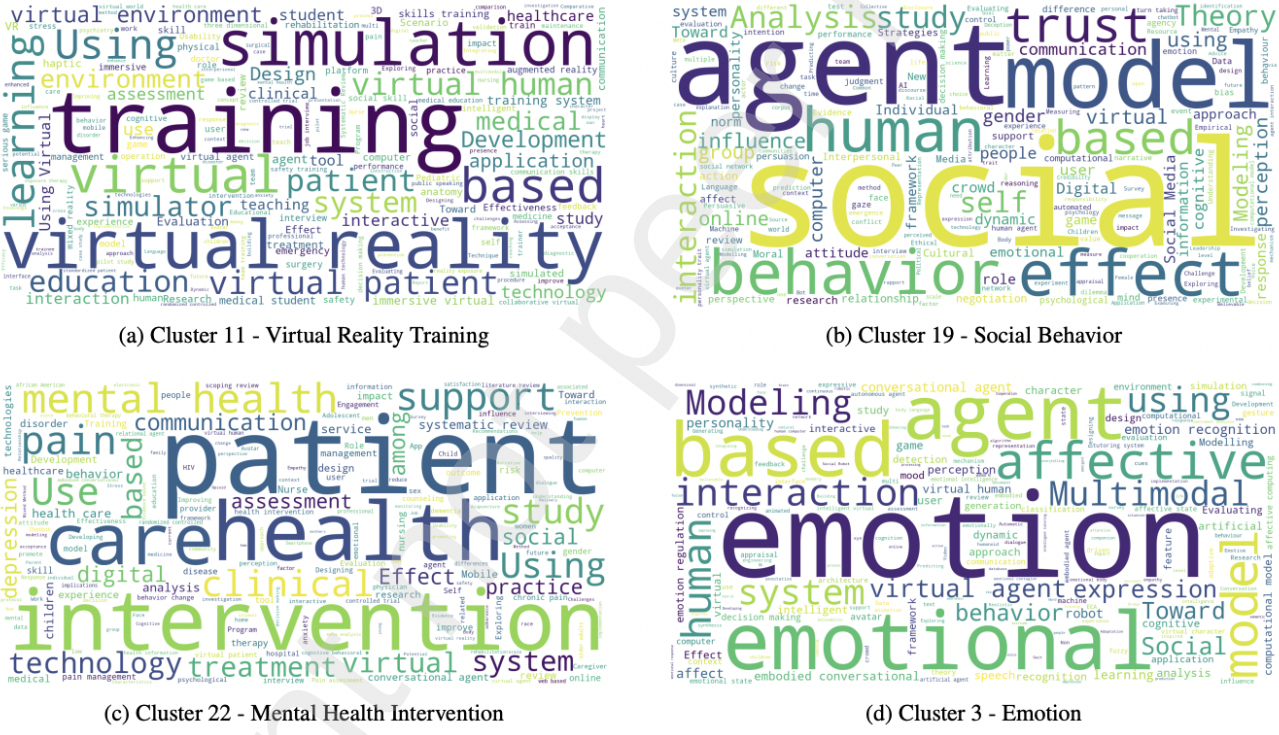
[3] https://github.com/amueller/word_cloud



(a) k-Elbow plot for *VH dataset*

(b) k-Elbow plot for *SA dataset*

Figure 3: Visualization of k-Elbow distortion metric to identify the optimal k for the k-means clustering given the document embeddings.



(a) Cluster 11 - Virtual Reality Training

(b) Cluster 19 - Social Behavior

(c) Cluster 22 - Mental Health Intervention

(d) Cluster 3 - Emotion

Figure 4: Visualization of four example word-clouds within VH dataset.

example, one might be interested how different demographic constraints (e.g., "military population") fit into the scoping review documents to understand if further research is warranted, a field is already crowded, or if a systematic review is even possible to dive deeper into the topic area of interest within the context of the scoping review's documents. These explorations require manual input to identify the keywords. We provide a few examples of such explorations in the evaluation section 5.

5. Evaluation

Here we present the outcomes of our document mining approach for both datasets. We first present the outcome of the topic clusters to summarize and map the state of evidence in Section 5.1. Then we identify key themes in the literature and present the breadth of research and yearly trends in each cluster in Section 5.2. Lastly, we explore preliminary steps towards identifying possible knowledge gaps within the literature as well as the feasibility to conduct a systematic review for specific topics within the realm of the scoping review datasets in Sections 5.3 and 5.4 .

Table 2
Mapping of Semi-automated Document-Mining Approach onto Scoping Review Goals.

Goal	Method	Automation Level	Description	Evaluation
Evidence Map	k-means Clustering	Automatic	4.2	5.1
Key Themes and Breadth of Research	Statistics/Word-Cloud	Automatic	4.3	5.2
Knowledge Gap Detection	Keyword Matching	Semi-Automatic	4.4	5.3
Feasibility of Systematic Review	Keyword Matching	Semi-Automatic	4.4	5.4

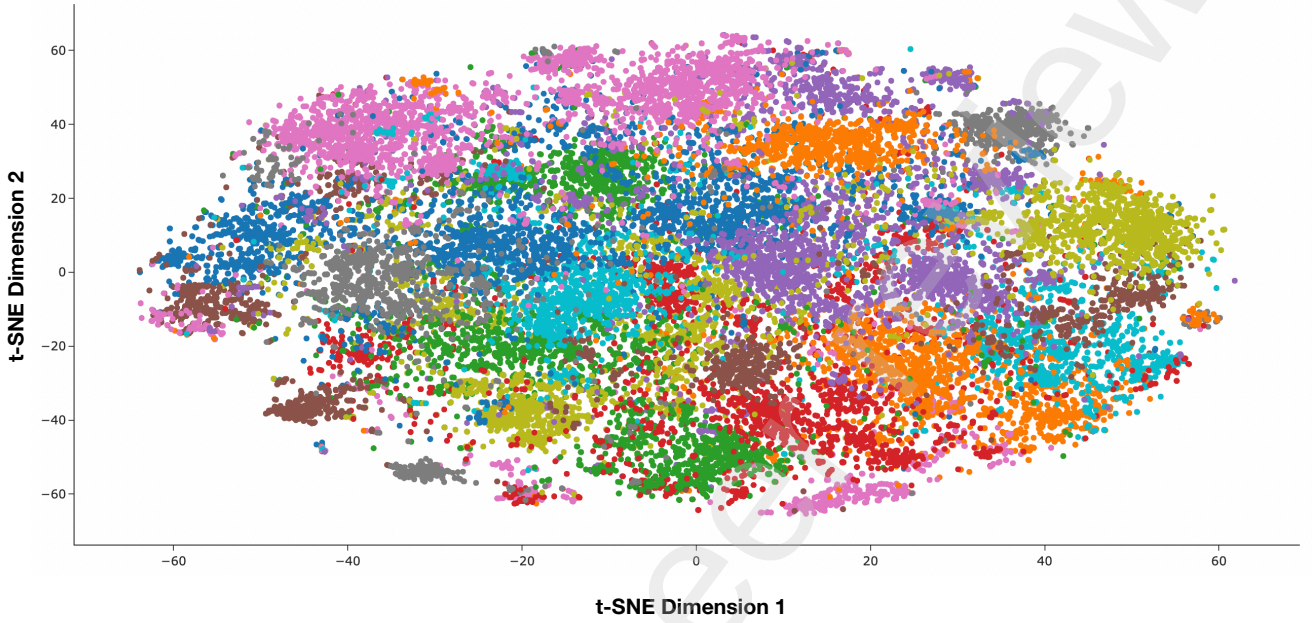


Figure 5: Visualization of the final k = 29 clustering for VH dataset. The axis correspond to the t-SNE projections. The clustering itself was computed by the high-dimensional document embeddings.

5.1. Evidence map

The optimal number of clusters was identified to be k = 29 for the *VH dataset* and k = 9 for the *SA dataset* respectively, as seen in Fig. 3. With these k values, we calculated a k-means clustering of the embeddings and visualized it in the two dimensional space for both datasets. The final cluster visualization is provided in Fig. 5 for the *VH dataset*[4]. Note that the clustering was computed with the 768-dimensional embedding vectors while the visualization in Fig. 5 is based on the t-SNE projections to only two dimensions. Based off of the Silhouette score, it is expected that the clusters are overlapping.

5.2. Key themes and breadth of research

To assign actual human interpretable topics to each of the clusters, we employ word-cloud analysis to each of the clusters and identify topics based on the keywords identified in each (Heimerl, Lohmann, Lange and Ertl, 2014). Figure 4 exemplifies clusters for the *VH dataset* from this analysis. Table 3 lists all clusters with accompanying keywords for the *VH dataset* and Table 4 for the *SA dataset*, respectively.

In order to assess the overall *health* (i.e., the growth in publications) of each cluster we calculate the average percentage growth (YoY) over the past 10 years in the number of publications. To illustrate the difference between a healthy and a lower trending cluster we provide histograms of the Healthcare cluster (Mean YoY growth = 35.3%) and the Animation cluster (Mean YoY growth = -3.4%) identified within the *VH dataset*. Fig. 6 visualizes the growth between the two clusters. Overall, the entire field of *VH dataset* appears to be growing at a steady pace of about 8% YoY for the past 10 years.

Due to the small number of documents in the *SA dataset* a YoY growth analysis is not possible as in some years no publications appeared in some clusters. Therefore we provide the bi-yearly growth for single clusters in Table 4. However, the field itself is rapidly growing overall at a 34.3% YoY growth, which is outpacing the larger field presented in the *VH dataset*.

5.3. Knowledge gap detection

One of the main purposes of scoping reviews is to determine knowledge gaps within the literature. To identify if a knowledge gap exists within the large and small scoping review datasets, we propose the use of keyword matching techniques applied to the title of the documents. To illustrate the approach we conduct a keyword matching example for the broader topic of *military* by searching for the keywords *military, soldier, veteran, weapon, army, air force, navy,*

[4]Interactive figures can be found at https://github.com/USC-ICT/VHFC

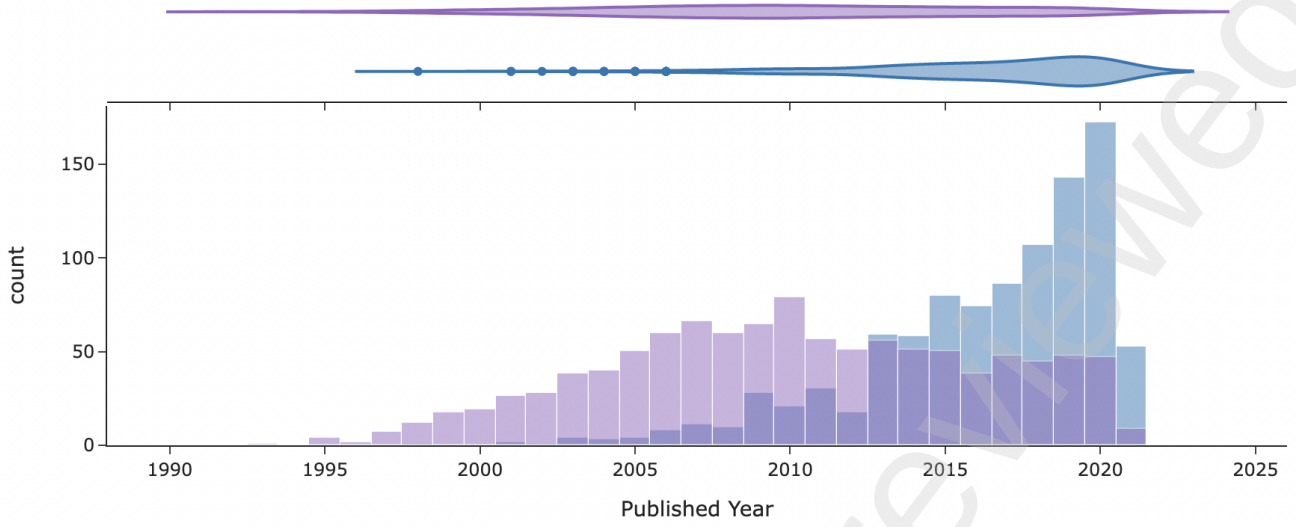


Figure 6: Year over year (YoY) growth comparison for *Healthcare* and *Animation* clusters respectively in VH dataset. *Healthcare* is shown in blue with a rapid growth trend and *Animation* is shown in purple with a longer history in the field and a slow decline over the years.

Table 3
Extracted clusters for the *VH dataset* with year over year (YoY) growth and total number of publications over the last ten years (i.e., 2010-2020).

ID	Name	Keywords	Growth	Pubs
0	virtual agents	interaction, social, realistic, multi agent, behavior	1.02%	982
1	biophysiology	cardiac, heart, arrhythmia, fibrillation	31.67%	211
2	human robot interaction	social robots, children, assistive robot, humanoid robot, multimodal, care	18.75%	1149
3	emotion	affective, facial expression, emotion recognition, behavior, speech, nonverbal behavior	0.64%	876
4	interfaces	user, interaction, multimodal, framework, model	14.06%	988
5	market research	design, digital, brand, loyalty, satisfaction, e commerce, service, customer	16.51%	786
6	biomechanics	ergonomics, force, muscle, joint, performance, motion, rehab	11.99%	789
7	system architecture	system, simulation, network, control, multi agent	11.19%	948
8	robot navigation	path planning, collision avoidance, swarm, tracking, navigation, UAV	2.21%	665
9	virtual interaction	avatar, virtual environments, virtual reality, agent, influence	8.92%	1166
10	vehicle	driving, ergonomics, passenger, seat, accident, driver simulation, car, safety	15.70%	678
11	VR training	training, virtual agents, virtual reality, simulation, medical, AR, surgery	20.23%	593
12	machine learning	representation learning, data, classification, prediction, reinforcement learning, neural network	7.30%	1165
13	animation	motion, 3d, model, character animation	-3.88%	570
14	gaming	serious games, storytelling, narrative, gamification, video game	6.27%	911
15	textiles	wearables, clothing, design, cloth, assembly, virtual fitting	22.95%	504
16	conversational agents	dialog, chatbot, conversation, communication, embodied conversational	13.23%	1178
17	modeling motion	movement, kinematic, motion simulation, motion tracking, posture	8.56%	882
18	medical imaging	patient, anatomy, CT, voxel, body, imaging, MRI, surgery	8.76%	598
19	social behavior	negotiation, trust, culture, empathy, social, presence, ethical, decision making, personality	11.91%	1172
20	multimodal interaction	speech, gesture, sign language, audiovisual, prosody	8.09%	814
21	virtual reality	virtual reality, virtual environment, augmented reality, immersive, mixed reality	11.91%	904
22	mental health	patient, treatment, depression, disorder, clinical, assessment, pain	35.39%	848
23	3d modeling	3d, model, shape, image, face, reconstruction, mapping, pose estimation	-0.82%	782
24	face	facial expression, face animation, perception, emotion, gender, eye, expression recognition	6.49%	483
25	ergonomics	design, evaluation, application, workplace, manufacturing, workstation, safety	14.64%	793
26	crowd simulation	pedestrian, crowd behavior, traffic, animation density	5.33%	776
27	learning	student, learner, education technology, children, pedagogical agent	7.82%	1209
28	pharmacokinetics	drugs, cardiac, atrial fibrillation, metabolism, physiology, treatment	13.07%	339

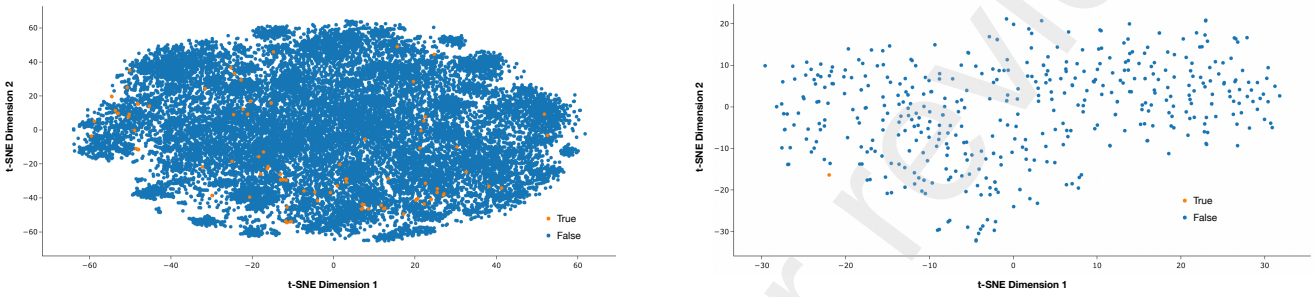
armor in both scoping review datasets. We visualize the result in Fig. 7. In the *VH dataset*, research regarding the military is spread out amongst topic clusters, providing evidence that there is diffuse research across topics, however, there is a lack of thematic concentration in a single domain. This example also clearly shows a significant knowledge gap within the literature on self-avatars with respect to the topic of *military*.

5.4. Feasibility of systematic review

The fourth aim of scoping reviews is to answer the question if a certain topic within the broader scoping review dataset warrants an in-depth systematic review to further our understanding of that particular topic (Tricco et al., 2016). As discussed in (Tricco et al., 2016, p. 9), one primary way to assess if an in-depth systematic review is feasible is for example "when at least ten studies are available on a specific

Table 4
Extracted clusters for the *SA dataset* with bi-yearly growth and total number of publications over the last ten years (i.e., 2010-2020).

ID	Name	Keywords	Growth	Publications
0	mental health	stress, therapy, body, addiction, emotion, image, anorexia, mindfulness	93.75%	26
1	virtual reality	HMD, virtual environment, egocentric distance, motion	39.52%	77
2	virtual worlds	media, social, second life, identity, community	57.78%	33
3	avatar creation	creation, intention, customization, identity, brand, consumer	50.16%	50
4	evaluation	participant, presence, experience, study, user	136.21%	79
5	data analysis	data, design, guidelines, learning, tangible, information	78.24%	23
6	gaming	player, game, massively multiplayer, gaming, Fortnite	117.22%	45
7	education	school, learning, student, science, mathematics	58.33%	24
8	therapeutic applications	rehabilitation, ASD, walking, Kinect, aging, effectiveness	46.30%	45



(a) Military category in *VH dataset*

(b) Military category in *SA dataset*

Figure 7: Visualization of the spread of the "military" topic within the two datasets. The topic is identified using a set of keywords related to the military, including: *military, soldier, veteran, weapon, army, air force, navy, armor*. Matching documents in the dataset are visualized with an orange dot and those which do not are visualized with a blue dot.

topic". Leveraging the semi-automatic clustering techniques in Section 4.2 and keyword spotting discussed in Section 5.3, we believe it is possible to, for example, conduct a systematic review of the use of virtual humans in the context of "military" simulation and training.

6. Discussion

As introduced above, the here proposed method seeks to follow the desiderata: (1) replicability, (2) objectivity, (3) automation & scalability, and (4) discovery to support the key objectives of scoping reviews (cf. Table 2). Within this section, we discuss how the proposed approach attempts to accomplish these characteristics, how they support the goals of scoping reviews, and where we identify shortcomings and possible improvements for future work.

Replicability. Replicability should be a core tenet of any study. Yet, due to the variations in approaches, absence of tools, and lack of consistency in focus of scoping reviews (Pham, Rajić, Greig, Sargeant, Papadopoulos and McEwen, 2014), replicability is traditionally not guaranteed. Within this work we follow a few strategies to facilitate replicability. First, we make all the data, search criteria, source code, and parameters used in the approach available through an open source repository hosted on GitHub.[5] Second, we leverage only open source software and well established algorithms within the work. Third, we further provide interactive interfaces whenever possible for readers and colleagues to investigate and experience the data themselves.

Objectivity. Whenever possible, we seek to reduce bias by eliminating subjective judgement from the approach. For example, we employ well defined automatic steps to evaluate and verify characteristics of the learned representations, such as the number of discovered topics within a scoping review dataset (cf. Section 4.2) and use the *elbow method* to identify an optimal number of k clusters for the k-means algorithm. However, it needs to be noted that this method is not perfect in the sense that it always identifies the *true* number of clusters in any given dataset, but rather identifies a trade-off between the number of clusters and their associated distortion score.

When analyzing the topic of each cluster, we also resorted to a method that only assists human subjective judgements rather than a fully automated term recognition (ATR) algorithm (Kageura and Umino, 1996). This is motivated by the fact that due to the nature of the collected dataset often ATR algorithms would identify terms that are present in the search criteria of the scoping review, such as "virtual human" instead of the actual topic of the cluster within the realm of virtual human research. When leveraging *pyate* (Lu, 2021) for ATR, for example, we found that the algorithm for Cluster 22 correctly identified "Mental Health" as the topic of the cluster, however, for Cluster 3 the automatically inferred term was "Virtual Agents", while "Emotion Recognition"

[5] https://github.com/USC-ICT/VHFC

came in a close second. We believe that these ATR methods have the potential to further refine the approach taken in this work, but this will require further iterations, subject to future investigations.

Scalability & Automation. The third desiderata seeks to show that the here presented approach can be applied to datasets of any size (including very large datasets) and help automate some of the most tedious steps of scoping reviews. Specifically, the approach was able to significantly speed up the identification of topics, their trends, and the presence of studies covering certain topics. There is no doubt about the immense workload of manual screening in scoping reviews. In both the abstract and full-text screening phases, two independent researchers must read and evaluate each resource for inclusion and exclusion variables. Even with small datasets this is a time-consuming and subjective process that can introduce bias, while large datasets exacerbate the issue of time commitments and efficiency. By reducing the amount of time dedicated to the screening process and by enabling the ability to rapidly discover patterns within the data, which can be a challenging process when employing manual screening procedures, we facilitate a more systematic and efficient scoping review process that can scale to even the largest of datasets.

Discovery. Last but not least, we seek to provide a set of tools and approaches that render scoping review datasets more accessible and discovery easier. Specifically, we would like to highlight the t-SNE algorithm (Van der Maaten and Hinton, 2008) that renders high dimensional document representations human consumable by projecting them to a 2D space. Together with an interactive system that allows the user to hover over titles[6] and explore similar fields with ease, this represents a significant step towards more accessible scoping review data discovery. Representation learning techniques and well established clustering algorithms, such as SPECTER and k-means employed in our work, transform text to meaningful representations and enable more targeted searching without increasing manual screening workload. Additionally, clustering in scoping reviews introduces a quantitative approach to the analysis of article datasets that can accelerate knowledge synthesis while hereto increasing objectivity and reducing some of the bias that can be introduced by subjective reviewers.

Together, these desiderata support scoping reviews on all aspects. The here introduced methods specifically support the mapping of the evidence, the identification of key themes and their breadth of research, the identification of knowledge gaps, and the assessment of feasibility to conduct a systematic review for a certain topic.

6.1. Limitations

Of course the here presented work is not without limitations. We discuss the main limitations of the present work in this section.

[6]To access the interactive maps go to our GitHub: https://github.com/USC-ICT/VHFC

While the semi-automated approach speeds up data synthesis, it still requires the qualitative analysis of clusters by domain experts. The here employed *elbow method* provides a mathematically optimal trade-off between the number of clusters k and the distortion score, however, it by no means always corresponds to the exact subjective clustering that a domain expert would potentially identify. Specifically, the domain experts need to verify and validate topics manually while inspecting the synthesized word-cloud plots to validate the cluster and identify specific cluster themes. Researchers should employ their own judgement when assessing the exact number of clusters and may need to discuss merging or splitting clusters given their observations. Within this work, we support this process of revising the clusters by enabling interactive analysis and word-cloud methods that allow rapid inspection of the articles associated with a certain cluster.

While our approach controls for the subjective bias introduced by manual screening, automated methods are certainly not without their own biases. We suggest to continuously screen for novel and improved representation learning algorithms that could replace the SPECTER model leveraged in this work. Further, while a full manual screening approach likely has a higher chance of identifying irrelevant articles that were introduced into the dataset during the initial search process (e.g., tables of content, articles with missing fields, irrelevant articles retrieved by keyword search, etc.), the automated process right now does not include a strong filtering approach. This could be introduced in a future update of the work, but is out of scope for this initial version.

7. Conclusions

Our work reports on the partnership of social science researchers, computer scientists, and librarians in the development of a document mining approach to support scoping reviews. We demonstrate the efficacy of our semi-automated technique in rapidly identifying patterns in both a large and small dataset of academic articles. This methodology can rapidly identify literature that should be further reviewed by researchers wishing to establish the current state of knowledge in a particular field or across multiple disciplines. As advances in information sciences increase the access to and volume of articles available to researchers, the application of validated semi-automated reviews will be a valuable tool that improves the efficiency of evidence synthesis projects and increases communication across disciplines.

Acknowledgments

This work was supported by National Institutes of Health National Institute on Alcohol Abuse and Alcoholism (NIAAA) grant R01 AA027225 and by the US Army under contract W911NF-14-D-0005. The content is the responsibility of the authors and does not necessarily reflect the official views of the NIAAA, nor the position or the policy of the Government, and no official endorsement should be inferred. Declarations of interest: none.

Appendix A. Supplementary data

Supplementary material related to this article can be found at https://github.com/USC-ICT/VHFC

References

Ahmed, M., Seraj, R., Islam, S.M.S., 2020. The k-means algorithm: a comprehensive survey and performance evaluation. Electronics 9, 1295.

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., Thomas, J., 2009. Supporting systematic reviews using text mining. Social Science Computer Review 27, 509–523.

Arksey, H., O'Malley, L., 2005. Scoping studies: towards a methodological framework. International journal of social research methodology 8, 19–32.

Bell, G., Hey, T., Szalay, A., 2009. Beyond the data deluge. Science 323, 1297–1298.

Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., et al., 2018. Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (icasr). Systematic reviews 7, 1–7.

Beltagy, I., Peters, M.E., Cohan, A., 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 .

Cassell, J., Sullivan, J., Prevost, S., Churchill, E., 2000. Embodied Conversational Agents edited by. Technical Report.

Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., et al., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175 .

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S., 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers, in: ACL.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X., Qu, H., 2010. Context preserving dynamic word cloud visualization, in: 2010 IEEE Pacific Visualization Symposium (PacificVis), IEEE. pp. 121–128.

Damashek, M., 1995. Gauging similarity with n-grams: Language-independent categorization of text. Science 267, 843–848.

Daudt, H.M., van Mossel, C., Scott, S.J., 2013. Enhancing the scoping study methodology: a large, inter-professional team's experience with arksey and o'malley's framework. BMC medical research methodology 13, 1–9.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.P., Scherer, S., 2017. Affect-lm: A neural language model for customizable affective text generation. arXiv preprint arXiv:1704.06851 .

Gusenbauer, M., Haddaway, N.R., 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. Research synthesis methods 11, 181–217.

Hamerly, G., Elkan, C., 2003. Learning the k in k-means. Advances in neural information processing systems 16, 281–288.

Hartholt, A., Fast, E., Reilly, A., Whitcup, W., Liewer, M., Mozgai, S., 2020. Multi-platform expansion of the virtual human toolkit: ubiquitous conversational agents. International Journal of Semantic Computing 14, 315–332.

Hartholt, A., Mozgai, S., Rizzo, A.S., 2019. Virtual job interviewing practice for high-anxiety populations, in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 238–240.

Hartholt, A., Traum, D., Marsella, S.C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.P., Gratch, J., 2013. All together now, in: International Workshop on Intelligent Virtual Agents, Springer. pp. 368–381.

Heimerl, F., Lohmann, S., Lange, S., Ertl, T., 2014. Word cloud explorer: Text analytics based on word clouds, in: 2014 47th Hawaii International Conference on System Sciences, IEEE. pp. 1833–1842.

Hey, A.J., Trefethen, A.E., 2003. The data deluge: An e-science perspective .

Jonnalagadda, S.R., Goyal, P., Huffman, M.D., 2015. Automating data extraction in systematic reviews: a systematic review. Systematic reviews 4, 1–16.

Kageura, K., Umino, B., 1996. Methods of automatic term recognition: A review. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 3, 259–289.

Kodinariya, T.M., Makwana, P.R., 2013. Review on determining number of cluster in k-means clustering. International Journal 1, 90–95.

Levac, D., Colquhoun, H., O'Brien, K.K., 2010. Scoping studies: advancing the methodology. Implementation science 5, 1–9.

Logan, J., Webb, J., Singh, N., Walsh, B., Tanner, N., Wall, M., Ayala, A.P., 2021. Scoping review search practices in the social sciences: A scoping review protocol .

Lu, K., 2021. kevinlu1248/pyate: Python automated term extraction. doi:10.5281/zenodo.5039289.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

Munn, Z., Peters, M.D., Stern, C., Tufanaru, C., McArthur, A., Aromataris, E., 2018. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. BMC medical research methodology 18, 1–7.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al., 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. Bmj 372.

Peters, M.D., Godfrey, C.M., Khalil, H., McInerney, P., Parker, D., Soares, C.B., 2015. Guidance for conducting systematic scoping reviews. JBI Evidence Implementation 13, 141–146.

Peterson, J., Pearce, P.F., Ferguson, L.A., Langford, C.A., 2017. Understanding scoping reviews: Definition, purpose, and process. Journal of the American Association of Nurse Practitioners 29, 12–16.

Pham, M.T., Rajić, A., Greig, J.D., Sargeant, J.M., Papadopoulos, A., McEwen, S.A., 2014. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Research synthesis methods 5, 371–385.

Raitskaya, L., Tikhonova, E., 2019. Scoping reviews: What is in a name? Journal of Language and Education 5, 4–9.

Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B., 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior, in: 2011 31st international conference on distributed computing systems workshops, IEEE. pp. 166–171.

Thomas, J., McNaught, J., Ananiadou, S., 2011. Applications of text mining within systematic reviews. Research synthesis methods 2, 1–14.

Torres, M.T., Adams, C.E., 2017. Revmanhal: towards automatic text generation in systematic reviews. Systematic reviews 6, 1–7.

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Kastner, M., Levac, D., Ng, C., Sharpe, J.P., Wilson, K., et al., 2016. A scoping review on the conduct and reporting of scoping reviews. BMC medical research methodology 16, 1–10.

Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F., Coiera, E., 2014. Systematic review automation technologies. Systematic reviews 3, 1–15.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Trikalinos, T.A., 2012. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr, in: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, pp. 819–824.

Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C., Schmid, C.H., 2010. Semi-automated screening of biomedical citations for systematic reviews. BMC bioinformatics 11, 1–11.

Wallach, H.M., 2006. Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd international conference on Machine learning, pp. 977–984.

Yamada, T., Yoneoka, D., Hiraike, Y., Hino, K., Toyoshiba, H., Shishido, A., Noma, H., Shojima, N., Yamauchi, T., et al., 2020. Deep neural network for reducing the screening workload in systematic reviews for clinical guidelines: algorithm validation study. Journal of medical Internet research 22, e22422.