Beyond CTRL F(ind): Exploring Insights Hidden in Abstracts through No-Code Text Mining Using the Example of Social Entrepreneurship

Nico Kling
Institute for Management and Information
University of Applied Sciences Zwickau
Zwickau, Germany
https://orcid.org/0000-0003-4987-9645

Chantal Kling
Institute for Management and Information
University of Applied Sciences Zwickau
Zwickau, Germany
chantal.kling@fh-zwickau.de

Anna-Maria Nitsche
University of Leipzig
Leipzig, Germany
https://orcid.org/0000-0003-3164-5066

Kevin Reuther
Innovation policy and transfer design
Fraunhofer Center for International Management and Knowledge Economics
Leipzig, Germany
https://orcid.org/0000-0003-3399-692

James B Johnston
School of Business and Creative Industries
University of West of the West of Scotland
Palsley, United Kingdom
https://orcid.org/0000-0002-1448-2392

***Abstract*—As the amount of scientific literature continues to grow, it becomes increasingly difficult for scientists to stay up to date within their own domains. Text mining has been suggested as a solution to this problem but has hardly established itself as an alternative to traditional literature reviews in management studies due to the requirement for coding knowledge and familiarity with algorithms. To address this issue, the authors of this paper introduce a no-code solution to text mining based on hierarchical clustering and cosine distance for the clustering of scientific abstracts and titles to create a fine-granular thematic clustering of a research field of choice. To demonstrate the approach, the authors applied it to 2,386 social entrepreneurship abstracts and titles, clustering them into 346 thematic clusters and further categorizing them into 17 different groups. These groups reflect different focus points of the literature, such as sustainability or diversity and inclusion. The authors believe that this approach is valuable both for early-stage researchers, as well as for experienced researchers, as it saves resources and is user-friendly, helping them tackle the ever-increasing amount of literature.**

Keywords—Text Mining; No-Code; Hierarchical Clustering; Machine Learning; Social Entrepreneurship

I. INTRODUCTION

There is widespread agreement in the academic community today that studying the published literature forms an essential part of scientific work, contributing to building a comprehensive body of knowledge [1]. It is used to identify existing knowledge gaps and consensus or disagreement in particular research fields [2] as well as to improve evidence-based decision-making processes [3]. It is also argued that literature reviews would form an important pillar of today's publication landscape: Several high-impact journals focus on the publication of such articles and are often highly cited, thus increasing the reputation of both the authors and the respective journals [4, 5]. Today, two phenomena related to the development of academic literature can be observed: Firstly, the number of publications in a variety of research fields is strongly increasing. Secondly, researchers' opportunities to access these publications may be better than ever. Both aspects jointly lead to the question of how to cope with this amount of available literature. As [6] argue, it was already considered a challenge for social scientists to keep up to date on and access primary research three decades ago. At the same time, they state that there is a general demand for literature analyses to be up-to-date, comprehensive and of high quality.

The use of text mining, machine learning algorithms and other IT-based solutions has already been suggested several times to tackle this issue [7]. Most of these articles originate from either medical science [8, 9] or computer science [10, 11]. Especially the affiliation of novel review techniques with medical research shows a strong similarity to the situation of systematic reviews when the article of [3] was published, which is still one of the most cited literature review methodologies in management to this date. Due to a variety of reasons, however, these approaches have hardly been able to assert themselves in the management research community so far, such as that traditional text mining techniques involve coding, which can be a time-consuming and resource-intensive process, as it requires a comprehensive understanding of programming languages as well as algorithms involved. As a result, researchers may encounter difficulties in analyzing large volumes of text data, which can hinder the advancement of research in various fields. One promising solution to the challenges associated with traditional text mining is the use of no-code platforms. No-code platforms often provide a user-friendly interface that allows users to analyze large volumes of different kinds of data without requiring a comprehensive understanding of programming languages and algorithms, while also offering an easy reproduction of results, as users only need the corresponding software version to mirror them. Additionally, it lowers the entry barrier for researchers who have not previously been exposed to such technologies [12].

II. TEXT MINING FOR LITERATURE REVIEWS – THE MACHINE LEARNING ALGORITHM BASED-REVIEW APPROACH

Indeed, the work of [13], and building on it the study of [14], show that no-code text mining is a variable alternative to traditional coded approaches for literature reviews, and promise a sound basis for algorithm-based literature reviews. Reuther used the Orange data mining software, which was developed at the University of Ljubljana in Slovenia in

1996/1997 [15]. It was chosen because it is open source, user-friendly, developed in a non-profit university research context, and constantly updated and developed. His integration of text mining using the Orange software follows a seven-step process developed based on systematic review approaches proposed by [3, 16, 17]. The author termed the approach machine learning algorithm-based reviews (MLR). The MLR approach is summarized in Figure 1.

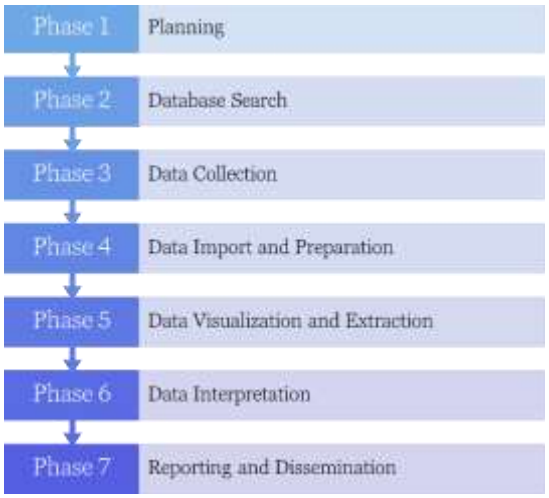

Figure 1 MLR process [13]

The first phase of the MLR approach is the identification and justification of the review scope, along with the formulation of the review problem and subsequent review question. This is similar to the approach suggested by [3]and includes the development of a review protocol. The second phase involves the database search, during which researchers select and define keywords, search terms, and databases for data collection. The third phase, data collection, involves generating a list of literature samples for the MLR, which is then exported from the database(s) and imported into Excel in preparation for the Orange text mining process. In the fourth phase, the data import and pre-processing of the text corpus are carried out using the Orange text mining software. The fifth phase of the MLR approach is the visualization of the data using dendrograms derived from hierarchical clustering based on Euclidean distance metric, and silhouette plots. At the end of this phase, researchers can either exit the MLR process by using the identified clusters to locate and justify a topic for a detailed conventional review or continue with the sixth phase to make interpretations based on the findings of the MLR. The sixth phase involves cluster analysis and synthesis, and the seventh phase involves the final cluster analysis and synthesis.

[13] suggests using titles for this procedure due to computing power restrictions typically faced by management scholars, as well as access restrictions for full texts and the lower thematic relevance of some full text sections such as reference lists.

III. AN ADAPTION OF THE MLR FOR THE USE OF SCIENTIFIC ABSTRACTS

While using titles only has its advantages, an adaptation of the approach using abstracts in addition to titles can also provide useful and relevant results. Abstracts provide a more comprehensive summary of the article, including key information such as the research objectives, methods, results, and conclusions, which can provide text mining algorithms with more information, leading to a more accurate analysis of the literature. Furthermore, using abstracts in text mining allows for the inclusion of relevant keywords and phrases that may not appear in the title, resulting in a more comprehensive search.

However, scientific abstracts contain some information like copyright or a hard textual division into Purpose, Methodology, Result and Summary. These represent noise for the algorithm, i.e. irrelevant information, which distorts the result, and must be removed accordingly. Additionally, when using abstracts as a basis for text mining, the distance metric needs to be changed due to the increased dimensionality of the data, as explained in the following chapter. The authors therefore recommend using the cosine distance. Cosine distance measures the angle between vectors, rather than their distance in space, and is thus a more accurate measure of the similarity between documents in high dimensionality, thus improving the effectiveness of the text mining analysis. Furthermore, the cosine metric was shown to be the most optimal to compute text similarity in comparative studies [18, 19].

Thus, the MLR process needs to be adapted if utilizing abstracts in addition to titles, as shown in Figure 2.

Figure 2 Adapted MLR process for use of abstracts

To underscore the utility of the adapted no-code text mining approach, this paper presents an exemplary case based on the research field of social entrepreneurship, comprising nearly 5000 papers on Scopus. The large number of papers offers a suitable test bed for the adapted approach, which results could also be used as a basis for a more detailed literature review. Additionally, the composition of the research team, which includes experts in the field of

entrepreneurship, provides a basis to assert the validity of the approach's results. The research question the authors attempt to answer is:

"How can no-code text mining based on abstracts be utilized efficiently to get a detailed overview of a research field?"

The structure of the exemplary case is based on the adapted MLR process displayed in Figure 2.

IV. The Application of the Adapted MLR on the Case of Social Entrepreneurship Research

In this chapter, the authors explore the application of the adapted MLR process for the case of social entrepreneurship literature from Scopus, while simultaneously providing an explanation of the text mining components utilized. The structure is based on figure 2.

A. Planning

As previously noted, the MLR approach involves an initial phase focused on identifying and justifying the review scope, formulating the review problem, and developing subsequent review questions. Given the exemplary nature of the investigation of the social entrepreneurship research field, a detailed explanation of this phase is not provided in this chapter, as it has already been broadly expounded in the preceding section.

B. Database Search

The database search was carried out in Scopus, as its multidisciplinary allows a holistic answer to the research question. In the preliminary literature search, different search terms, term combinations and synonyms were tested to identify and integrate relevant theoretical and empirical work. However, to ensure the inclusiveness and comprehensiveness of the literature search, the authors decided to use the broad search term "social entrepreneur*" rather than limiting the search to specific terms, combinations, or synonyms. This decision was made to avoid reducing the informative value of the results and to capture a wide range of relevant theoretical and empirical work, as well as to showcase the capabilities of the adapted MLR process. Thus, the search string used was the following:

Scopus: (TITLE-ABS-KEY ("social entrepreneur")*

On February 17, 2023, the search yielded 4,957 results, which were subsequently filtered to include only English language papers due to limitations in comparing multilingual works in Orange. In addition, the search was further refined by selecting journal articles only and limiting the search to papers in their final publication phase, resulting in 2,874 results. The corresponding CSV was exported, and transformed into an XLSX file, suitable for Orange. This reference list was further manually reduced by 42 due to missing abstracts and citation information. The file was then scanned on duplicates, further reducing the list by 446. The reference list therefore consisted of 2,386 journal publications.

C. Data Collection

Through copying the titles and abstracts from the source file, a second file was afterwards created. The source file contains all the meta-data, abstracts, and titles from the original data source (Scopus) without modifications, while the analysis file only includes the necessary text data (i.e., abstracts and titles) for text mining and clustering. This approach was taken based on the authors' experience and testing, which has shown that Orange achieves more precise clustering results when using only the necessary text data, while the inclusion of meta information such as authors or journal names negatively impacts the clustering process. Both the master file and analysis file are made available for easy reproduction of the results.

D. Data Cleaning

The abstracts in the analysis file were cleaned afterwards. The cleaning process included a removal of the copyright information (©*; copyright*), as well as removing structural elements in abstracts (Abstract: ; Purpose: ; Design/methodology/approach: ; Findings: ; Practical implications: ; Originality/value: ; Purpose –; Design/methodology/approach – ; Findings – ; Research limitations/implications – ; Originality/value –) to reduce noise within the data, and to align the abstracts.

E. Data Import and Preparation

The analysis file was then uploaded to Orange and used for creating a text corpus, which is a collection of text data. The corpus was then pre-processed in a second step, a process consisting of several different stages. Systematic preprocessing is necessary as it affects validity, interpretability as well as reliability of unsupervised learning models [20]. Due to the higher dimensionality of abstracts, this step significantly differs from [13] approach.
First, the corpus was tokenized, which describes the process of dividing text into word units, so called tokens. The software offers several tokenizer options, Regexp with a \w+ pattern was chosen to create tokens consisting only of alphanumeric matches. Afterwards, the tokens were converted to lowercase for term unification. Filtering was then applied to remove special characters such as periods, as well as stop words. Stop words are usually functional words with low-level information in any language, in this case English, such as articles or prepositions [21]. The researchers additionally added a specific stop word list consisting of words which appear often in scientific abstracts, such as researcher, approach, novel or paper. This list consists of 79 words in total.

Next, a term unification method was applied, specifically the Snowball stemmer [22]. Stemming is a process that involves converting the various morphological forms of a word into its base form or stem, with the assumption that each form carries the same meaning. The stem may not necessarily

correspond to an existing word in the dictionary, but all word variants should map to the same stem [23]. Thus, a stemmer can effectively reduce the number of word forms in a corpus, making it easier to analyze and process, and reducing noise in the process, increasing accuracy. The Snowball stemmer is based on the Porter stemmer by [24] and is an improved version of the original.

The last difference to the original MLR in the pre-processing was that the authors did not incorporate n-grams, a group of adjacent words in a given text that are comprised of n number of words [25], in the adapted approach. While they can be strong indicators for relevant themes, they also increase the feature space complexity and thus make the interpretation of the model more difficult. Since titles have a lower dimensionality, the addition of n-grams as an enrichment of text analysis is useful, however, this is not the case for the comparatively high-dimensional abstracts.

The third and final step in the Data Import and Preparation phase is the creation of a bag of words. Thus, a new corpus is created consisting of the entities (rows) and word counts for each data instance (columns). Rather than considering the order in which words appear in a document, a bag of words represents t he entity as a combination of series of words [26]. Unstructured text data is therefore converted to structured numeric data in the form of a vector (vectorization), which can then be used for a calculation of similarity between the different entities. As already mentioned, the adaption of the MLR utilizes the cosine similarity, which is of importance for the choice of count-based term frequency as a basis for the bag of words in this paper.

As displayed in Figure 4, the cosine similarity between two vectors is calculated using the dot product of the vectors A and B, divided by the magnitude of vector A times the magnitude of vector B. The choice of count-based term frequency in this paper over binary or sublinear term frequency is based on the necessity for computing the dot product. This is not possible with the binary-based term frequency, as it only considers whether a term appears or not in a document, without any regard for its frequency. The sublinear-based term frequency on the other hand is a type of term weighting that scales and biases the term weights to prevent over-weighting and to address the under-weighting problem and would in theory be usable. However, if the sublinear function scales down term weights excessively or if the bias term is too large, the ratios between term weights become too small, leading to the under-weighting problem [27]. Therefore, in the context of clustering, the use of sublinear term frequency is not suitable as it may lead to an ineffective representation of the data, i.e. making it difficult for the clustering to distinguish between different instances, and thus lead to unusable results.

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

Figure 3 Cosine distance formula

Furthermore, the authors used the document frequency (df) to calculate the inverse document frequency (idf) as a weighting scheme. The document frequency is a measure of the frequency of a term appearing in a collection of documents. In contrast, inverse document frequency is a measure of the importance of a term in the document. The corresponding formula can be seen in Figure 5.

$$idf(w) = log(\frac{N}{df_t})$$

Figure 4 IDF formula

Through calculating the logarithm of the ratio of the total number of documents in the collection to the number of documents containing the term, the weights of frequently appearing terms are weighted down [28]. However, when the standard IDF is used to calculate the weights, terms that appear in all documents in the collection, as compared to the documents used in the creation of this model, will have an IDF of zero, causing a division by zero error. To avoid this issue and ensure that the trained model can be used for future classification of social entrepreneurship abstracts, it is necessary to use a smoothed version of IDF. The implementation of IDF in Orange is based on the TfidfTransformer from scikit-learn [29], which uses a modified version of the IDF formula, as seen in Figure 6.

$$idf(w) = [log(\frac{N}{dft})] + 1$$

Figure 5 IDF formula utilized in Orange [29]

The addition of 1 in the equation ensures that terms that occur in all documents do not get a weight of 0, as this can cause issues with further calculations, as explained above. While this solves the division by zero problem, it can still result in a situation where the weights of the more frequent terms are not effectively reduced. Therefore, the researchers decided to base their bag of words model on the smooth IDF calculation, as seen in Figure 7.

$$idf(w) = [log(\frac{1+N}{1+dft})] + 1$$

Figure 6 Smoothed IDF formula utilized in Orange [29]

This formula has a plus 1 on both the numerator and denominator, leading to more effective reduction of weights of frequent terms. This resulted in a more accurate and meaningful ranking of search results, making it a better choice for the clustering of scientific abstracts in this case.
Finally, the authors introduce regularization, as it helps with reducing the problem of overfitting [30]. It does so by penalizing unnecessary model complexity and focusing on the most relevant features. The two most commonly used methods are L1 regularization, which adds the sum of the absolute values of the model parameters to the objective function, and L2 regularization, which adds the sum of the squares of the parameters [31]. Both are available in Orange. The formula for the regularization penalty term of L1 is displayed in Figure 8, while the same for L2 is displayed in Figure 9.

$$L1Regularization : \lambda \sum_{j=1}^{p} |\beta_j|$$

Figure 7 L1 Regularization penalty term formula

$$L2Regularization : \lambda \sum_{j=1}^{p} \beta_j^2$$

Figure 8 L2 Regularization penalty term formula

Therefore, L1 is able to drive some parameters to exactly zero, thus possibly reducing the dimensionality of the problem, while the L2 regularization tends to spread the error across all parameters, keeping the dimensionality [31]. As the authors decided against a reduction of dimensionality, L2 regularization was implemented to reduce overfitting.

F. Data Extraction and Visualization

Based on the bag of words vectorization, the authors calculated a distance metric needed for the creation of a hierarchical clustering. As already mentioned, this adapted MLR approach utilizes the cosine similarity as distance metric for the hierarchical clustering, rather than the Euclidean distance. The Euclidean distance is displayed in Figure 8, while the later can be seen in Figure 5.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Figure 9 Euclidean distance formula

In the Euclidean distance formula, the distance between two points is calculated based on the absolute differences between their corresponding values in each dimension. These dimensions are represented by the different tokens in text mining, thus the number of dimensions is equal to the size of the vocabulary utilized. This means that the more terms a document has, the higher the dimensionality of its feature space. However, if the dimensionality of the feature space increases, the Euclidean distance becomes less effective [32]. This so called “curse of dimensionality” leads to distances between points becoming less meaningful with an increase of dimensions, which tend to converge to a constant value. Additionally, if the length of documents varies, which oftentimes is the case for abstracts, the Euclidean distance calculation is furthermore affected. The result is that documents with more token will have higher values in each dimension, thus making them appear more distant from the documents with fewer words. This, however, does not necessarily indicate that they are semantically or conceptually distant, i.e. them not belonging to the same thematic cluster. Thus, the Euclidean distance is suitable for the clustering of titles, but not for abstracts.

The cosine similarity on the other hand, as displayed in Figure 4, is based on an angle of two vectors, rather than the absolute difference between their corresponding values in each dimension, leading to a robustness to the curse of dimensionality, which means that it can handle the high-dimensionality feature space typical for text data. Additionally, due to the measurement of similarity through the orientation between two vectors not being affected by the length of vectors, it is specifically useful for text mining, and does not have the same shortcomings as Euclidean distance. Therefore, the cosine similarity is the metric of choice for the adapted MLR approach.

After the distance metric was calculated, the hierarchical clustering was carried out. Hierarchical clustering, as explained by [33] is a data mining technique that groups data objects into a tree of clusters. There are two approaches, agglomerative, where clusters are merged starting from the bottom, or a divisive, where clusters are divided starting from the top. The hierarchical clustering algorithm creates a sequence of nested partitions with a single, all-inclusive cluster at the top and individual objects at the bottom. Each intermediate level of the hierarchy combines two clusters from the lower level or splits a cluster from the higher level. The result can be displayed as a dendrogram, which graphically shows the merging process and the intermediate clusters, and how points can be merged into a single cluster. In Orange, hierarchical clustering is based on the agglomerative approach.

A prerequisite for hierarchical cluster analysis is a linkage function, which is a measure of the distance between two groups of objects, i.e. between two clusters [34]. There are, however, different options for linkage functions, which each affect the result by utilizing different distance calculations for each subset. Single-linkage uses the shortest distance between two subsets, average-linkage uses the average distance between them, and complete-linkage uses the largest distance between them [33]. Additionally, there is the Ward linkage, which however mathematically requires Euclidean distance as the utilized distance metric [35], and is thus not suitable for this use case. However, there is no universally optimal choice as each linkage function has their drawbacks [33]. Thus, researchers often have to utilize different techniques and compare their results to determine the most appropriate linkage function for their specific dataset and clustering objective [36].

After carefully comparing the different results, the researchers decided to use the complete linkage for the clustering of the social entrepreneurship abstracts. Its function as displayed in Figure 11.

$$d(C_i, C_j) = \max_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

Figure 10 Complete linkage formula

The main reasons for this choice were that the resulting clusters were well separated and distinguishable in the dendrogram, as well as the results of the silhouette plot. Additionally, the complete linkage was able to capture the global structure of the data and produce clusters that are compact and have a clear separation from one another, which was important for the interpretability of the results.

As the final step of this phase, a silhouette plot was created. The silhouette plot is a valuable tool for evaluating

the quality of the clustering result [37]. It is based on a calculation of a silhouette score for each object, which measures how well it is assigned to its cluster as compared to other clusters. Objects with a high silhouette score are considered well-clustered, while objects with low scores may not belong to their assigned cluster. The plot shows the distribution of silhouette scores for all objects in the clustering, providing a visual representation of the clustering quality. By analyzing the silhouette plot, the researcher can determine whether the chosen number of clusters is appropriate and adjust it if necessary. The authors used the height ratio of the dendrogram, a threshold at which the dendrogram is cut to form clusters, to create a varying number of clusters, and then compared the resulting numbers of clusters, average silhouette scores and median silhouette scores to determine an appropriate number of clusters for the social entrepreneurship abstracts. The calculation of the metrics was done through exporting the silhouette plot results through the Data Table widget, and a subsequent calculation in an adequate program. It was observed that these metrics increased significantly until 346 clusters were formed, corresponding to a dendrogram height of 96.5 percent. Subsequently, a decrease in the height ratio only led to a minor improvement in the metrics, with the number of clusters continuing to increase but the cluster size gradually diminishing in terms of the number of documents per cluster.

G. Data Interpretation

The clusters are subsequently analyzed for clarity and research object. The respective clarities are derived based on the silhouette plot. Clear clusters consist of > 70 percent positive scored paper, clear-fuzzy clusters are mixed (< 70 %, > 50 %), and fuzzy clusters consist of < 50 percent positive scored papers. Out of a total of 346 clusters, 243 are determined to be clear, while 43 are identified as clear-fuzzy, and the remaining 60 are classified as fuzzy. The fuzzy clusters, which encompass a total of 417 papers, are subsequently excluded from the analysis, as the resulting research object would not accurately describe at least 50 percent of the documents within the clusters and would thus reduce the significance of the example. The researchers utilize a multi-faceted approach to extract the research objects from the clusters, using the select rows function of Orange. They employ a title-scanning method based on the Data Table widget, which proved effective in approximately 80% of the cases, the Word Cloud widget for 15% of the remaining clusters, and again scan abstracts for the final 5% using the Data Table widget. However, 13 clusters are excluded because the authors were unable to identify a theme using any of these methods. In total, the analysis includes 1,911 papers divided into 274 clusters. Of these, 173 clusters consist of at least four papers, with an average cluster size of 9, while 101 clusters consist of three papers or less, with an average cluster size of 2.5. Therefore, the approach is successful in identifying topics at a very granular level. The research objects are further categorized into 17 different groups to provide a more comprehensive representation of the results. They are displayed, based on the number of documents within each group, in Figure 12.

Figure 11 Social entrepreneurship literature groups

The focus of the largest group of documents of the social entrepreneurship literature can be assigned to the group of sustainability. The clusters are located within all three pillars of sustainability, i.e. environment (e.g. C208 Country emissions), social responsibility (e.g. C309 Philanthropy), and economic (e.g. C15 Solidarity Economy). Other groups consist, for example, of clusters that focus on social entrepreneurship in the context of different regions (Group 6 Regions), or of work that focus with the necessary competencies and educational opportunities for social entrepreneurs (Group 9 Competencies and Education). To exemplify the fine granularity of the approach, the authors show a detailed breakdown of the clusters that together form Group 8 Diversity and Inclusion in Figure 13.

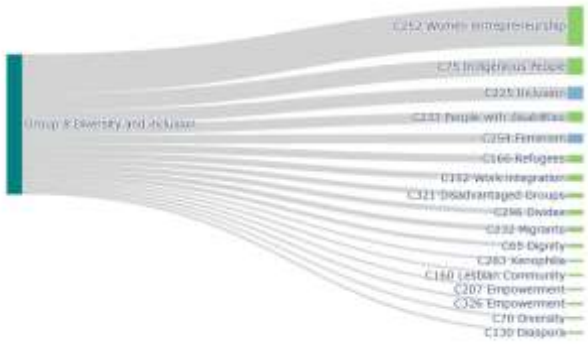

Figure 12 Detailed view of Group 8 Diversity and Inclusion

This chart visualizes the distribution of clear research object clusters related to diversity and inclusion, represented by the green nodes, and their connection to the group label, Group 8 Diversity and Inclusion, represented by the teal node at the top. The blue nodes represent research object clusters that are considered clear-fuzzy. In this chart, the blue nodes are “C254 Inclusion” and “C70 Feminism”. The size of each node corresponds to the amount of research documents related to that research object. The largest node is 'Women entrepreneurship', which has 30 research documents, and the smallest node is 'Diaspora', which has only 2 research documents.

The complete workflow in Orange can be seen in Figure 14.

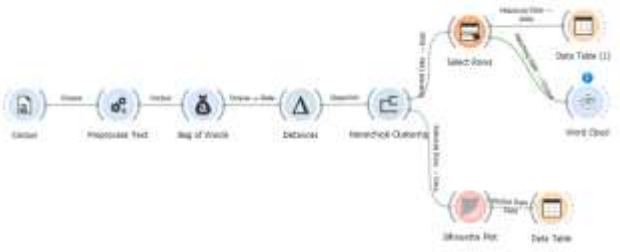

Figure 13 Adapted MLR approach workflow in Orange

V. CONCLUSIONS AND LIMITATIONS

In this paper, the authors introduce a no-code text mining approach adapted from the MLR approach by [13], which offers a new way to explore and analyze research literature. It is designed to create a fine-granular thematic clustering of a chosen research field based on scientific titles and abstracts, which can afterwards be further analyzed and categorized to provide a detailed overview of the literature on a given topic. To demonstrate the capabilities of the approach, the authors apply it to the research field of social entrepreneurship, an area of growing interest in recent years.

The proposed methodology involves the creation of a corpus comprising abstracts and titles after the original data has undergone cleaning. Subsequently, the text undergoes preprocessing, wherein the authors advocate the use of a regular expression with a \w+ pattern to tokenize the data, followed by conversion to lowercase, removal of stop words (including both common English words and a specific list), and normalization using the snowball stemmer. Next, a bag of words is constructed using term frequency as the basis for count, smooth IDF as document frequency, and L2 regularization. Cosine distance is then computed between different instances of the data set, which are hierarchically clustered using a complete linkage function. The number of clusters is determined by comparing various cluster height ratios based on cluster quantity, average silhouette score, and median silhouette score. In the final step, the clusters are themed using a three-step process that involves title scanning, word clouds, and abstract scanning. In the given example, the authors clustered 2,386 papers into 346 thematic clusters and subsequently categorized them into 17 distinct groups, reflecting various areas of focus within the literature such as sustainability and diversity and inclusion.

The authors highlight the benefits of the approach, particularly for new researchers, who can quickly and easily recognize what literature is already available and support interest finding for future research, which helps them tackle the evermore increasing amount of literature. However, the fine granularity can also be beneficial for established researchers, as it provides an objective approach that helps remove blinders, leading to the recognition of potentials for trans- and interdisciplinary approaches. Additionally, text mining makes it easier for researchers to sort relevant and irrelevant literature, saving time and resources.

Despite the potential of the no-code text mining approach, some limitations need to be acknowledged. For example, that the unsupervised machine learning algorithm hierarchical clustering is not customizable, meaning that it may cluster over some meaningless words that the program cannot interpret. Moreover, no-code is generally not customizable, which limits its flexibility. Finally, abstracts are not as meaningful as whole texts, which can limit the scope of analysis.

Nevertheless, the no-code text mining approach presents a promising method for researchers to explore existing literature and identify potential research avenues. As researchers continue to explore new techniques and approaches to analyzing research literature, the no-code text mining approach is a valuable addition to the research methodology toolkit. With further development and refinement, the approach has the potential to make significant contributions to the field of research methodology, particularly in supporting the identification of research gaps and new directions.

VI. MATCH AND CONTRIBUTION

The paper fits the call for papers and the field of interest of the IEEE Technology & Engineering Management Society because it utilizes innovative technology to help scientists become more productive and furthermore discusses sustainable entrepreneurship.

[1] S. Kunisch, M. Menz, J. M. Bartunek, L. B. Cardinal, and D. Denyer, "Feature Topic at Organizational Research Methods: How to Conduct Rigorous and Impactful Literature Reviews?," Organizational Research Methods, vol. 21, no. 3, pp. 519–523, Juli 2018, doi: 10.1177/1094428118770750.

[2] A. Booth, D. Papaioannou, and A. Sutton, Systematic Approaches to a Successful Literature Review. 2012.

[3] D. Tranfield, D. Denyer, and P. Smart, "Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review," Br J Management, vol. 14, no. 3, pp. 207–222, September 2003, doi: 10.1111/1467-8551.00375.

[4] J. L. Callahan, "Writing Literature Reviews: A Reprise and Update," Human Resource Development Review, vol. 13, no. 3, pp. 271–275, September 2014, doi: 10.1177/1534484314536705.

[5] D. Denyer, "Doing a Literature Review.," presented at the Presentation to the British Academy of Management Doctoral Symposium, Bristol, UK.

[6] W. R. Schumm and D. W. Crawford, "Evaluating the Quality of Literature Reviews in the Social Sciences: Developing a Measure of Quality with an Illustration," PRA, vol. 1, no. 2, June 2019, doi: 10.22606/pra.2019.12003.

[7] L. Feng, Y. K. Chiam, and S. K. Lo, "Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review," in 2017 24th Asia-Pacific Software Engineering Conference (APSEC), Nanjing, December 2017, pp. 41–50. doi: 10.1109/APSEC.2017.10.

[8] L. Orgeolet et al., "Can artificial intelligence replace manual search for systematic literature? Review on cutaneous manifestations in primary Sjögren's syndrome," Rheumatology, vol. 59, no. 4, pp. 811–819, April 2020, doi: 10.1093/rheumatology/kez370.

[9] A. Korhonen, D. Ó Séaghdha, I. Silins, L. Sun, J. Högberg, and U. Stenius, "Text Mining for Literature Review and Knowledge Discovery in Cancer Risk Assessment and Research," PLoS ONE, vol. 7, no. 4, p. e33427, April 2012, doi: 10.1371/journal.pone.0033427.

[10] D. Yang and A. N. Zhang, "Performing literature review using text mining, Part III: Summarizing articles using TextRank," in 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, December 2018, pp. 3186–3190. doi: 10.1109/BigData.2018.8622408.

[11] K. R. Felizardo, E. F. Barbosa, R. M. Martins, P. H. D. Valle, and J. C. Maldonado, "Visual Text Mining: Ensuring the Presence of Relevant Studies in Systematic Literature Reviews," Int. J. Soft. Eng. Knowl. Eng., vol. 25, no. 05, pp. 909–928, Jun. 2015, doi: 10.1142/S0218194015500114.

[12] N. Kling, C. Runte, S. Kabiraj, and C.-A. Schumann, "Harnessing Sustainable Development in Image Recognition Through No-Code AI Applications: A Comparative Analysis," in Recent Trends in Image Processing and Pattern Recognition, vol. 1576, K. Santosh, R. Hegadi, and U. Pal, Eds. Cham: Springer International Publishing, 2022, pp. 146–155. doi: 10.1007/978-3-031-07005-1_14.

[13] K. Reuther, “A Systems Theory Perspective of Interconnected Influence Factors on Front-End innovations: The role of Organisational Structures,” University of West of Scotland, 2019.

[14] A.-M. Nitsche, C.-A. Schumann, B. Franczyk, and K. Reuther, “Mapping supply chain collaboration research: a machine learning-based literature review,” International Journal of Logistics Research and Applications, pp. 1–29, November 2021, doi: 10.1080/13675567.2021.2001446.

[15] J. Demsar et al., “Orange: Data Mining Toolbox in Python,” Journal of Machine Learning Research, no. 14, pp. 2349–2353, 2013.

[16] H. Cooper, Synthesizing research: A guide for literature reviews, 3rd ed. Thousand Oaks, CA, US: Sage Publications, Inc, 1998, pp. xii, 201.

[17] R. B. Briner and D. Denyer, “Systematic Review and Evidence Synthesis as a Practice and Scholarship Tool,” in The Oxford Handbook of Evidence-Based Management, 1st ed., D. M. Rousseau, Ed. Oxford University Press, 2012, pp. 112–129. doi: 10.1093/oxfordhb/9780199763986.013.0007.

[18] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, “A Comparison of Semantic Similarity Methods for Maximum Human Interpretability,” 2019, doi: 10.48550/ARXIV.1910.09129.

[19] L. Zahrotun, “Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method,” ComEngApp, vol. 5, no. 1, pp. 11–18, February 2016, doi: 10.18495/comengapp.v5i1.160.

[20] M. J. Denny and A. Spirling, “Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It,” Polit. Anal., vol. 26, no. 2, pp. 168–189, April 2018, doi: 10.1017/pan.2017.44.

[21] D. Maier et al., “Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology,” Communication Methods and Measures, vol. 12, no. 2–3, pp. 93–118, April 2018, doi: 10.1080/19312458.2018.1430754.

[22] M. F. Porter, “Snowball: A language for stemming algorithms,” 2001.

[23] A. Jivani, “A Comparative Study of Stemming Algorithms,” Int. J. Comp. Tech. Appl., vol. 2, pp. 1930–1938, 2011.

[24] M. F. Porter, “An algorithm for suffix stripping,” Program, vol. 14, no. 3, pp. 130–137, March 1980, doi: 10.1108/eb046814.

[25] M. Schonlau, N. Guenther, and I. Sucholutsky, “Text Mining with n-gram Variables,” The Stata Journal, vol. 17, no. 4, pp. 866–881, December 2017, doi: 10.1177/1536867X1801700406.

[26] J. Wang and Y. Dong, “Measurement of Text Similarity: A Survey,” Information, vol. 11, no. 9, p. 421, August 2020, doi: 10.3390/info11090421.

[27] H. Wu and X. Gu, “Balancing Between Over-Weighting and Under-Weighting in Supervised Term Weighting,” 2016, doi: 10.48550/ARXIV.1604.04007.

[28] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval. New York: Cambridge University Press, 2008.

[29] scikit-learn, “TfidfTransformer.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[30] K. Skianis, F. Rousseau, and M. Vazirgiannis, “Regularizing Text Categorization with Clusters of Words,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 2016, pp. 1827–1837. doi: 10.18653/v1/D16-1188.

[31] O. Demir-Kavuk, M. Kamada, T. Akutsu, and E.-W. Knapp, “Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features,” BMC Bioinformatics, vol. 12, no. 1, p. 412, December 2011, doi: 10.1186/1471-2105-12-412.

[32] S. Xia, Z. Xiong, Y. Luo, WeiXu, and G. Zhang, “Effectiveness of the Euclidean distance in high dimensional spaces,” Optik, vol. 126, no. 24, pp. 5614–5619, December 2015, doi: 10.1016/j.ijleo.2015.09.093.

[33] Y. Rani and D. H. Rohil, “A Study of Hierarchical Clustering Algorithm,” 2013.

[34] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of Document Clustering Techniques,” in Proceedings of the International KDD Workshop on Text Mining, 2000.

[35] S. Miyamoto, R. Abe, Y. Endo, and J. Takeshita, “Ward method of hierarchical clustering for non-Euclidean similarity measures,” in 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Fukuoka, Japan, November 2015, pp. 60–63. doi: 10.1109/SOCPAR.2015.7492784.

[36] O. Yim and K. T. Ramdeen, “Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data,” TQMP, vol. 11, no. 1, pp. 8–21, February 2015, doi: 10.20982/tqmp.11.1.p008.

[37] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, November 1987, doi: 10.1016/0377-0427(87)90125-7.