

**VILNIUS GEDIMINAS TECHNICAL UNIVERSITY**

**FACULTY OF FUNDAMENTAL SCIENCES**

**DEPARTMENT OF INFORMATION SYSTEMS**

**Mahesh Kumar Pratihar (20234738)**

**Textual Data Clusterization Based on Self-organizing Map and Word Embeddings**

**Master Graduation Thesis**

**Supervised by: Assoc. Prof. Dr. Pavel Stefanovič**

VILNIUS 2025

**CONTENT**

[Introduction 5](#_Toc189422202)

[Investigation Object 7](#_Toc189422203)

[The Aim and Tasks of the Thesis 7](#_Toc189422204)

[Novelty of the Topic 7](#_Toc189422205)

[Relevance of the Topic 8](#_Toc189422206)

[Research Methodology 8](#_Toc189422207)

[Scientific Value of the Thesis 8](#_Toc189422208)

[Main Results of the Thesis 9](#_Toc189422209)

[Structure of the Work 9](#_Toc189422210)

[1. Related work Analysis 10](#_Toc189422211)

[1.1 Main Concepts 10](#_Toc189422212)

[1.2 Related works on Improving the Text Quality using SOMs and Word embeddings. 11](#_Toc189422213)

[2. Conclusion 30](#_Toc189422214)

[3. References 30](#_Toc189422215)

**List of Images**

[Figure 1:A simple Kohonen network (Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.) 10](#_Toc189571270)

[Figure 2: Simple SOM View 1 (Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.) 11](#_Toc189571271)

[Figure 3: Simple SOM View 2 (Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.) 11](#_Toc189571272)

[Figure 4: SOM weights (Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.) 12](#_Toc189571273)

[Figure 5: BMU neighbourhood (Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.) 12](#_Toc189571274)

**List of Tables**

[Table 1: Data Extraction Template 19](#_Toc189571385)

|  |  |  |
| --- | --- | --- |
|  |  |  |

**Abbreviations**

The abbreviations used in this document are described as follows:

**SOM**: Self organizing maps

**LaRoSeDa:** Large Romanian Sentiment Data.

**BERT**: Bidirectional Encoder Representations from Transformers

**TF**: Term Frequency

**IDF**: Inverse Document Frequency

**HISK:** Histogram Intersection String Kernel

**BOWE**: Bag-of-word-embeddings

**NLP**: Natural Language Processing.

**LSA**: Latent Symantic Analysis

**GANs**: Generative Adversarial Nets

**SOMGAN:** Self-Organizing Map Generative Adversarial Nets

**DESOM:** Deep Embedded Self-Organizing Map

# **Introduction**

The problem of text quality presents a considerable challenge in Natural Language Processing (NLP), impacting various tasks like text classification, clustering, and sentiment analysis. Ensuring high-quality text data is vital for achieving accurate outcomes in these applications. Often, poor text quality stems from problems such as grammatical mistakes, informal phrases, ambiguity, and noise within the data. Furthermore, numerous text datasets experience high dimensionality and sparsity, complicating the extraction of meaningful patterns. For example, when addressing large-scale text clustering, conventional clustering techniques such as K-means, hierarchical clustering, and density-based clustering encounter difficulties due to the high-dimensional characteristics of textual data (Araújo et al., 2020). These methodologies frequently struggle to produce precise clusters as they cannot effectively manage sparse matrices and the semantic relationships among words. Therefore, improving text quality is essential for enhancing the efficiency and effectiveness of NLP models.

Addressing text quality issues is vital for a range of real-world applications. In sentiment analysis, for instance, the precision of sentiment classification significantly relies on the quality of the input text. If a dataset comprises noisy or ambiguous text, the model may incorrectly classify sentiments, resulting in misinterpretations of user opinions. Likewise, in text clustering, inadequate text quality can lead to improper grouping of documents, diminishing the efficiency of document organization systems. Another critical domain where text quality is significant is information retrieval. Search engines depend on high-quality text representations to deliver relevant search results, and if the text is ambiguous or erroneous, the retrieval process can become ineffective. Machine translation also requires high-quality text, as low-quality inputs may produce inaccurate or misleading translations. Given these dependencies on high-quality textual data, researchers have devised various text representation strategies to address the challenges posed by poor-quality text.

Initial methods for tackling text quality concerns predominantly utilized statistical techniques such as the Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) models (Mehta et al., 2021). BoW represents text by transforming words into numerical features based on their occurrences, while TF-IDF enhances this by weighing words according to their significance in a document corpus. Although these approaches established the foundation for text representation, they exhibited several shortcomings. A major limitation was the loss of contextual meaning, as these models treated words separately, disregarding their relationships within sentences. Additionally, the high dimensionality resulting from these representations rendered them computationally intensive. Moreover, they did not resolve the polysemy issue, where a single word may have multiple meanings depending on its context.

The proposed solution to address some of these challenges, researchers developed n-gram models, which take into account sequences of words instead of treating individual words independently. These models improved contextual comprehension by partially capturing word dependencies. However, they still faced dimensionality issues, as increasing the n-gram size led to an exponential rise in the number of features. Furthermore, n-gram models struggled to effectively capture long-range dependencies in text, rendering them inadequate for complex NLP tasks. This propelled the advancement of word embedding techniques such as Word2Vec and GloVe (Ravi & Kulkarni, 2023), which represented words as dense vectors in a continuous space. These embeddings enabled models to grasp semantic similarities between words, addressing some limitations of previous methods. Nonetheless, despite their benefits, Word2Vec and GloVe came with their own set of challenges. One of the most significant shortcomings was their failure to produce context-sensitive representations. These models assigned a fixed vector to each word, meaning that the same representation was applied, regardless of the context in which the word appeared. Consequently, they struggled to differentiate between various meanings of the same word.

This thesis explores how SOMs and word embeddings can improve the clustering of textual data within the context of education. Word embeddings are numerical representations of words that encode their meanings and relationships in dense vectors, enabling semantic understanding of textual data. Advanced models like Word2Vec, GloVe, and especially BERT (Bidirectional Encoder Representations from Transformers) have revolutionized natural language processing by offering context-aware embeddings (Tache et al., 2021)​​. BERT stands out as it captures the nuances of words based on their surrounding context, rather than assigning a fixed meaning to each word. This dynamic approach is critical for understanding complex educational content and drawing connections across diverse learning materials. SOMs are a type of artificial neural network introduced by Teuvo Kohonen, designed specifically for clustering and visualization. They map high-dimensional data onto a two-dimensional grid while preserving the structural relationships within the data. Unlike traditional clustering methods such as k-means, SOMs effectively identify and preserve patterns, making them particularly well-suited for handling the complexity of textual data in education (Tache et al., 2021)​. Educational content, characterized by domain-specific language, diverse subject matter, and varied data types, presents unique challenges for clustering. By focusing on the field of education, this study applies SOMs and word embeddings to organize learning resources, categorize academic texts, and discover patterns in research publications. English datasets are utilized in this research for their richness in linguistic diversity and their relevance to global educational platforms, facilitating robust analysis and comparisons (Ravi & Kulkarni, 2023)​.

The key issue addressed in this research is the inability of traditional clustering techniques to manage the high dimensionality of textual data and linguistic intricacies, such as polysemy (words with multiple meanings) and synonymy (different words with similar meanings). SOMs, when paired with word embeddings, offer a powerful solution by capturing these linguistic complexities while preserving the data's topological properties​. This research is novel in applying SOMs specifically for clustering word embeddings in the domain of education, which remains an underexplored area. By integrating SOMs pattern-recognition capabilities with BERT’s advanced semantic understanding, this study aims to develop efficient and meaningful clusters of educational texts. The outcomes could advance text mining research and benefit various applications in the educational sector, such as adaptive learning systems, automated grading tools, and content curation platforms.

## **Investigation Object**

The investigation object of this research is the clustering of textual data in the educational domain using Self-Organizing Maps (SOMs) and word embeddings.

## **The Aim and Tasks of the Thesis**

The research aims to improve text clustering by addressing high dimensionality and enhancing contextual meaning to improve text quality.

To achieve this aim, the following tasks are identified:

1. To analyze related works on unstructured data classification in the context of clustering techniques and embedding methods.
2. To propose an approach for classifying textual data using SOMs and word embeddings.
3. To implement the proposed approach as a prototype and evaluate its efficiency.

## **Novelty of the Topic**

The novelty of this topic lies in its innovative use of self-organizing maps (SOMs) combined with advanced word embedding techniques to tackle the complex challenges of clustering textual data (Mehta et al., 2021; Tache et al., 2021). Unlike traditional clustering methods that often struggle with high dimensionality and lack contextual understanding, this approach integrates the semantic richness of word embeddings with the adaptability of SOMs to deliver more accurate and contextually meaningful groupings. By focusing on Hindi-language datasets, this research addresses the underexplored area of clustering in non-English languages, contributing valuable insights to multilingual natural language processing(Ravi & Kulkarni, 2023). The work stands out for its potential to refine clustering techniques, making them more effective for large, diverse, and semantically complex textual corpora.

## **Relevance of the Topic**

The problem of clustering textual data efficiently while addressing semantic complexity and high dimensionality remains unresolved. Conventional methods like K-means and TF-IDF struggle with sparsity, polysemy, and synonymy, particularly when dealing with large-scale datasets (Mehta et al., 2021; Ravi & Kulkarni, 2023). Emerging techniques, such as leveraging advanced models like BERT for context-sensitive word embeddings, have shown promise by improving clustering accuracy and scalability (Mehta et al., 2021; Tache et al., 2021). Despite progress, few studies explore the integration of clustering innovations, such as Self-Organizing Maps (SOMs), with word embeddings for more nuanced semantic grouping, leaving a gap for scalable, context-aware clustering solutions (Tache et al., 2021). This thesis seeks to bridge this gap by developing a novel methodology for clustering text data using advanced embedding techniques, contributing to both theory and practice in large-scale text mining.

## **Research Methodology**

This research employs a systematic and experimental approach to enhance text clustering using SOMs and word embeddings. A thorough literature review identifies limitations of traditional clustering methods and opportunities for improvement with SOMs and embeddings like Word2Vec, GloVe, and BERT. A labelled English-language dataset is selected and pre-processed with tokenization, stop word removal, and stemming to create semantically rich vector representations. SOMs are implemented for clustering, with experiments conducted on parameters such as learning rates, map dimensions, and neighbourhood functions. The performance of SOMs is compared to K-means and hierarchical clustering using metrics like cluster purity and adjusted Rand index. Dimensionality reduction techniques, such as latent semantic analysis, are explored to optimize scalability. Iterative refinement ensures improved clustering by validating results against dataset labels and using evaluation metrics like silhouette scores and precision-recall. Results are benchmarked against traditional clustering approaches to assess robustness and generalizability.

## **Scientific Value of the Thesis**

The scientific value of this thesis lies in its contribution to improving how textual data is grouped and understood using modern techniques like self-organizing maps (SOMs) and word embeddings. Unlike traditional clustering methods, this research focuses on capturing the context and meaning of words, which is especially important for languages like Hindi that are less represented in current studies. By exploring and refining these advanced methods, the thesis provides a better way to handle large and complex datasets, making it easier to identify meaningful patterns and relationships. The findings not only help improve text analysis but also open the door for further research and practical applications in areas like multilingual language processing and big data analysis.

## **Main Results of the Thesis**

Analyzing Self-Organizing Maps (SOMs) and word embeddings like BERT shows how they can work together to enhance text quality. SOMs are great for clustering because they take complex, high-dimensional data and map it onto a simpler, lower-dimensional grid while keeping the data’s structure intact. This makes it easier to spot patterns, group similar data, and identify outliers, leading to more meaningful clusters. Meanwhile, BERT excels at capturing the context and meaning of words within sentences, providing rich, nuanced embeddings. It also helps reduce dimensionality without losing important details. When used together, SOMs and BERT complement each other—SOMs handle clustering effectively, and BERT ensures the data retains deep semantic meaning. This combination results in better text representation, more accurate clustering, and overall improved text quality.

## **Structure of the Work**

The introduction section highlights the significance of text clustering for handling high-dimensional textual data in under-researched domains like Hindi, focusing on the potential of self-organizing maps (SOMs) and word embeddings for context-based clustering.

The literature review section reviews traditional clustering methods and embedding models, evaluates SOM's strengths and weaknesses, and identifies gaps in existing approaches for semantic clustering.​

# **Related work Analysis**

## **Main Concepts**

For the main concepts of this research includes SOMs and word embeddings. All the related papers are chosen based on the SOMs and word embeddings related work. They have used hybrid approaches using SOMs and embeddings or comparing traditional clustering techniques with new techniques.

Unstructured textual data, such as reviews, news articles, and social media posts, lacks a predefined structure, making it inherently difficult to analyze and cluster effectively (Ravi & Kulkarni, 2023). In the domain of Natural Language Processing (NLP), clustering this data poses significant challenges, including high dimensionality, semantic ambiguity, and sparse representations (Stefanovič & Kurasova, 2022). Traditional clustering methods, such as K-means and hierarchical clustering, struggle to handle the complexities of large-scale datasets, especially in capturing semantic relationships between words​​. Emerging approaches that combine Self-Organizing Maps (SOMs) with word embeddings have demonstrated substantial potential to address these issues by offering dimensionality reduction, context-sensitive clustering, and superior topological preservation​.

Word embeddings, such as those generated by Word2Vec, GloVe, and BERT, encode words as dense vectors in a high-dimensional space, capturing semantic and contextual nuances. BERT excels at generating context-aware embeddings by considering the bidirectional context of words, making it highly effective for clustering tasks​​. For instance, (Mehta et al., 2021) introduced WEClustering, which integrates BERT embeddings with statistical models to enhance clustering quality by mitigating the curse of dimensionality and improving semantic coherence​. Similarly, (Ravi & Kulkarni, 2023; Rejeb et al., 2022) showcased BERT's superiority over traditional embedding models like TF-IDF in clustering Twitter data, achieving high accuracy and semantic alignment using metrics such as Purity and Adjusted Rand Index (ARI).

SOMs, on the other hand, are neural network-inspired clustering tools that preserve the topological structure of data while providing visualization capabilities. Unlike traditional methods like K-means, SOMs model data density effectively, enabling more natural cluster formation. (Tache et al., 2021) demonstrated the utility of SOMs when paired with advanced embeddings, showing improved alignment with Zipf's law—a distribution characteristic of natural language—leading to better cluster coherence​. Furthermore, combining SOMs with techniques like Latent Semantic Analysis (LSA) has shown promise in reducing dimensionality without losing semantic depth​​.

The aim of SOM is to offer a data visualization method that aids in comprehending high-dimensional data by lowering its dimensions for mapping. Additionally, SOM embodies the clustering idea by aggregating similar data points (Paper et al., 1990). As a result, it can be stated that a Self-Organizing Map minimizes the dimensionality of data and visually represents items that are similar.

Architecture of self-organizing maps, we simplicity we focus on a two-dimensional Self-Organizing Map (SOM). The network consists of a two-dimensional arrangement of ‘nodes’, where each node is completely linked to the input layer. The figure 1 below illustrates a compact Kohonen network comprising 4 X 4 nodes that are connected to the input layer (depicted in green) representing a two-dimensional vector.

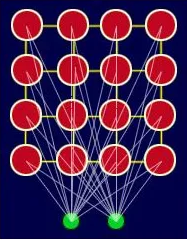


Figure 1:A simple Kohonen network **(Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.)**

Every node is assigned a distinct topological location (defined by an x, y coordinate in the lattice) and holds a weight vector that matches the dimensionality of the input vectors. In other words, if the training dataset comprises vectors, V, with n dimensions:

V1, V2, V3…Vn

Then each node will contain a corresponding weight vector *W,*of *n*dimensions:

W1, W2, W3…Wn

The lines linking the nodes in the figure above solely illustrate adjacency and do not indicate a connection, as typically described in the context of a neural network. There are no lateral connections among nodes within the lattice.

A Self-Organizing Map (SOM) operates without the need for a predefined target output, in contrast to many other neural network types. Rather, when the weights of a node align with the input vector, that specific region of the lattice is fine-tuned to better represent the data corresponding to the class of the input vector. Starting from a randomly assigned weight distribution, and through numerous iterations, the SOM ultimately converges into a map featuring stable areas. Each area functions as a classifier for certain features, allowing you to view the graphical output as a form of feature map for the input space.

How the SOMs learn? We can have a simple self-organizing map. We have three input vectors which mean three features, and we have nine output vectors. Each input is like a column(dimension) in a data set and there might be thousands of rows. So, data could be size 1000 X 3. The output nodes are always two-dimensional. Figure 2 shows a simple SOM view, with three inputs and 9 outputs.

A diagram of a network

Description automatically generated

Figure 2: Simple SOM View 1 **(Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.)**

Figure 3 shows another way visualizing the SOM, which is more understandable.

A diagram of a network

Description automatically generated

Figure 3: Simple SOM View 2 **(Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.)**

Example X1 = 0.7, X2 = 0.6 and X3 = 0.9

The training of SOMs occurs in several steps and over many iterations:

1. Each node’s weights are initialized.

2. A vector is chosen at random from the set of training data and presented to the lattice.

3. Every node is examined to calculate which one’s weights are most like the input vector. Figure 4 shows the winning node is commonly known as the Best Matching Unit (BMU) calculated using Euclidean distance.

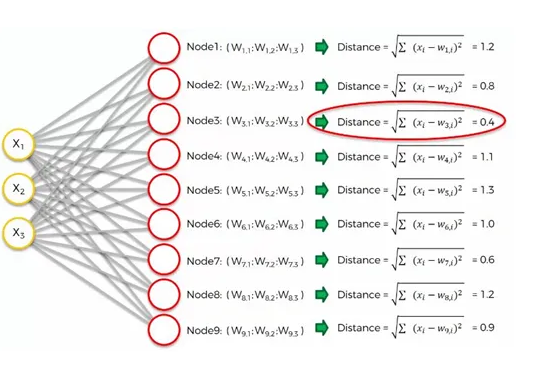


Figure 4: SOM weights **(Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.)**

1. The radius of the neighborhood of the BMU is now calculated. This is a value that starts large, typically set to the ‘radius’ of the lattice, but diminishes each time-step. Figure 5 shows that any node found within this radius are deemed to be inside the BMU’s neighborhood.

A red circle with black dots

Description automatically generated

Figure 5: BMU neighbourhood **(Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium, n.d.)**

5. Each neighboring node’s (the nodes found in step 4) weights are adjusted to make them more like the input vector. The closer a node is to the BMU; the more its weights get altered.

6. Repeat step 2 for N iterations.

Clustering unstructured textual data presents challenges such as high dimensionality, semantic complexity, and sparse representations, making traditional methods like K-means inefficient. Implementing SOMs requires careful tuning of parameters (e.g., grid size, learning rates) and is sensitive to noise in sparse datasets. Embedding models like BERT, while context-rich, are computationally expensive and require fine-tuning for domain-specific tasks. Evaluating clustering quality remains difficult, as metrics like Purity and ARI may not fully capture semantic accuracy.

Recent research highlights that hybrid approaches combining SOMs with word embeddings outperform conventional methods, particularly in representing the semantic and syntactic structure of textual datasets. For example, models using Bag-of-Word Embeddings (BOWE) (Araújo et al., 2020) alongside SOMs have achieved better clustering quality by leveraging the strengths of both numerical computation and semantic alignment​​. Performance metrics such as Purity, ARI, and silhouette coefficients(Ravi & Kulkarni, 2023) have been widely adopted to evaluate clustering outcomes, and studies consistently indicate that embedding-based approaches achieve higher accuracy and computational efficiency​.

The integration of SOMs with modern embeddings, particularly BERT, offers a powerful solution for clustering unstructured English textual data. By addressing the limitations of traditional clustering techniques and leveraging semantic richness and topological preservation, this study aims to establish a scalable and robust clustering framework for English datasets. Insights from existing research provide a strong foundation for experimentation, contributing to advancements in the field of text clustering.

## **1.2 Related works on Improving the Text Quality using SOMs and Word embeddings.**

In (Mehta et al., 2021), The publication introduces the WEClustering algorithm, which leverages word embeddings and clustering techniques to analyze large-scale textual datasets, aiming to improve the organization and interpretation of unstructured text data. The main contribution of this approach is its ability to combine semantic understanding through embeddings like BERT with clustering methods, enabling more context-aware text grouping and uncovering latent structures in data. One advantage is its precision in handling polysemous words and capturing semantic relationships due to the use of advanced embeddings like BERT. However, a disadvantage is the computational complexity associated with processing large embeddings and datasets, which may require significant computational resources. The research applies word embedding techniques such as BERT and TF-IDF for vectorization and employs clustering methods like K-means to group the data. Verification methods include experiments that evaluate clustering quality using metrics like silhouette coefficient, purity, and adjusted Rand index across diverse datasets.

In (Ravi & Kulkarni, 2023), The publication introduces a comparative study of various text embedding techniques, including TF-IDF, Word2Vec, GloVe, Doc2Vec, and BERT, for clustering Twitter data, with the goal of improving clustering accuracy and efficiency in unstructured, context-sensitive social media datasets. The main contribution of this research is its empirical evaluation of these techniques, showcasing BERT as the superior model for generating meaningful representations and enhancing clustering outcomes. One advantage is BERT's ability to capture both statistical and contextual relationships within tweets, which significantly improves clustering quality. However, a disadvantage is its computational intensity, requiring considerable resources for embedding generation and model execution. The research applies text embedding techniques to vectorize tweets and clusters the resultant vectors using the K-means algorithm. Verification methods include internal metrics like silhouette scores and Davies-Bouldin index, and external metrics such as accuracy and F1-score, all of which validate the clustering performance of each embedding method. Experiments were conducted on tweets collected from Indian news channels on topics like Afghanistan's regime change and the Omicron COVID-19 variant.

In (Tache et al., 2021), The publication introduces the use of Self-Organizing Maps (SOMs) as a clustering technique for word embeddings, demonstrating superior performance compared to K-means on tasks such as sentiment classification and topic categorization. The main contribution of this approach is its ability to generate clusters that closely follow Zipf’s law, resulting in improved semantic representation and enhanced accuracy. The research applies SOMs to cluster word embeddings produced by models like Word2Vec and BERT, creating bag-of-word embeddings (BOWE) for classification tasks. Verification methods include experiments on the LaRoSeDa and MOROCO datasets, with SOMs achieving higher accuracy than K-means, including a 90.9% test accuracy for sentiment classification on LaRoSeDa. The tools used include Python, Scikit-learn for SVM classification, and SOM libraries for clustering. The datasets consist of 15,000 Romanian product reviews (LaRoSeDa) and news articles from the MOROCO corpus, validating the generalization of SOMs across multiple datasets. This paper contributes to my Master Thesis by demonstrating how SOMs can outperform traditional clustering algorithms, offering a powerful framework for improving clustering quality and advancing text analysis through density-preserving clustering methods.

In (Stefanovič & Kurasova, 2022), The publication introduces a method for handling multi-label text data classification by combining Self-Organizing Maps (SOM) and Latent Semantic Analysis (LSA), aimed at improving the accuracy of multi-label text classification by enhancing the semantic alignment between labels and data points. The main contribution of this approach is its iterative adjustment of label relationships based on the semantic structure of the data, facilitating better classification outcomes. One advantage is its ability to capture contextual relationships in high-dimensional textual datasets through SOM, while a disadvantage could be its reliance on computational resources and preprocessing, which may affect scalability. The research applies multi-label classification techniques, LSA for dimensionality reduction, cosine similarity metrics, and SOM to group semantically related data points and adjust labels iteratively. Verification methods include experiments that evaluate performance on real-world datasets, where the proposed approach is compared to traditional methods like k-Nearest Neighbors and decision trees using metrics such as Precision, Recall, and F1-score. Several experiments assess the effectiveness of the methodology in assigning accurate labels to multi-label text datasets.

In (Motegi & Seki, 2023), introduces the Shrinking Maximum Likelihood Self-Organizing Map (SMLSOM) algorithm for clustering and dimensionality reduction of high-dimensional data, aimed at enhancing efficiency by dynamically adapting the network size based on data distribution. The main contribution of this approach is its ability to adjust the structure of the SOM during training, reducing redundancy and focusing computational resources on meaningful areas of the data. One advantage is the reduction in memory usage and computational overhead by shrinking unused nodes, while a disadvantage is its sensitivity to initialization parameters, which may affect robustness across diverse datasets. The research applies probabilistic modeling, likelihood estimation, and dynamic SOM topology adjustments to process and analyze high-dimensional input data. Verification methods include experiments on synthetic and real-world datasets, where SMLSOM is compared to traditional SOM and k-means clustering based on performance metrics like accuracy, cluster compactness, and computation time.

In (Wang et al., 2023), The publication introduces a deep clustering framework that combines distributed text representations with clustering algorithms, focusing on improving the representation and clustering performance for textual datasets. The main contribution of this approach is its use of contrastive learning with pre-trained language models (PLMs) like BERT, enabling the generation of meaningful, low-dimensional embeddings that enhance clustering accuracy. One advantage is its effectiveness in capturing both semantic and syntactic nuances of text, while a disadvantage is the computational intensity of training PLMs and clustering large datasets. The research applies deep learning techniques, such as contrastive learning and clustering loss functions, combined with algorithms like Self-Organizing Maps (SOM) and k-means for organizing the text clusters. Verification methods include experiments on real-world datasets (AgNews and StackOverflow) using metrics like Normalized Mutual Information (NMI) and Accuracy (ACC).

In (Rejeb et al., 2022), The paper "Self-Organizing Maps for Exploration of Partially Observed Data and Imputation of Missing Values" introduces the missSOM algorithm, which simultaneously handles clustering and missing data imputation. This approach improves data visualization and clustering accuracy by integrating imputation into the map learning process. Its main contribution is generating topological maps that are robust to missing data, offering better representation and actionable insights. An advantage is its computational efficiency, especially with the accelerated version, while a drawback is the reliance on initial imputation quality. Verification includes experiments on real and synthetic datasets, measuring quantization error, topographic error, and imputation accuracy. Comparisons with state-of-the-art methods, like KNN and missForest, highlight missSOM's superior performance in clustering and data imputation.

In (Drakopoulos et al., 2022), publication introduces an advanced approach using Self-Organizing Maps (SOMs) to cluster and analyze user behavior for cultural content delivery systems. This method incorporates a tensor-based distance metric that includes behavioral attributes, enabling more effective segmentation of users and improving the relevance of delivered content. The main contribution of this approach is its ability to enhance clustering quality through the integration of multidimensional data, which allows for tailored cultural recommendations and insights into user preferences. One advantage of this method is its flexibility in handling complex datasets with behavioral, social, and topical attributes, providing a deeper understanding of user behavior. A potential disadvantage is its computational complexity, as tensor-based methods require significant processing power and high-quality, feature-rich data for optimal results

In (Yoshioka & Dozono, 2022), This research introduces the Spherical Tree-Structured Self-Organizing Map (S-TS-SOM), a novel framework for hierarchical clustering and data visualization. By extending traditional SOMs to a spherical topology and tree-structured architecture, the method provides enhanced scalability and improved representation of complex datasets. The key contributions include efficient multi-level clustering and the ability to handle high-dimensional data while preserving its topology. A significant advantage is its scalability for large datasets, while a potential drawback is the computational complexity of managing the hierarchical structure.

In (Delgado et al., 2021), the research leverages Self-Organizing Maps (SOM) to cluster and analyze student behavior in online learning environments, uncovering patterns that can enhance engagement and academic performance. The approach integrates clustering and visualization to provide actionable insights into student interactions with Learning Management Systems (LMS). Key contributions include the ability to process high-dimensional educational data, offering interpretable clusters that correlate with student activity and outcomes. A notable advantage is its effectiveness in identifying at-risk students for targeted interventions, while a limitation is its dependency on well-structured, pre-processed datasets. The research validates its approach through experiments on large-scale student activity data, analyzing patterns across timelines and engagement levels. This study enriches my Master Thesis by demonstrating the practical application of SOM for behavior analysis, contributing to personalized learning strategies and extending the potential of machine learning in education analytics

The research papers collectively explore advanced clustering techniques, particularly leveraging word embeddings, self-organizing maps (SOMs), and deep learning for improving text analysis. (Mehta et al., 2021) introduce WEClustering, combining BERT embeddings with clustering for efficient text organization, though it faces computational challenges. (Ravi & Kulkarni, 2023) compare multiple text embedding techniques, concluding that BERT improves clustering quality but is resource intensive. (Tache et al., 2021) propose using SOMs instead of K-means for clustering word embeddings, enhancing semantic representation while requiring careful parameter tuning. (Stefanovič & Kurasova, 2022) integrate SOMs with Latent Semantic Analysis (LSA) for multi-label classification, improving label alignment but demanding high preprocessing effort. (Motegi & Seki, 2023) develop SMLSOM, a dynamic SOM model that reduces computational overhead but is sensitive to initialization. (Wei et al., 2023) present a deep clustering framework using contrastive learning and pre-trained language models, improving representation quality but requiring extensive computation. (Rejeb et al., 2022) introduce missSOM for handling missing data while clustering, optimizing imputation but relying on initial data quality. (Drakopoulos et al., 2022) apply SOMs to user behavior clustering for cultural content recommendations, offering deeper insights but requiring complex feature-rich datasets. (Yoshioka & Dozono, 2022) extend SOMs to a hierarchical, spherical structure for scalable clustering, though its complexity increases with data size. (Delgado et al., 2021) apply SOMs in education analytics, effectively identifying learning patterns but needing structured input data.

**Table 1.2** provides a structured data extraction template for analyzing research studies related to textual data clusterization based on Self-Organizing Maps (SOM) and word embeddings. The table is designed to systematically categorize key aspects of each study, beginning with the reference column, which lists the citation of the analyzed research paper. It then outlines the main research question or problem, followed by the methodology or approach employed, particularly focusing on SOM and word embeddings for clustering textual data. The table further details the field of study or application domain, specifying whether the research pertains to text mining, natural language processing, sentiment analysis, or other relevant areas. It also includes information on datasets used, such as benchmark corpora, social media text, or domain-specific documents, along with the attributes considered for clustering, including document vectors, word embeddings, and textual features. Additionally, the table captures evaluation metrics such as silhouette scores, clustering accuracy, and external validation measures to assess the effectiveness of each approach. Comparisons with other clustering techniques, including traditional k-means, hierarchical clustering, and deep learning-based models, are documented to highlight the advantages and limitations of SOM-based methods. Finally, the results section summarizes key findings, demonstrating the impact of integrating SOM and word embeddings in improving textual data clustering, enhancing interpretability, and overcoming limitations of traditional vector-space models. This structured format enables a comprehensive review of existing studies, providing a clear understanding of the advancements and challenges in the field of SOM-based textual data clustering.

**Table 1: Data Extraction Template**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Reference (in APA style)** | **Main research question / problem** | **Used approach** | **Field Studied / Application domain** | **Dataset used** | **Attributes used for prediction** | **Evaluation of the approach** | **Comparison with other works** | **Result** |
| (Mehta et al., 2021) | How to perform clustering on large text datasets? | WEClustering based on word embeddings derived from a recent deep learning model named BERT | Text Mining, web search results clustering, browsing, social news clustering | Articles-253, Scorpus, 20NG, Classic4, Scorpus-long, Classic4-long, 20NG-long | Each document consists of title, Abstract and reference, and labels for categories of documents, Vectorized documents | Silhouette coefficient – the performance of techniques,  Purity–to measure clustering results when labels are available | BERT, K-means algorithm, and its minibatch version for handling large datasets and agglomerative clustering algorithm. | Document clustering is an important task in field of text mining. Existing clustering techniques have some limitations when applied to textual datasets based on TF-IDF based term-document matrix. Results shows that WEClustering outperforms all the compared techniques. |
| (Ravi & Kulkarni, 2023) | How the different word embedding techniques performs on twitter data? | BERT with K-means clustering | Text embedding, Twitter analysis | Taliban Tweets, Omicron Tweets | Document vectors created through different embedding techniques. | External clustering metrics: Accuracy, F1-Score and Adjusted Rand Score and internal clustering metric: Silhouette Score and Davies Bouldin Index for the evaluation of document embedding techniques and subsequent KMeans clustering | Comparison of BERT with Word2Vec, GloVe, Doc2Vec | BERT models on Tweets of news channels achieved 98% accuracy and gives the best performance of clustering result when the document vecotrs are clustered using K-means algorithm. |
| (Tache et al., 2021) | Sentiment analysis of Romanian language for the development of NLP. | Low-level features (character n-grams),  High level features (bag-of-word-embeddings generate by clustering word embeddings with SOMs | NLP | LaRoSeDa, which is composed of 15,000 positive and negative reviews collected from the one of the largest Romanian e-commerce platforms. MOROCO, a data set with Moldavian and Romanian news articles. | Word embedding vectors | 10-fold cross-validation procedure and using the train-test split. | HISK,  BOWE-BERT with k-means,  BOWE-BERT with SOMs,  HISK + BOWE-BERT with k-means,  HISK+BOWE-BERT with SOMs (BEST results) | SOMs, a clustering approach that preserves the density of words in the embedding space, resulting in a more effective bag-of-word-embeddings representation. Our top accuracy rates on LaRoSeDa are 89.54% for the cross-validation procedure and 90.90% on the test set. |
| (Stefanovič & Kurasova, 2022) | How to make semi-automated revisions of class assignments to improve the quality of Data using SOMs and LSA? | SOMs and LSA | Clustering | The newly collected data from four leading financial news websites in Lithuania have been experimentally analyzed. | Text data with assigned classes. Each text data cannot have more than two classes. | Correct Assignment Ratio = (Accept + Possible)/(Accept + Decline + Possible). | NONE | The experimental investigation has proved that the proposed approach can be used for multi-label text class adjustment and verification. The dimensionality reduction analysis using LSA has shown that the highest number of new assignments is made when the dimensionality is reduced to D = 40. |
| (S. Jamil et al., 2024) | The study addresses the challenge of clustering multi-view datasets, which often have high dimensionality, noise, and divergence across views. | Utilizes Self-Organizing Maps (SOMs) to map high-dimensional data to low -dimensional spaces. Apply dimensionality reduction and fusion techniques to merge complementary information across views. | The study is situated in the domain of unsupervised machine learning and applies specifically to multi-view clustering across various datasets, including. Image processing:Text clustering, Multimedia content analysis | Synthetic Datasets: Two-Moon and Two-Ring datasets. Real-World Datasets: YaleB,WebKB,Mfeat,Caltech-20,Caltech-7. | Input features: Derived from different views of the datasets.Cluster representations: Created using SOMs with adaptive weighting and relevance determination. | Accuracy (ACC) Normalized Mutual Information (NMI) | The paper benchmarks MSOMPA\_MV against state-of-the-art multi-view clustering methods. | Average accuracy: 61.8% to 93.6% across datasets. Demonstrated effectiveness in fusing multi-view data and handling high dimensionality. Consistently removed noisy data and maintained consensus information. |
| (Li et al., 2024) | The research aims to improve mode exploration capability by introducing SOMGAN, which integrates topological constraints and multiple discriminators. | Combines Self-Organizing Map (SOM) clustering with multiple discriminators to capture diverse data modes,  Introduces topological constraints over multiple discriminators to ensure diverse sample generation and robust mode coverage. | Computer vision | MNIST dataset  CIFAR-10  STL-10  CelebA  ImageNet | Input data modes,  Discriminator specialiazation,  Topological constraints,  Vectors | Visual comparison of mode coverage on synthetic datasets, Diversity assessment of generated samples across different categories, Use of Fréchet Inception Distance (FID) to measure the quality of generated images. | Standard GAN,UnrolledGAN, D2GAN, MGAN, DropoutGAN,VEEGAN,PacGAN. | SOMGAN is compatible with various GAN architectures and losses while maintaining computational efficiency. It outperforms standard models in both diversity and quality metrics, achieving near-uniform category coverage on challenging datasets like ImageNet. |
| (Palomino Mariño & Tenorio de Carvalho, 2024) | The paper addresses the challenge of effectively handling multi-view dissimilarity data in unsupervised learning, proposing new SOM (Self-Organizing Map) algorithms. | Multi-Medoids SOM (MBSOM-MMdd): Utilizes weighted medoids as cluster representatives and adapts relevance weights of dissimilarity matrices globally or locally. | Unsupervised clustering tasks.  Data visualization and mapping for multi-view datasets in diverse domains, including biomedical and image-based applications. | The algorithms were evaluated on 14 datasets, including:  Standard datasets: Iris, Wine, Mfeat, and Phoneme. | Dissimilarity matrices from multiple data views. Adaptive relevance weights are assigned either globally (for all clusters) or locally (for each cluster). Cluster representatives in the form of weighted medoids or linear combinations. | Internal Indices: Topographic Error (TE) for SOM quality. Silhouette Coefficient (SIL) for clustering quality.  External Indices:  F-measure and Normalized Mutual Information (NMI) for assessing clustering against known labels. | Single-view: SBSOM-CMdd, SBSOM-MMdd, and SRBSOM.  Multi-view: MBSOM-CMdd (global and local variants). | MBSOM-MMdd-L achieved the best Silhouette scores, while MRBSOM-G excelled in F-measure and NMI metrics. |
| (Cong et al., 2023) | How can the conceptual design of Smart Product-Service Systems (Smart PSS) be improved by incorporating user-generated emotions and feedback from online reviews, particularly in addressing gaps in current methodologies that prioritize functional requirements over emotional aspects? | Data Collection: Identifying reference products and collecting user-generated data from online reviews.  Self-Organizing Map (SOM): Used for clustering and analyzing review data. | Smart Product-Service Systems (Smart PSS), specifically focusing on enhancing user experience by addressing emotional and functional needs. | User reviews of similar traditional products (electric bicycles in the case study).  4080 online comments after cleaning and preprocessing. | User emotions (e.g., comfort, compactness).  Design elements (e.g., materials, riding mode, product interfaces). | The approach evaluates initial and improved conceptual design solutions through a user satisfaction scoring system. Weights for evaluation criteria are calculated using AHP. The final solution is assessed using Likert scale scores from target users. | Traditional conceptual design methods rely on functional requirements and designer intuition, often overlooking emotional aspects. This work bridges the gap by systematically incorporating user-generated emotions into the conceptual design process using SOM and interactive tools. | The proposed method demonstrated improved user satisfaction in a case study involving a smart electric bicycle service system (SEBSS). The dissatisfaction rate decreased, and average satisfaction scores increased across evaluated interaction interfaces after applying the new method. |
| (Motegi & Seki, 2023) | The paper addresses the limitations of conventional Self-Organizing Maps (SOMs) by proposing a novel variation, the Shrinking Maximum Likelihood Self-Organizing Map (SMLSOM). | The authors develop the SMLSOM algorithm, which uses a shrinking mechanism and maximum likelihood estimation. | The research is primarily focused on machine learning and neural network models, with applications in unsupervised learning, clustering, and data visualization. | The paper does not explicitly mention a specific dataset in the accessible content. It discusses synthetic and real-world datasets. | The approach focuses on general-purpose attributes relevant to clustering and data visualization tasks. | The evaluation is performed by comparing the performance of SMLSOM with traditional SOMs in terms of:Map topology preservationReconstruction error. | The paper compares SMLSOM with traditional SOM algorithms and highlights its advantages, including better adaptability, elimination of boundary effects. | The results show that SMLSOM outperforms traditional SOMs in representing data with complex distributions and is particularly effective for clustering tasks. |
| (Kaur et al., 2024) | The paper focuses on improving the retrieval and annotation of in vivo medical images using a hybrid self-organizing map (HSOM). The main goal is to facilitate efficient analysis and interpretation of medical images. | The hybrid self-organizing map (HSOM) approach combines clustering with neural network-based image annotation and retrieval. | The application domain is medical imaging, particularly in vivo imaging. | The dataset consists of vivo images; however, specific details about the dataset (e.g., size or source) were not included in the visible snippets. If needed, I can analyze the dataset details further from the document. | Image features derived from spatial distributions. Intensity-based patterns and other clustering-relevant parameters. | The performance of the HSOM is evaluated by: Its accuracy in retrieving relevant medical images. Computational efficiency in clustering and annotation. Comparison with traditional retrieval systems. | Conventional SOM methods. Standard image retrieval and clustering tools, highlighting the limitations of those methods in handling complex, in vivo image datasets. | The HSOM showed superior performance in: Effectively clustering and organizing medical images. Achieving higher accuracy and efficiency compared to conventional approaches. |
| (Javed et al., 2024) | The study aims to explore the use of Self-Organizing Maps (SOMs) for time series clustering, specifically in the context of understanding patterns in serious illness conversations. | The paper employs Self-Organizing Maps (SOMs), an unsupervised neural network-based clustering method. | Field: Machine Learning and Data Analysis. Application Domain: Healthcare communication, particularly focused on serious illness conversations between patients and healthcare providers. | The study uses conversational datasets containing transcripts of serious illness discussions. These datasets include time-stamped sequences reflecting communication dynamics. | Attributes include temporal sequences and contextual features of the conversations, such as emotional tone, sentiment, and other linguistic markers. | The effectiveness of SOM clustering is evaluated by assessing the quality and interpretability of the resultant clusters. Metrics include within-cluster similarity and domain expert validation. | The paper compares SOM-based clustering with traditional clustering techniques (like k-means) and highlights the advantages of SOMs. | The SOM-based clustering approach successfully identifies meaningful conversational patterns, which are validated by domain experts. The findings demonstrate the potential of SOMs to aid in understanding and improving serious illness communication. |
| (Wang et al., 2023) | How to learn distributed representations and perform deep embedded clustering of texts effectively using neural network-based methods. | The study introduces a model that combines distributed representations learning with deep embedded clustering, leveraging the latent representations of text data for clustering tasks. | Text analysis and natural language processing. | The specific datasets are not detailed in the snippet but would typically include benchmark text corpora for clustering tasks. | Latent features of the text derived from the neural network-based distributed representations. | Evaluation metrics are not explicitly mentioned in the provided snippet, but standard clustering metrics like adjusted Rand index (ARI) or normalized mutual information (NMI) are likely used. | The study compares its approach against other state-of-the-art text clustering methods, focusing on the benefits of integrating distributed representations with deep clustering. | The proposed method demonstrates superior performance over traditional clustering methods in terms of effectiveness and clustering accuracy. |
| (Rejeb et al., 2022) | How can self-organizing maps (SOMs) be adapted to handle partially observed data and simultaneously impute missing values effectively? | The authors propose an extension of the classical Kohonen algorithm, named miss SOM, to compute self-organizing maps for incomplete data. | Primary Domain: Chemometrics. Other Fields of Application Mentioned: Data visualization and clustering in high-dimensional datasets. Fraud detection, Geological data classification, financial data analysis, Health monitoring of aircraft engines. | Wines Dataset, Simulated Multivariate Gaussian Mixture Data | Wines Dataset: Attributes pertain to the chemical composition of wines. Gaussian Mixture Dataset: Five dimensions with simulated correlation and Gaussian mean differences. | Criteria:Quantization Error: Measures the squared distances between data points and their closest prototype vectors, assessing representation quality. | Deletion Method: Removes incomplete data before SOM computation. Cottrell’s Approach: Adapts the SOM algorithm to incomplete observations by restricting calculations to observed data entries. | missSOM consistently provides high-quality maps, particularly excelling in topographic error preservation. For imputation accuracy, missSOM outperformed most methods, especially when the missingness rate is high, though MissForest showed superior accuracy for pure imputation tasks. |
| (A. Jamil et al., 2023) | The study addresses two main challenges in conventional Self-Organizing Maps (SOM): Achieving high accuracy with fast convergence. Reducing topological error (TE) while preserving topology throughout training iterations. | Introduced a Variable Learning Rate SOM (VLRSOM), which adapts the learning rate dynamically based on the error response during training. | The research focuses on unsupervised learning algorithms, particularly in clustering and dimensionality reduction. | Synthetic Data: Generated in a 2D feature space for initial testing. MNIST Dataset: 70,000 samples of handwritten characters divided into training (60,000) and testing (10,000) subsets. | Features vary by dataset:  Synthetic data: Coordinates in a 2D space.  MNIST: Pixel intensities of digit images. | Accuracy: Classification success rate.  Quantization Error (QE): Measures how well input data maps to neurons in the SOM.  Topology Error (TE): Assesses topology preservation by checking if neighboring neurons represent similar inputs. | Conventional SOM  PLSOM2 (Parameter-Less SOM) | VLRSOM showed faster convergence and lower QE and TE compared to other models.  Example (200 iterations): VLRSOM (QE: 4.6×10⁻⁴, TE: 1.0×10⁻⁴) outperformed PLSOM2 (QE: 5.6×10⁻⁴, TE: 3.57×10⁻²). |
| (Drakopoulos et al., 2022) | How can a tensor distance metric, coupled with self-organizing maps (SOMs), be used to cluster user bases for a cultural content delivery system effectively? | Developed a tensor distance metric that incorporates behavioral attributes along with typical attributes in recommender systems. | Cultural Content Delivery Systems: Systems designed to recommend and deliver cultural items (e.g., art, music, history). | A benchmark dataset containing:  Behavioral Attributes:  Number of sessions per month.  Average session duration.  Average keywords per query.  Average queries per session. | Behavioral Attributes:  Number of sessions per month (min: 1, max: 52).  Average session duration (min: 0.83 hours, max: 5.1 hours).  Average keywords per query (min: 2, max: 7).  Average queries per session (min: 4, max: 21). | Evaluated using metrics such as:  Topological Error (TE): Measures the percentage of misplaced data points in clusters.  Quantization Error (QE): Assesses the average distance between data points and their cluster centroids.  Correction Decay (CD): Tracks the stabilization of cluster centroids over epochs. | Compared the proposed tensor distance metric with:  Traditional SOMs without behavioral attributes.  SOMs using standard distance metrics like cosine similarity and infinity norm. |  The inclusion of behavioral attributes:  Improved topological error and quantization error compared to baseline methods.  Achieved better cluster separation (increased inter-cluster distance).  Reduced the number of training epochs, indicating faster convergence. |
| (Yoshioka & Dozono, 2022) | How to improve the self-organizing map (SOM) for high-dimensional data visualization and clustering by reducing learning time and eliminating learning unevenness caused by map edges. | Proposal of the Spherical Tree-Structured SOM (S-TS-SOM):  Nodes are arranged on a spherical surface. | Machine learning, data visualization, and clustering in high-dimensional data analysis. | MNIST dataset: 60,000 training images and 10,000 test images of handwritten digits, converted to 784-element vectors. Zoo dataset (UCI Machine Learning Repository): 101 animal species classified into 7 types based on 16 Boolean attributes. | MNIST dataset: Pixel intensity values of 28x28 images.  Zoo dataset:  16 attributes such as "hair", "feathers", "eggs", normalized "legs", etc. | Visualization Experiment:  Visualized MNIST data using the S-TS-SOM, clustering similar images together on a spherical map.  Quantitative Clustering Performance:  Evaluated using Purity and Normalized Mutual Information (NMI) metrics.  S-TS-SOM achieved comparable results to Spherical SOM (S-SOM) but required significantly less computation time. | Spherical SOM (S-SOM):  Similar clustering performance in terms of Purity and NMI.  S-TS-SOM reduced learning time, especially with increased node counts.  Improved clustering hierarchy due to tree structure. | Visualization:  S-TS-SOM effectively grouped similar items and displayed clustering on a spherical surface.  Performance:  Comparable clustering metrics (Purity and NMI) to S-SOM.  Learning time significantly reduced as the number of nodes increased (e.g., 12% of S-SOM time for 5120 nodes).  Hierarchical Clustering:  Successfully demonstrated multi-layer clustering granularity using the Zoo dataset. |
| (Ferles et al., 2021) | The study investigates the design and evaluation of a Deep Self-Organizing Convolutional Map (SOCOM) for clustering and visualizing image data. | The authors hybridized Convolutional Neural Networks (CNNs) and Self-Organizing Maps (SOMs) with gradient backpropagation optimization to develop the SOCOM framework. SOCOM leverages CNNs for feature extraction and SOMs for clustering and visualization in an end-to-end unsupervised architecture. | The study focuses on the field of unsupervised deep learning with applications in clustering and visualization of image datasets. | The experiments primarily utilized the STL-10 dataset. | SOCOM operates on the learned high-dimensional feature representations generated by the CNN backbone. The features are clustered into SOM neurons based on topological relationships and activation responses. | Visualization Experiments: Neural Map Visualization (NMV) provided insights into the clustering results and neuron activations. | The paper benchmarks SOCOM against various state-of-the-art models:  It achieved 78.7% accuracy on the STL-10 dataset without requiring labeled data during the clustering process. | SOCOM successfully produced accurate and interpretable clustering outputs on the STL-10 dataset without supervised labels.  The Neural Map Visualization (NMV) demonstrated that neurons encode meaningful patterns related to the input data. |
| (Forest et al., 2021) | The study explores how to integrate representation learning and clustering into a single framework using DESOM | The authors propose DESOM, a model that combines:  An autoencoder for representation learning. | This work is in the field of unsupervised learning. | The model was evaluated on:  MNIST (grayscale handwritten digits, 28×28 pixels). | The prediction utilizes latent space representations extracted by the encoder. | Evaluation criteria included:  Clustering Performance: Metrics like purity and normalized mutual information (NMI). |  DESOM outperformed traditional SOM, AE+SOM, and other SOM variants | DESOM achieved higher purity and NMI across datasets compared to baseline SOM methods. |
| (Delgado et al., 2021) | The research focuses on analyzing student behavior in online learning environments to identify clusters of user behaviors. | The study employs Self-Organizing Maps (SOMs), an unsupervised neural network, to cluster and analyze over 1.7 million records of student activity from an online university | Educational Data Mining and Learning Analytics, specifically applied to online learning environments. | Source: Data from the Universidad Internacional de La Rioja (UNIR), covering 2015-2019. | The dataset was preprocessed to generate activity vectors consisting of:100 features: Counts of 10 event types (e.g., forum posts, resource creation) across 10 course periods. | SOM clustering was assessed based on:  Topographic function for topology preservation.  CONNIndex and Davies-Bouldin Index (DBI) for cluster quality.  3D Grade Maps were used to visually correlate activity patterns with student performance. | SOM-based clustering offered advantages over traditional clustering techniques like k-means, including better visualization and handling of complex data shapes. | Clusters Identified: 13 distinct activity patterns, including:  Clusters linked to high interaction (e.g., frequent forum participation) showed better performance. |
| (Kotyrba et al., 2021) | The study aims to investigate the use of hybrid clustering methods that combine Self-Organizing Maps (SOM) with conventional clustering algorithms (CLARA, CURE, K-means) to improve clustering efficiency, especially when initial data knowledge is vague. | The researchers developed a hybrid clustering methodology. Initially, SOM was applied for preprocessing and topological structuring of data. Following this, selected clustering algorithms (CLARA, CURE, or K-means) were employed for further clustering based on the preprocessed outputs. | The research falls within artificial intelligence and clustering analysis. It addresses clustering techniques applicable to diverse datasets, especially in complex scenarios where conventional clustering struggles. | The Fundamental Clustering Problems Suite (FCPS) dataset was used, which consists of multiple benchmark datasets with varying complexities. | Attributes of the FCPS datasets included dimensions (2D/3D), the number of classes/clusters, and object count. Clustering outputs were evaluated using topological structuring | The Rand index was the primary evaluation metric, quantifying clustering performance by comparing the similarity between predicted clusters and actual patterns. | The proposed hybrid methods (SOM + CLARA, SOM + CURE, SOM + K-means) were compared to standalone clustering methods and existing approaches in the literature. | The hybrid approaches improved clustering accuracy, especially for complex datasets like Atom, which was fully resolved using SOM + CLARA and SOM + K-means. |
| (Balaji et al., 2020) | The study investigates the development of a clustering algorithm for mixed datasets (numerical and categorical attributes). The primary focus is to enhance clustering performance and address challenges such as data diversity and scalability. | A novel Density-Peaks and Self-Organizing Generative Adversarial Networks (DP-SO-GAN) model was proposed: Preprocessing: Categorical attributes were transformed using one-hot encoding, and numerical attributes normalized. | This research is in the field of machine learning and clustering with applications in:  Medicine (e.g., clustering cardiovascular disease data)  Image processing  Pattern recognition | The study employed five datasets, including:  Cardiovascular Disease Dataset: A mixed dataset with numerical and categorical attributes from the UCI repository. | For the Cardiovascular Disease Dataset, the attributes included:  Numerical: Age, resting blood pressure, serum cholesterol, maximum heart rate, etc. | The clustering performance was evaluated using accuracy (ACC) as the metric, comparing predicted cluster labels with true labels.  Structural Similarity (SS) and Feature Similarity (RFS) metrics were used for image-related datasets. | The proposed DP-SO-GAN model was compared with several existing clustering algorithms (e.g., K-Prototypes, KL-FCM-GM, DPC-MD). It consistently outperformed these models in clustering accuracy. | The DP-SO-GAN demonstrated higher clustering accuracy and efficiency across all datasets.  It reduced computational complexity by 18% compared to classical GANs. |
| (Gorzałczany & Rudziński, 2020) | The paper investigates the evolution of Self-Organizing Maps (SOMs) to address challenges in high-dimensional data visualization and clustering of complex data. | The study presents generalized SOMs with advanced structural and learning algorithms:  Dynamic SOMs (DSOMs): Feature splitting-merging neuron chains. | The research lies in the domains of artificial neural networks, machine learning, and clustering. Applications include gene expression analysis, document clustering, and electricity consumption profiling. | Various benchmark datasets were employed for evaluation:  Gene expression datasets (e.g., leukemia, lymphoma, colon cancer). | Attributes were context-dependent, such as:  Numerical gene expression levels in biological datasets. | Performance was assessed using:  Clustering accuracy: Measured the percentage of correct cluster assignments. | The proposed methods (DSOMs and GeSOMs with T-LSs) were compared with traditional SOMs and other clustering techniques | DSOMs and GeSOMs outperformed alternative approaches in detecting clusters, even in highly complex datasets. Demonstrated flexibility in handling datasets with diverse shapes and densities. |
| (Araújo et al., 2020) | How to effectively perform clustering on high-dimensional and multi-view datasets where traditional distance metrics lose discriminant power? | Introduction of LARFDSSOM2 (Local Adaptive Receptive Field Dimension Selective Self-Organizing Map 2), a self-organizing map with a time-varying structure. | Data mining  Gene expression  Multi-view categorization  Computer vision  Text clustering | UCI Machine Learning Repository datasets  Gene expression datasets (e.g., Leukemia, Prostate Cancer, Brain\_Tumor2 | High-dimensional features, such as gene expression levels, pixel intensities, or extracted features for categorization in multi-view and computer vision tasks. | Metrics used:  Rand Index (RI)  Adjusted Rand Index (ARI)  Clustering Accuracy (ACC) | Compared with algorithms such as:  RANSAC, GPCA, SSC, ALC for motion segmentation.  EWKM, FSC, LAC, AWA, FWKM for single-view clustering. | LARFDSSOM2 achieved higher accuracy and lower error rates across a wide variety of datasets, often outperforming its predecessor (LARFDSSOM) and other state-of-the-art methods. |
| (Samaranayaka & Wimalaratne, 2020) | How can Self-Organizing Maps (SOMs) effectively cluster and visualize high-dimensional call data to identify and interpret sudden call hikes? | The study employs Self-Organizing Maps (SOMs) as the core clustering technique. SOMs, a type of neural network, are used to reduce data dimensionality and create visual representations of patterns and anomalies in call data. The system integrates SOM-based visualizations into a wallboard interface to enable real-time monitoring and interpretation of call behavior. | The application domain is contact center analytics, specifically for monitoring and managing sudden increases in call volumes. The approach is aimed at improving operational efficiency and resource allocation in contact centers. | Call timestamps  Call durations  Nature of inquiries  Call resolution statuses. | Temporal information: Time and date of the calls. Behavioral patterns: Call durations and inquiry types.Volume metrics: Frequency and intensity of call spikes. | Quantization Error: Measures how well the SOM represents the data.  Topological Error: Assesses the preservation of data relationships in the clusters. Visual Interpretability: Based on qualitative feedback from contact center managers, indicating how easily patterns and anomalies could be understood. | K-means clustering: Found less effective for high-dimensional and non-linear data.  Hierarchical clustering: Not as visually intuitive as SOM-based wallboards. | The SOM-based wallboard system effectively visualizes call volume anomalies.  Managers were able to detect sudden call hikes and interpret underlying patterns in real-time.  The SOM approach achieved lower quantization and topological errors compared to k-means and other methods, proving its robustness in clustering multi-dimensional call data. |

**Summary of Literature Review Based on** **Textual Data Clusterization Based on Self-organizing Map and Word Embeddings**

The primary focus of this research is on how self-organizing maps (SOM) combined with word embeddings can improve textual data clustering. Traditional clustering methods struggle with high-dimensional, unstructured text data, leading to inaccurate groupings. This study investigates the effectiveness of SOM in capturing semantic relationships and improving the organization of textual information.

The research employs a hybrid approach integrating self-organizing maps with word embeddings to enhance textual data clustering. Word embeddings transform text into numerical vectors, preserving semantic meaning, while SOM organizes the vectors into clusters based on similarity. This method ensures more interpretable and meaningful text groupings compared to traditional clustering techniques.

The research is conducted is the field of SOM-based text clustering in domains such as information retrieval, sentiment analysis, and document classification. It is particularly useful in fields like customer feedback analysis, news categorization, and social media monitoring, where large volumes of unstructured text need efficient organization.

The datasets used for this research includes Publicly available and domain-specific datasets to evaluate the proposed method. Common datasets include 20 Newsgroups for document classification, IMDb for sentiment analysis, and Wikipedia text corpora for semantic clustering. Custom datasets may also be developed for specific applications requiring specialized text processing.

The attributes used for prediction include word vector representations obtained from word embedding models like Word2Vec, GloVe, and FastText. Other attributes include sentence structure, term frequency-inverse document frequency (TF-IDF), and contextual relationships, which help improve clustering accuracy and semantic organization.

The approach is evaluated using clustering performance metrics such as silhouette score, Davies-Bouldin index, and purity. These metrics measure how well SOM-based clustering structures textual data compared to traditional clustering methods. Computational efficiency and scalability are also assessed to determine real-world applicability.

The research includes a comparative analysis with SOM-based clustering with traditional techniques like k-means, hierarchical clustering, and DBSCAN. Unlike conventional methods, SOM provides a more interpretable representation of text clusters, capturing nonlinear relationships and improving the contextual organization of textual data.

The results indicate that integrating self-organizing maps with word embeddings significantly improves text clustering performance. The proposed method enhances semantic grouping, reduces noise in classification, and provides better insights into textual data patterns, making it useful for various real-world applications.

# **Conclusion**

Based on the performed analysis of the integration of Self-Organizing Maps (SOM) and word embeddings, such as BERT and Word2Vec, offers a scalable and semantically rich clustering framework that addresses high-dimensional data challenges, enabling contextually accurate insights while aligning with natural language patterns like Zipf’s law.

# **References**

Araújo, A. F. R., Antonino, V. O., & Ponce-Guevara, K. L. (2020). Self-organizing subspace clustering for high-dimensional and multi-view data. *Neural Networks*, *130*, 253–268. https://doi.org/10.1016/j.neunet.2020.06.022

Balaji, K., Lavanya, K., & Mary, A. G. (2020). Clustering algorithm for mixed datasets using density peaks and Self-Organizing Generative Adversarial Networks. *Chemometrics and Intelligent Laboratory Systems*, *203*(June), 104070. https://doi.org/10.1016/j.chemolab.2020.104070

Cong, J., Chen, C. H., Meng, X., Xiang, Z., & Dong, L. (2023). Conceptual design of a user-centric smart product-service system using self-organizing map. *Advanced Engineering Informatics*, *55*(November 2022), 101857. https://doi.org/10.1016/j.aei.2022.101857

Delgado, S., Moran, F., Jose, J. C. S., & Burgos, D. (2021). Analysis of Students’ Behavior through User Clustering in Online Learning Settings, Based on Self Organizing Maps Neural Networks. *IEEE Access*, *9*, 132592–132608. https://doi.org/10.1109/ACCESS.2021.3115024

Drakopoulos, G., Giannoukou, I., Mylonas, P., & Sioutas, S. (2022). Self organizing maps for cultural content delivery. *Neural Computing and Applications*, *34*(22), 19547–19564. https://doi.org/10.1007/s00521-022-07376-1

Ferles, C., Papanikolaou, Y., Savaidis, S. P., & Mitilineos, S. A. (2021). Deep Self-Organizing Map of Convolutional Layers for Clustering and Visualizing Image Data. *Machine Learning and Knowledge Extraction*, *3*(4), 879–899. https://doi.org/10.3390/make3040044

Forest, F., Lebbah, M., Azzag, H., & Lacaille, J. (2021). Deep embedded self-organizing maps for joint representation learning and topology-preserving clustering. In *Neural Computing and Applications* (Vol. 33, Issue 24). https://doi.org/10.1007/s00521-021-06331-w

Gorzałczany, M. B., & Rudziński, F. (2020). Evolution of SOMs’ structure and learning algorithm: From visualization of high-dimensional data to clustering of complex data. *Algorithms*, *13*(5). https://doi.org/10.3390/A13050109

Jamil, A., Hameed, A. A., & Orman, Z. (2023). A faster dynamic convergency approach for self-organizing maps. *Complex and Intelligent Systems*, *9*(1), 677–696. https://doi.org/10.1007/s40747-022-00826-2

Jamil, S., Rehman, E. I. D., Shahzad, T., Ishtiaq, M., Mazhar, T., Ghadi, Y. Y., & Ahmed, A. (2024). *Multi Self-Organizing Map ( SOM ) Pipeline Architecture for Multi-View Clustering*. *12*(May).

Javed, A., Rizzo, D. M., Lee, B. S., & Gramling, R. (2024). Somtimes: self organizing maps for time series clustering and its application to serious illness conversations. *Data Mining and Knowledge Discovery*, *38*(3), 813–839. https://doi.org/10.1007/s10618-023-00979-9

Kaur, P., Malhi, A., & Pannu, H. (2024). Annotate and retrieve in vivo images using hybrid self-organizing map. *Visual Computer*, *40*(8), 5619–5638. https://doi.org/10.1007/s00371-023-03126-z

Kotyrba, M., Volna, E., Jarusek, R., & Smolka, P. (2021). The use of conventional clustering methods combined with SOM to increase the efficiency. *Neural Computing and Applications*, *33*(23), 16519–16531. https://doi.org/10.1007/s00521-021-06251-9

Li, W., He, Y., & Tang, Y. (2024). Improving mode exploring capability of generative adversarial nets by self-organizing map. *Neurocomputing*, *576*, 127244. https://doi.org/10.1016/J.NEUCOM.2024.127244

Mehta, V., Bawa, S., & Singh, J. (2021). WEClustering: word embeddings based text clustering technique for large datasets. *Complex and Intelligent Systems*, *7*(6), 3211–3224. https://doi.org/10.1007/s40747-021-00512-9

Motegi, R., & Seki, Y. (2023). SMLSOM: The shrinking maximum likelihood self-organizing map. *Computational Statistics and Data Analysis*, *182*, 107714. https://doi.org/10.1016/j.csda.2023.107714

Palomino Mariño, L. M., & Tenorio de Carvalho, F. de A. (2024). Self-organizing maps with adaptive distances for multiple dissimilarity matrices. *Machine Learning*, *113*(10), 7783–7806. https://doi.org/10.1007/S10994-024-06607-X/TABLES/16

Paper, I., Introduction, I., Models, N., & Maps, S. (1990). *The Self-organizing Map*. *78*(9), 1464–1480.

Ravi, J., & Kulkarni, S. (2023). Text embedding techniques for efficient clustering of twitter data. *Evolutionary Intelligence*, *16*(5), 1667–1677. https://doi.org/10.1007/s12065-023-00825-3

Rejeb, S., Duveau, C., & Rebafka, T. (2022). Self-organizing maps for exploration of partially observed data and imputation of missing values. *Chemometrics and Intelligent Laboratory Systems*, *231*(August), 104653. https://doi.org/10.1016/j.chemolab.2022.104653

Samaranayaka, J. R. A. C. P., & Wimalaratne, P. (2020). Self-Organizing Map based Wallboards to Interpret Sudden Call Hikes in Contact Centers. *International Journal of Advanced Computer Science and Applications*, *11*(11), 208–215. https://doi.org/10.14569/IJACSA.2020.0111127

*Self Organizing Map(SOM) with Practical Implementation | by Amir Ali | Aorb Tech | Medium*. (n.d.). Retrieved February 3, 2025, from https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117

Stefanovič, P., & Kurasova, O. (2022). Approach for Multi-Label Text Data Class Verification and Adjustment Based on Self-Organizing Map and Latent Semantic Analysis. *Informatica (Netherlands)*, *33*(1), 109–130. https://doi.org/10.15388/22-INFOR473

Tache, A. M., Gaman, M., & Ionescu, R. T. (2021). Clustering word embeddings with self-organizing maps. Application on LaRoSeDa - A large romanian sentiment data set. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 949–956. https://doi.org/10.18653/v1/2021.eacl-main.81

Wang, S., Beheshti, A., Wang, Y., Lu, J., Sheng, Q. Z., Elbourn, S., & Alinejad-Rokny, H. (2023). Learning Distributed Representations and Deep Embedded Clustering of Texts. *Algorithms*, *16*(3), 1–17. https://doi.org/10.3390/a16030158

Wei, X., Wang, Z., & Yang, S. (2023). An Automatic Generation and Verification Method of Software Requirements Specification. *Electronics (Switzerland)*, *12*(12). https://doi.org/10.3390/electronics12122734

Yoshioka, K., & Dozono, H. (2022). Spherical Tree-Structured SOM and Its Application to Hierarchical Clustering. *Applied System Innovation*, *5*(4), 1–12. https://doi.org/10.3390/asi5040076