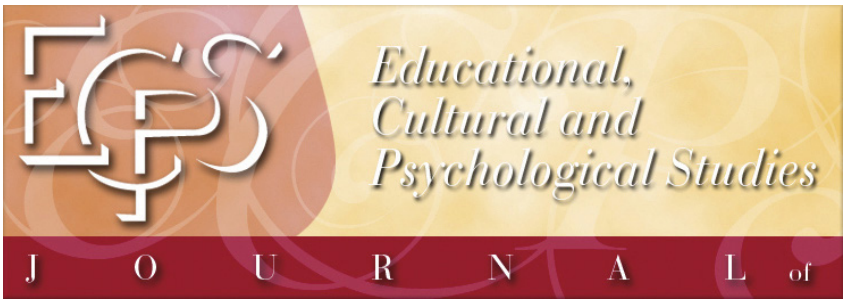

30
December 2024

Special Issue on

The Contribution of Artificial Intelligence
to the Qualification of Educational Processes
Il contributo dell'intelligenza artificiale
alla qualificazione dei processi di istruzione

Edited by
Gaetano Domenici

ECPS Journal – 30/2024 - https://www.ledonline.it/ECPS-Journal/
Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

ECPS Journal – 30/2024 - https://www.ledonline.it/ECPS-Journal/
Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

Commenti, Riflessioni, Presentazioni, Resoconti, Dibattiti, Interviste

Comments, Reflections, Presentations, Reports, Debates, Interviews

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

Artificial Intelligence to Enhance Qualitative Research: Methodological Reflections on a Pilot Study

Maria Luongo [1] - **Michela Ponticorvo** [1]
Maria Beatrice Ligorio [2] - **Pietro Crescenzo** [2]
Giuseppe Ritella [3]

[1] *Università di Napoli «Federico II» - Department of Humanities, Natural and Artificial Cognition Laboratory «Orazio Miglino» (Italy)*

[2] *Università di Bari «Aldo Moro» - Department of Educational Sciences (Italy)*

[3] *Università della Campania «Luigi Vanvitelli» - Department of Psychology (Caserta, Italy)*

DOI: https://doi.org/10.7358/ecps-2024-030-luon

maria.luongo@unina.it
michela.ponticorvo@unina.it
bealigorio@hotmail.com
pietro.crescenzo@uniba.it
giuseppe.ritella@unicampania.it

L'INTELLIGENZA ARTIFICIALE PER POTENZIARE LA RICERCA QUALITATIVA: RIFLESSIONI METODOLOGICHE SU UNO STUDIO PILOTA

Abstract

Qualitative analysis is essential in research across diverse fields, offering in-depth insights that often cannot be captured through quantitative methods. However, managing large volumes of qualitative data presents challenges, including its labour intensive nature and the potential for interpretive biases. In this study, we introduce and show a methodology step by step that integrates artificial intelligence (AI) in the analysis of qualitative data, with a focus on textual responses extracted from survey questions. Specifically, our

approach employs AI techniques, utilizing Word2Vec for word embedding extraction and K-Means clustering to streamline the analysis of qualitative textual data, while ultimately integrating the researcher's interpretation of the identified clusters to improve the relevance of the analysis. Moreover, the present article discusses the relevance and significance of this approach as well as its ethical and methodological challenges by means of an empirical illustration taken from a study on teachers' sensemaking regarding a range of different educational activities.

Keywords: Artificial intelligence; Clustering; Embeddings; Qualitative research; Textual data.

1. Introduction

The interplay between qualitative research and advanced computational techniques like machine learning and natural language processing (NLP) represents a transformative shift in how researchers can approach text analysis (Dey, 2003). Traditional qualitative methods, while rich and nuanced, come with inherent limitations due to their labor-intensive nature and potential for subjective bias. Researchers' backgrounds, experiences, and perspectives can inadvertently influence the themes and codes they derive from text, which could affect the consistency and validity of the findings (Manning, 2022). To address these issues, qualitative researchers have historically developed methodological solutions that significantly improve the validity of qualitative research. For example, in many qualitative investigations, multiple analysts code the data independently to identify consensus and discrepancies in their interpretations. Inter-rater reliability is often considered in this context to assess the quality of the coding system adopted and the validity of the findings (Cohen, 1960; Fleiss, 1971). Some qualitatively minded scholars, instead, adopt an iterative and dialogic revision of the coding until complete agreement on the interpretation of the data is reached, based on the assumption that the dialogue between researchers on the interpretation of the data can support the emergence of more nuanced interpretations, improving the relevance and significance of the conclusions (e.g. Ritella *et al.*, 2022). Such measures, while ensuring the accuracy and rigor of the research, add another layer of complexity and effort to the research process, often also leading to potential increase of the number of researchers involved in the research and/or to a temporal extension of the data analysis phase. The advent of machine learning, particularly in the realm of NLP, offers a compelling complement to these

traditional methods (Abram *et al.*, 2020; Clarke *et al.*, 2020; Natural Language Processing, 2020). Large language models (LLMs) like ChatGPT represent a significant leap forward (Valdenegro, 2023). These models, built on vast neural network architectures, transform textual data into numerical forms, enabling the machine to process language in a way that has a great potential for supporting qualitative research. These LLMs are not just static tools; they learn and evolve. Trained on extensive corpora of text and fine-tuned through human feedback, they can develop an ability to contribute to the analysis of nuances, contexts, and even the subtleties of human preferences (Wang *et al.*, 2023). This dynamic learning process is what makes LLMs particularly valuable for qualitative research. They can handle large volumes of text swiftly, identifying patterns at a scale and speed unattainable for human researchers.

Nevertheless, the role of the human researcher remains central. Qualitative researchers bring a depth of understanding, interpretive nuance, and contextual awareness that machines struggle to replicate. When researchers partner with LLMs, they can guide the analysis, ensuring that the themes and codes generated align with the research objectives and contextual realities of the data. In addition, as will be discussed at the end of this article, the patterns identified by means of AI tools can provide further inputs for the researchers' interpretation of the data, potentially leading to more nuanced findings. The integration of human insight with machine processing can enhance the robustness of qualitative analysis. Researchers can use LLMs to conduct initial coding or to cross-check and complement their manual coding efforts, thereby reducing the potential for subjective bias and increasing the reliability of the findings. In essence, the confluence of qualitative research and NLP represents a symbiotic relationship. Each domain contributes its strengths, with qualitative researchers providing interpretive depth and contextual sensitivity, and LLMs offering computational power and efficiency. This partnership not only has the potential to accelerate and facilitate the research process but also to enhance the capacity to conduct rigorous analysis with large textual datasets that are usually difficult to handle for single research groups, opening new horizons for understanding complex social phenomena through text. The objective of this research is to describe step by step a methodology aimed at categorizing and analyzing qualitative data. The article illustrates how the adoption of AI can support researchers to conduct qualitative analysis on large amounts of textual data.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

1.1. *AI integration in qualitative research stages*

As mentioned in the introduction of this article, artificial intelligence (AI), especially through machine learning and natural language processing, can play a multifaceted role in various stages of qualitative research, opening up new pathways for data collection, processing, analysis, and interpretation (Christou, 2023). This integration has the potential of both facilitating the implementation of labor intensive research processes and contributing to unveiling complex insights from large sets of qualitative data, potentially enhancing the overall quality and significance of qualitative studies. During data collection, AI can facilitate the organization and initial categorization of textual data, sifting through voluminous and diverse text sources to identify preliminary patterns or themes. This automation not only has the potential to save time but can also introduce a level of preliminary analysis that can support researchers in their subsequent research efforts (Creswell & Poth, 2017). The pre-processing stage can benefit from AI for tasks such as text cleaning, normalization, and automated feature extraction, ensuring that the data is primed for in-depth analysis. By timely handling such traditionally labor-intensive aspects of data preparation, AI allows researchers to focus on the more nuanced and meaningful aspects of their study. Nevertheless, it is in the data analysis phase that the contribution of AI in the research process is most promising, as it offers more and more sophisticated tools allowing to uncover underlying themes and relationships within the data. AI's role can extend to the validation of research findings, where it can serve as a tool for cross-verifying themes and patterns identified through traditional qualitative methods, enhancing the credibility and robustness of the research.

2. Methodology

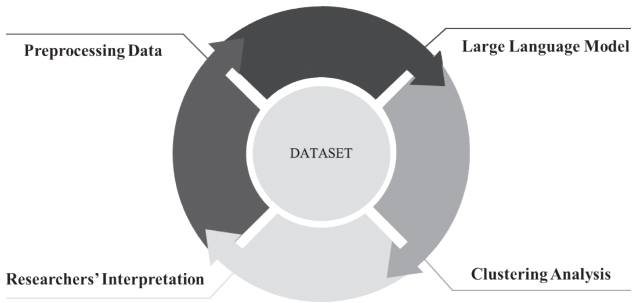

Figure 1. – Cyclic workflow for text analysis and clustering.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

In this section, we describe the step-by-step methodology used for analyzing and clustering qualitative data. The process involves several key phases, each of which plays a crucial role in ensuring the accuracy and relevance of the analysis (*Fig. 1*). In the following paragraphs, we will present these steps, applying them to a real-world dataset as an empirical illustration.

2.1. *Dataset creation and preprocessing*

The first step in this approach is the creation of the dataset for analysis. In the example discussed in this article, data were collected through a set of seven questions designed to explore the meanings teachers associate with various didactic activities, including group discussions, lecturing, collaborative learning, problem-solving, activities involving the creation or manipulation of physical and/or virtual objects, and those utilizing technology to distribute teaching materials or facilitate interaction between students and teachers. The questions were part of a questionnaire that also included inquiries about teachers' sensemaking, emotional responses, and a scale measuring burnout risk. Given the focus of this article, we will concentrate solely on the analysis of the seven aforementioned questions as an illustration of AI adoption in qualitative research. These questions asked participants to describe each didactic activity using a single word. This approach was chosen because open-ended questions requiring lengthy responses can be time-consuming for participants, potentially leading to higher rates of skipped questions or participant drop-out. By requesting a one-word answer, we aimed to gather a greater number of responses, allowing researchers to examine the range of meanings expressed by the participants. Ultimately, the research objectives prioritized collecting data on the diversity of meanings from the entire sample, rather than delving into the nuances of meaning-making that longer answers might provide. The survey was filled in by 379 Italian teachers.

Once the dataset is established, the next phase involves data cleaning. This process includes handling missing or null values, correcting typographical errors, removing duplicates, and filtering out irrelevant information. Specifically, in our case, we removed duplicate entries and converted all text to lowercase to avoid discrepancies due to capital letters or unusual formatting. After data cleaning, the dataset contained 384 words. For the analysis described in this article, part of the preparatory step involved translating terms from Italian to English. This decision was made because, despite increasing attention to training large language models in languages other than English, these models are predominantly trained on English

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

texts. As a result, they generally perform better in terms of accuracy, comprehension, and generating meaningful outputs in English. Since the aim of this pilot study was to establish the analysis procedure and evaluate the performance of AI as a tool for qualitative research – typically published in English – the research group concluded that this choice was appropriate. However, they acknowledged that some cultural nuances inherent in the Italian language might be lost in translation. Given the rapid advancements in AI, it is anticipated that, in the near future, LLMs will be trained in additional languages, enabling analyses to be conducted in the original language of the data, thus overcoming this problem.

In the next section we will discuss how we employed language models to cluster words and expressions that belonged to the same overarching categories of meaning. As mentioned above, the aim was to determine to what degree LLMs could cluster words in a way that was useful for the generation of meaningful categories.

2.2. *Model selection for embedding extraction and cluster analysis*

After preparing the data, an important step is the selection of the language model for embedding extraction. Choosing the right language model for a specific NLP task is crucial to achieving optimal performance and accurate results. Different models are tailored to excel in distinct types of applications, leveraging their unique architectures and training objectives. For instance, when the task involves semantic similarity analysis, such as clustering sentences by meaning or determining how closely two texts are related, models like Sentence-BERT (SBERT) are particularly well-suited (Reimers & Gurevych, 2019). On the other hand, tasks requiring text generation, causal reasoning, or open-ended question answering are better handled by models like Mistral, a general-purpose causal language model trained to generate coherent and contextually appropriate text (Jiang *et al.*, 2023). For more foundational tasks, such as word-level understanding, translation, or simple clustering, models like Word2Vec remain effective (Church, 2017). Word2Vec, with its simpler architecture, captures basic semantic relationships between words. By aligning the model's strengths with the task requirements, researchers and practitioners can ensure they are leveraging the most suitable tools to achieve their desired outcomes in NLP applications.

For this study, we employ Word2Vec, a model known for its simplicity and effectiveness in generating word embeddings by learning vector representations based on word co-occurrences in large text corpora (Church,

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

2017). Unlike other models, Word2Vec operates using an architecture specifically aimed at capturing the semantic relationships between words based on their co-occurrences in the data used for the training. Specifically, in our study, we use Word2Vec Google News 300 model, a pre-trained word embedding model that generates dense vector representations for words using the Word2Vec technique. This model was trained on approximately 100 billion words from Google News and produces 300-dimensional vectors for each word. Word2Vec was adopted to assist researchers in the task of generating interpretive categories able to capture the meanings that the sample of 379 teachers associates to the different didactic activities described in the questions.

By means of the clustering analysis, each word in the dataset is assigned to a cluster based on its vector representation. Once the clustering is computed, the researchers evaluate the similarity between words within the same cluster and between words in different clusters to assess the semantic coherence of the clusters. Words that are semantically similar are expected to cluster together, demonstrating the effectiveness of the Word2Vec model and the K-Means algorithm in capturing the underlying structure of the data. K-Means algorithm is a clustering algorithm that partitions data into *K* groups, assigning each point to the cluster whose centroid (mean point) is closest, with the goal of minimizing the sum of squared distances between points and their centroids. The process involves an initial selection of centroids, assigning points to the nearest cluster, and continuously updating centroids until they stabilize. However, the main challenge is choosing the optimal number of clusters *K*. The Elbow Method helps to determine the optimal *K* by plotting the sum of squared intra-cluster distances (inertia) against the number of clusters (k) (Bholowalia & Kumar, 2014). This method exists upon the idea that one should choose a number of clusters so that adding more clusters doesn't contribute to the analysis. The optimal number of clusters is found at the point where the inertia reduction slows down, forming an «elbow» in the graph. Once the optimal number of clusters is identified using this method, K-Means clustering can be applied to group the data points into the selected number of clusters.

For our dataset, we calculated the inertia across a range of clusters from 2 to 14. Our findings indicated a significant decrease in inertia with an increasing number of clusters. Lower inertia values suggest that the data points are more tightly grouped around their centroids, indicating more compact and well-defined clusters. Upon analyzing the differences in inertia between consecutive clusters, the elbow was identified at 7 clusters. Finally, K-Means Clustering was performed. In the present pilot study, we applied

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

K-Means clustering to group words based on their Word2Vec embeddings. These embeddings represent words as high-dimensional vectors, capturing their semantic meanings in the context of the data. K-Means allows us to group words with similar semantic properties. This method enabled us to identify word clusters that reflect shared semantic features. In particular, we retrieved the embeddings for each word by using them as inputs to the Word2Vec Google News 300 model. The list of *N* words serves as input for the embedding model, which returns an output of dimension (*N*, 300), where 300 represents the fixed dimension of the embedding space in the Word2Vec model. This matrix represents the projection of words into the model's 300-dimensional embedding space, with each row corresponding to the embedding vector for a specific word.

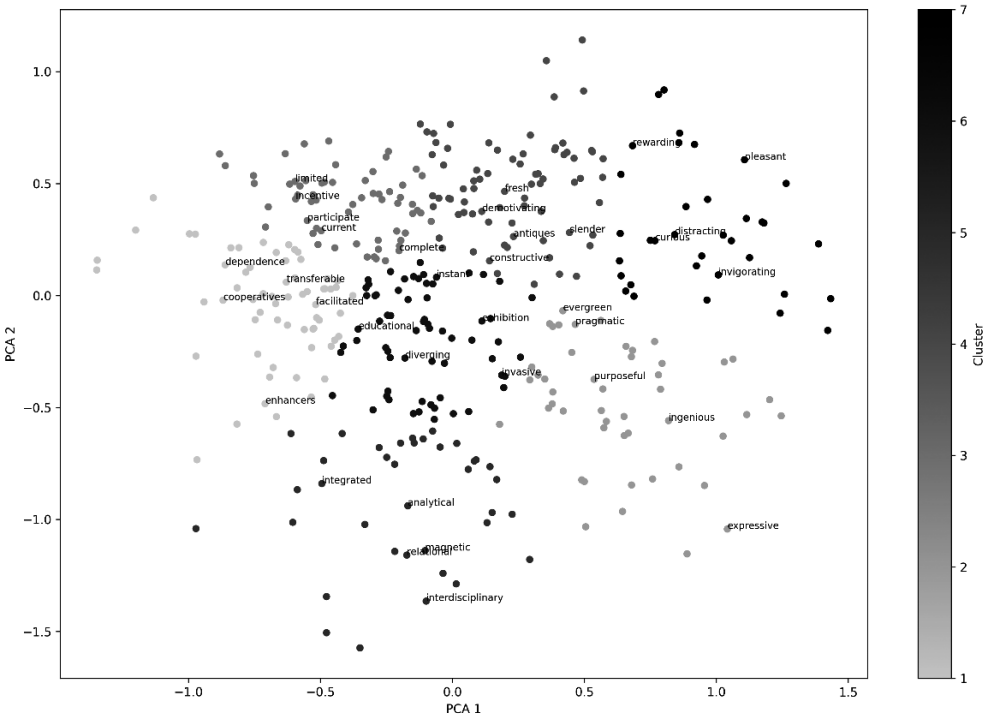

Figure 2. – Two-dimensional representation of word clustering using PCA and K-Means, based on Word2Vec embeddings. Each point represents a word, colored according to its assigned cluster. Labels highlight a few words from each cluster.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

The results of the words clustering with 7 clusters are visualized in *Figure 2*, which presents a two-dimensional representation of the word clusters. Indeed, we used PCA (Principal Component Analysis) to reduce the high-dimensional Word2Vec embeddings to two dimensions, facilitating the visualization of the clusters (Hasan & Abdulazeez, 2021). The plot displays each word as a point, differentiated by color according to their respective clusters, with a selection of words from each cluster labeled for enhanced interpretability.

2.3. *Assessing semantic similarity through cosine similarity and human interpretation*

The semantic similarity between each pair of words was assessed both with quantitative and qualitative methods. First, the cosine similarity coefficient was applied. Cosine similarity is a measure used to determine how similar two non-zero vectors are in an inner product space (Rahutomo *et al.*, 2012). This metric calculates the cosine of the angle between the two vectors, providing an indication of their orientation to each other, irrespective of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. In *Figure 3 (A)*, we present the average correlation between words belonging to the same cluster and between words belonging to different clusters. The results reveal that words within the same cluster exhibit the highest levels of similarity, demonstrating that words grouped into the same cluster share significant semantic overlap. This pattern of similarity suggests that the Word2Vec model seem to effectively capture relationships in word meanings, as reflected by the proximity of their vector embeddings. Additionally, the fact that words in different clusters show lower correlation highlights the model's ability to distinguish between distinct semantic constructs. These findings indicate that the language model to some degree groups semantically similar words together and separates different concepts, based on the underlying word embeddings.

Subsequently, the emerging clusters were qualitatively assessed based on the researchers' interpretation of the words in each cluster. This assessment involved coding the words within each cluster to identify the range of meanings present. However, from a qualitative research perspective, interpreting the meaning within each cluster proved challenging. Indeed, several codes were repeated across clusters and all the clusters seemed to contain several distinct semantic fields making it challenging to identify overarching categories able to synthesize effectively the range of meanings present within each cluster.

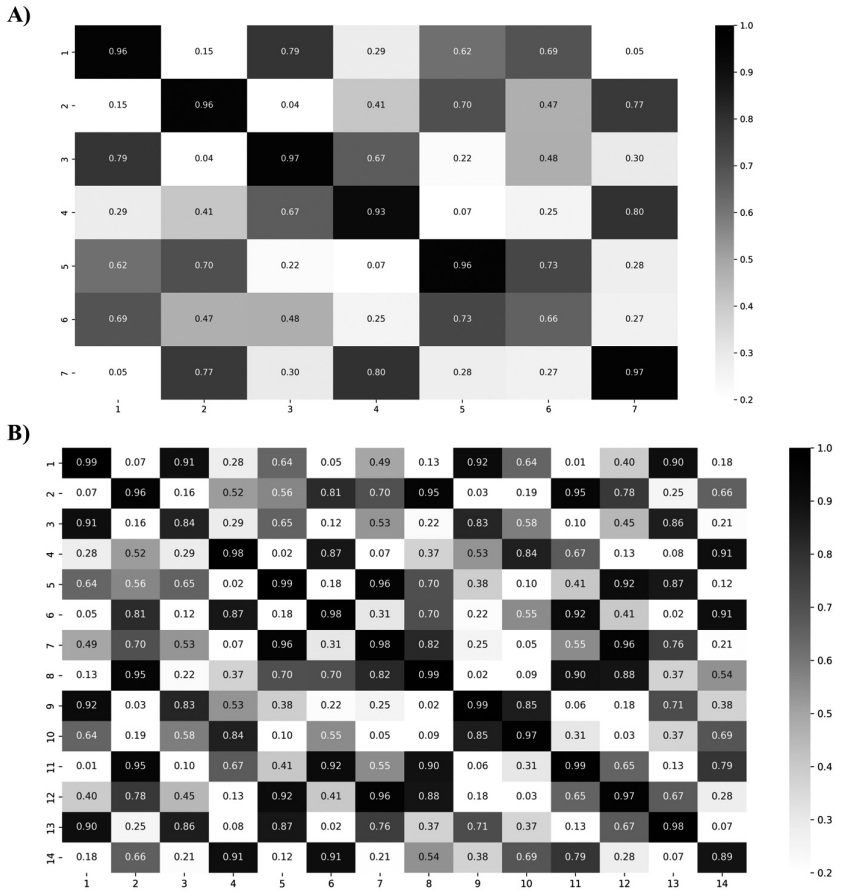

Figure 3. – Average Cosine similarity matrix across words for both the seven and fourteen clusters. The heatmap uses color gradients to depict different levels of similarity: darker shades represent higher similarity between clusters, while lighter shades indicate lower correlation. (A) shows the average Cosine similarity matrix for the seven clusters, where words within the same cluster exhibit stronger correlations compared to words across different clusters. (B) presents the average Cosine similarity matrix for the fourteen clusters.

Table 1. – Codes associated with each of 14 clusters.

Cluster 1	improvement
	facilitation
	participation
Cluster 2	productivity vs sterility
	classicity vs modernity
Cluster 3	newness vs obsolescence
	social
	essential vs comprehensive
Cluster 4	emotional response
Cluster 5	innovation
	funtionality
Cluster 6	exciting vs boring
Cluster 7	systematic vs disordered
	experiential vs transmissive
Cluster 8	empathic vs monotone
	multidimensional vs reductive
Cluster 9	limits
	necessities
	incentives
	participation
Cluster 10	challenge vs routine
	profitable vs ineffective
Cluster 11	interesting vs banal
Cluster 12	dynamic vs static
Cluster 13	facilitation
	active methods
Cluster 14	motivating vs demotivating

It was hypothesized that the number of clusters identified with the Elbow Method was too low for an effective contribution to qualitative interpretation by the researchers. To address this, an iterative process was initiated to refine the number of clusters used in the K-Means algorithm. In each iteration, the number of clusters was incrementally increased (N+1), until a configuration was reached that produced more interpretable results. The iteration with 14 clusters was deemed most satisfactory (*Tab. 1*) for

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

facilitating qualitative analysis although some overlap in meaning across clusters persisted (e.g., «participation» was coded in both Cluster 1 and Cluster 9), and a few outlier words remained. This similarity in meaning between clusters also became evident when average cosine similarity between words belonging to the same cluster and between words belonging to different clusters for 14 clusters, as reported in *Figure 3 (B)*. Indeed, the results showed good cohesion within the clusters (intra-cluster similarity), with high values along the diagonal indicating strong semantic affinity between words in the same cluster, as expected. However, some clusters showed high correlations with each other (cross-cluster correlation), suggesting the presence of shared concepts; for example, Cluster 1 exhibited high similarity with Cluster 9 and Cluster 3. On the other hand, some pairs of clusters, such as Cluster 1 and Cluster 5 or Cluster 2 and Cluster 10, displayed very low correlations, indicating that the words in these groups are semantically distinct and reflect separate concepts. While some overlaps in meaning were present, researchers were able to identify up to four distinct categories of meaning for each cluster, adequately representing the range of meanings in the data. Nevertheless, further efforts will be required to refine the categorial structure, aligning it more coherently with the study's goals and research questions. As follows we briefly discuss the categories of meaning that emerged from the analysis and the methodological issues emerging for the qualitative interpretation of the results.

First, while some clusters contained words associated with two opposing poles of meaning (e.g., modernity vs classicism), others seemed to include multiple independent categories of meaning (e.g., improvement, facilitation, and participation in Cluster 1). In *Table 1*, the notation «vs» was used to indicate when two opposing poles within the same category of meaning were identified. Depending on the research objectives, these opposing poles could either be treated as a unified category or considered as distinct categories.

Second, the clusters varied significantly in size, with the smallest (Cluster 11) containing only nine words and the largest (Cluster 3) containing 53 words. This variation influenced the number of categories identified within each cluster. Initially, the research group hypothesized that increasing the number of clusters might resolve this issue; however, this was not the case. Even when the number of clusters was increased to 20, the sizes remained uneven, ranging from 5 to 35 words, and several clusters still contained multiple categories of meaning, similar to the iteration with 14 clusters. This analysis suggests that establishing a generalizable method for determining the optimal number of clusters remains difficult. Nonetheless, the pilot study suggests that it is unlikely that a number of clusters

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

lower than that identified by the Elbow Method would suffice for a coherent identification of meaning categories. In addition, the matrix of cosine similarity seems to provide a good quantitative measure able to inform and/or confirm the validity of the decision on the number of clusters as discussed above.

Third, some categories overlapped across clusters. One approach to address this issue is to merge overlapping categories into overarching categories within the final category tree. However, the fact that words from the same category of meaning appear in different clusters may indicate that they are used in distinct contexts, or that they tend to systematically co-occur with different words, which could provide valuable insights for researchers. For example, in comparing Cluster 1 and Cluster 9 (*Tab. 1*), «participation» appears in both clusters. While it is possible to merge the words related to «participation» into a single category, it might be useful for some research purposes to further explore the difference between the words belonging to this category by identifying subcategories that capture subtle differences in the meanings associated with «participation» across the two clusters. The fact that these words appear in different clusters might reflect the fact that some words linked to «participation» are typically connected to improvement and facilitation, while others are typically associated with limitations, necessities, or incentives. These nuanced differences may hold relevance for the research aims, which could be overlooked in analyses conducted without AI support. In this sense, AI can enrich the researchers' interpretation if the qualitative analysis conducted by the researchers leads to confirmation of the validity of this AI-augmented interpretive process.

In sum, although the use of AI for the qualitative analysis of one-word responses in surveys presents complexities and limitations, it can be valuable for two main reasons. First, the AI-generated clusters may reveal implicit associations between word meanings, enriching the interpretation process. In this respect, AI can be viewed as an additional coder offering a different perspective that provides insights that researchers can reflect on to identify relevant categories of meaning. These categories would then be based on a synergy of human interpretation and automated clustering, potentially highlighting relationships that researchers might not readily detect. Second, the experience of the research group shows that AI-supported data preprocessing and clustering streamline the coding process by partially automating data cleaning and organization. This allows researchers to focus more on interpretive processes and less on labor-intensive tasks such as the cleaning and organizing of data taking place before the generation of categories.

These findings emphasize the importance of researchers' qualitative examination of clusters to assess whether they support the identification

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-184-1

of key themes and concepts and to identify patterns or discrepancies that may require refinement or adjustment. For example, words that appear to be incorrectly clustered based on their original context might need reclassification, or they may point to limitations in the model's ability to capture subtle meanings. This combination of machine-driven clustering and human interpretation facilitates a robust understanding of the semantic relationships in the dataset, ensuring that the final analysis is both data-driven and contextually meaningful.

3. Discussion and conclusions

In this article, we aimed to present a methodology for supporting the analysis of textual data. Our goal was to provide an approach that can be applied to similar datasets, discussing the potential and limitations of the approach. We demonstrated how these techniques can be applied in practice by using a real-world dataset, showcasing each step from data preparation to the final clustering of the words. The methodology involved several key stages: data creation, data cleaning, model selection for embedding extraction, application of the Elbow Method, K-Means clustering analysis and Qualitative interpretation of the results. The analysis showed that a model like Word2Vec can support the development of insights for qualitative interpretation. However, it is important to recognize both the advantages and limitations that artificial intelligence can bring to the analysis (Jafari *et al.*, 2024). On the positive side, AI can enable the analysis of large volumes of data and allows to enrich interpretation through the identification of patterns and themes that might be difficult for humans to detect, especially in extensive datasets. However, the use of AI in qualitative research is not without its drawbacks. One major concern is the potential loss of context and depth that is often central to qualitative analysis. AI algorithms, while powerful, may not fully grasp the nuanced meanings and subtleties inherent in human language and interactions. This limitation can lead to an oversimplification of complex social phenomena or the overlooking of critical insights that a human researcher might discern. For this reason, in this article the importance of the qualitative interpretation of the data by the researchers is emphasized. Reflection is needed to assess relevance and significance of AI clustering for enhancing and/or facilitating interpretation. In this sense, there is a risk of over-reliance on AI-generated results, which could lead to a detachment from the data and a reduced emphasis on the interpretive role of the researcher. Furthermore, ethical considerations

arise regarding data privacy and the transparency of AI processes. The «black box» nature of some AI systems can obscure the decision-making process, making it challenging to understand how conclusions are derived and to ensure the ethical handling of sensitive information (Pedreschi *et al.*, 2019). While AI can greatly aid in managing and analyzing qualitative data, it is imperative to strike a balance, leveraging AI's strengths while remaining vigilant about its limitations. Researchers should use AI as a tool to complement, not replace, the critical and interpretive skills that are at the core of qualitative research. Further research is needed to assess the generalizability of the method for similar research projects (for example by refining the criteria for the determination of the number of clusters) and to test the adoption of AI with different types of data such as open-ended answers that involve more complex texts than single words.

References

Abram, M.D., Mancini, K.T., & Parker, R.D. (2020). Methods to integrate Natural Language Processing into qualitative research. *International Journal of Qualitative Methods*, *19*. https://doi.org/10.1177/1609406920984608

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on Elbow Method and k-means in WSN. *International Journal of Computer Applications*, *105*(9).

Christou, P.A. (2023). How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? *Qualitative Report*, *28*(7).

Church, K.W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.

Clarke, N., Foltz, P., & Garrard, P. (2020). How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *129*, 446-463. https://doi.org/10.1016/j.cortex.2020.05.001

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Creswell, J.W., & Poth, C.N. (2017). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). London: Sage.

Dey, I. (2003). *Qualitative data analysis: A user friendly guide for social scientists*. London: Routledge.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Hasan, B.M.S., & Abdulazeez, A.M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20-30.

Jafari, F., & Keykha, A. (2024). Identifying the opportunities and challenges of artificial intelligence in higher education: A qualitative study. *Journal of Applied Research in Higher Education*, *16*(4), 1228-1245.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D.L., …, & Sayed, W.E. (2023). Mistral 7B. *arXiv preprint*. https://doi.org/10.48550/arXiv.2310.06825

Manning, C.D. (2022). Human language understanding & reasoning. *Dædalus*, *151*(2), 127-138. https://doi.org/10.1162/daed_a_01905

Natural Language Processing (2020). https://www.ibm.com/cloud/learn/natural-language-processing

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. *Proceedings of the AAAI (Association for the Advancement of Artificial Intelligence) Conference on Artificial Intelligence*, *33*(1, July), 9780-9784.

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic cosine similarity. *International Student Conference on Advanced Science and Technology – ICAST. Proceedings*, *4*(1, October), 1. University of Seoul (South Korea).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint*. https://doi.org/10.48550/arXiv.1908.10084

Ritella, G., Loperfido, F.F., De Giglio, G., Scurani, A., & Ligorio, M.B. (2022). Adopting educational robotics and coding to open dialogic spaces in lower secondary education. *Dialogic Pedagogy: A Journal for Studies of Dialogic Education*, *10*, DT41-DT58.

Valdenegro, D. (2023). A large language models digest for social scientist. *SocArXiv*. https://doi.org/10.31235/osf.io/m74vs

Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv*, 2304.04339.

Riassunto

L'analisi qualitativa è essenziale nella ricerca in diversi campi, offrendo approfondimenti approfonditi che spesso non possono essere catturati tramite metodi quantitativi.

Tuttavia, la gestione di grandi volumi di dati qualitativi presenta delle sfide, tra cui la sua natura laboriosa e il potenziale di pregiudizi interpretativi. In questo studio, introduciamo e mostriamo una metodologia passo dopo passo che integra l'intelligenza artificiale (IA) nell'analisi dei dati qualitativi, con un focus sulle risposte testuali estratte dalle domande del sondaggio. Nello specifico, il nostro approccio impiega tecniche di IA, utilizzando Word2Vec per l'estrazione degli embedding dalle parole e il clustering K-Means per semplificare l'analisi dei dati testuali qualitativi, integrando infine l'interpretazione del ricercatore dei cluster identificati per migliorare la pertinenza dell'analisi. Inoltre, il presente articolo discute la pertinenza e il significato di questo approccio, nonché le sue sfide etiche e metodologiche, mediante un'illustrazione empirica tratta da uno studio sul sense-making degli insegnanti in merito a una gamma di diverse attività educative.

Parole-chiave: Clustering; Dati testuali; Embeddings; Intelligenza artificiale; Ricerca qualitativa.

How to cite this paper: Luongo, M., Ponticorvo, M., Ligorio, M.B., Crescenzo, P., & Ritella, G. (2024). Artificial intelligence to enhance qualitative research: Methodological reflections on a pilot study [L'intelligenza artificiale per potenziare la ricerca qualitativa: riflessioni metodologiche su uno studio pilota]. *Journal of Educational, Cultural and Psychological Studies (ECPS)*, *30*, 118-135. https://doi.org/10.7358/ecps-2024-030-luon