Quality & Quantity
https://doi.org/10.1007/s11135-026-02894-9

Natural language processing for data analysis: an application on the well-being

Samuele Boi[1] · **Nicola Tedesco**[1] · **Luisa Salaris**[1]

Received: 31 December 2025 / Accepted: 28 May 2026
© The Author(s) 2026

Abstract
This article presents a case study on the analysis of semi-structured interviews conducted in Italy, using a small dataset. Two supervised approaches are applied to identify key questions to retrieve information and compare responses to the same questions across different respondents. The first approach is based on a bag-of-words model, while the second relies on embeddings. These approaches are compared with two topic modeling methods (LDA and BERTopic). The results highlight the differences between the methods: key-question-based approaches seem to be more suitable when the goal is to compare responses to specific questions, whereas topic modeling techniques are better suited for identifying latent topics.

Keywords NLP · Semi-structured interviews · Conversation analysis · LDA · BERTopic · Well-being

1 Introduction

Text analysis is often based on comparisons among words within and between corpora. For this reason, it is often preferable for documents to be homogeneous in format and length, such as collections of tweets or patents. Moreover, a large number of observations enables analyses such as topic identification and comparisons of term distributions. The homogeneity and size of the corpora are the bases of many classification-based tasks. Classification-based tasks do not refer only to the process of labeling documents but also to identifying the

All the authors have contributed equally to this work.

✉ Samuele Boi
samuele.boi@unica.it

Nicola Tedesco
tedesco@unica.it

Luisa Salaris
salaris@unica.it

[1] Department of Political and Social Sciences, University of Cagliari, Cagliari, Italy

Published online: 10 June 2026

sentiment of texts in a corpus (sentiment analysis) or determining the main topic of a group of articles (topic classification). Corpus homogeneity and size also play crucial roles in unsupervised methods such as topic modeling, where document topics are identified without prior labels.

The situation becomes more problematic when documents are not homogeneous. For example, combining short stories and tweets within the same dataset is generally not advisable. This is not only due to differences in language style but also to the large discrepancy in document length and, consequently, in term distributions. Indeed, when metrics based on term occurrences are applied, tweets and short stories exhibit fundamentally different statistical properties. As a result, methods such as term frequency–inverse document frequency (TF–IDF) (Sparck Jones 1972; Salton and Buckley 1988; Goldberg 2017; Vechtomova 2009) may not perform well in such cases.

Semi-structured interview corpora represent a particularly challenging data structure in data analysis. Treating each interview as a single independent document may not be optimal. In fact, interview-based analyses can yield superficial results that fail to capture certain types of information. This is mainly because surveys typically include a limited number of interviews, while each interview contains a large amount of information. Furthermore, interviews are structured around questions, and ignoring the relationship between questions and answers can result in information loss or misleading interpretations. For example, some responses do not repeat the subject mentioned in the question. In such cases, treating only the responses can lead to information loss. However, merging questions and answers may introduce bias or noise, since each response would include similar questions, and the responses may submerge the information in the questions.

Moreover, to retrieve high-quality information, it may be useful to capture differences among responses within each interview. However, to do so, interviews cannot be treated as independent units of a corpus.

Therefore, selecting an appropriate level of analysis is crucial.

For these reasons, in this analysis, interviews are treated as collections of responses, allowing comparisons within each interview, rather than treating interviews as independent units.

This approach is later compared with two topic modeling methods: Latent Dirichlet Allocation (LDA) (Blei 2012) and BERTopic (Grootendorst 2022). This is because the general subjects of an interview are often known in advance and may correspond directly to the questions posed to respondents. In other words, it is worth investigating whether the topics identified by topic modeling align with the subjects addressed in each question, or whether latent topics emerge.

This article proposes a case study comparing two Natural Language Processing (NLP) pipelines for topic modeling in the context of structured domain-specific interviews with a limited dataset. In particular, the NLP pipelines aim to:

(i) Facilitate information extraction, particularly when comparing answers from different interviews to the same questions;
(ii) Preserve information about which interview question each sentence originates from when applying methods such as clustering or topic modeling.

The article is structured as follows: first, the literature is reviewed; second, an introduction to the data is provided, and the main challenges involved in analysing interview data are discussed; then, two methodologies for matching interview responses with template responses are presented, followed by two standard topic modeling approaches. These methods are then applied to the data in the analysis section. Finally, the results of these methods are compared.

2 Literature review

Qualitative analysis plays a crucial role in social and health-related research (Miles et al. 2013; Kuckartz 2014; Merriam and Tisdell 2015; Leeson et al. 2019; Brinkmann and Kvale 2018; Talmy 2010; Braun and Clarke 2006). Traditional qualitative analysis allows researchers to delve into interview responses through content analysis, narrative analysis, hermeneutics, and discourse analysis (Brinkmann and Kvale 2018).

However, because standard qualitative analyses are labour-intensive, researchers use NLP models or qualitative text analysis software to facilitate their work.

Regarding NLP techniques, Guetterman et al. (2018) compare traditional qualitative analysis conducted by researchers with NLP approaches. They find that traditional qualitative text analysis is better at identifying nuances compared to automated NLP approaches. However, NLP approaches are useful when the dataset involves a large number of interviews. The authors also note that few methodological researchers have examined NLP from a qualitative perspective.

Regarding Semi-Structured Interview (SSIs) Analysis, Parfenova (2022) compares different topic modeling techniques. The best method involves using Bidirectional Encoder Representations from Transformers (BERT) embeddings, dimensionality reduction (UMAP), and clustering (HDBSCAN).

Other articles focus on the role of the interviewer (Quillivic and Payet 2024). Indeed, sometimes the analysis focuses solely on the interviewees' responses, while ignoring the interviewer's questions. For example, the interviewer's role is analysed by considering the use of pronouns from the interviewer (Bonneau and Dister 2010), which raises the possibility of different ways the interview can be conducted. Another study on the role of the interviewer focuses on differences by using Reinert's method Scelles 1997, a popular textual clustering method, across corpora with and without interviewer interventions (Dalud-Vincent 2010).

Quillivic and Payet (2024) conducted 926 semi-structured interviews to study the linguistic coordination (or linguistic alignment), which refers to the alignment of communication among participants in a conversation. The authors quantify the alignment of discourse between the interviewer and the interviewee using a few-shot learning classifier based on cosine similarity between word embeddings.

The authors specifically focus on gender configuration (interviewee - interviewer) differences and also compute quantitative measures to examine other associations, for example metadata usage, using t-tests. The use of metadata is possible due to the large number of interviews in the dataset.

Other similar articles on linguistic coordination study gender differences (Danescu-Niculescu-Mizil and Lee 2011), linguistic style (Doyle and Frank 2016), and semantic alignment using word embeddings (Nasir et al. 2019).

Regarding limited datasets, one popular solution is data augmentation (Zou 2019; Kobayashi 2018; Sennrich et al. 2016). Textual data augmentation can be performed using different approaches. Zou (2019) propose four operations: synonym replacement, random insertion, random swap, and random deletion. Kobayashi (2018) proposes contextual augmentation, where words are stochastically replaced with others based on context. Sennrich et al. (2016) use back-translation of monolingual paired training data as additional parallel training data.

Data augmentation is now also extended to generative AI (Chintagunta et al. 2021; Argyle et al. 2023; Hämäläinen et al. 2023). However, if the original interviews are conducted with a panel of selected experts, it would not be a good strategy to mix experience-based, context-specific information with generic AI-generated information. Even if the generative process is only intended to expand the interview dataset, it should be noted that many NLP processes rely on comparisons, and AI-generated information may influence the analysis of the original data. This, however, does not account for potential biases in the generative process, which may be reduced through more specific prompting.

Regarding NLP and semi-structured interviews, Argyle et al. (2023) use language models to create silicon samples from thousands of socio-demographic backstories and compare silicon and human samples.

A third way to improve performance with limited data involves few-shot learning (Brown et al. 2020). Instead of generating synthetic data for training, few-shot learning aims to provide the model with a few demonstrations of the task to improve performance. Unlike fine-tuning, few-shot learning does not allow weight updates.

Lison et al. (2021) propose a solution to address both the limited availability of textual data and privacy concerns by developing an automatic de-identification model.

Researchers also make use of qualitative text analysis software such as Atlas.ti, NVivo and MAXQDA. These tools usually allow thematic analysis through coding, which involves labeling text segments and grouping them by theme.

Finally, another approach may involve using LLMs to summarise and retrieve information from interviews. LLMs may be fine-tuned for qualitative and interview analysis. Such methods must address ethical issues related to data management (privacy). However, this problem may be mitigated by running such models locally. Certain qualitative data analysis software is already AI-powered.

3 Data

The interviews used in this analysis were conducted with managers and/or professionals from third-sector organizations that provide goods and services to people 65 years and older (hereafter referred to as seniors). These interviews constitute qualitative data from

the SENIOR project.[1] The project included two questionnaires. Questionnaire 1 focuses on monitoring the needs of older adults, while Questionnaire 2 concentrates on monitoring the purchasing behaviours of goods and services.

The survey structure for the interviews in the SENIOR project included 14 questions, divided into a general and a specific part. The general part consists of six questions that explore the background and experience of the respondents, the hierarchy of needs of seniors, the main difficulties faced by the elderly, and the changes that have occurred in recent years as well as those that may occur in the future. As mentioned, the respondents are managers and/or professionals from third-sector organizations, so their experience is highly informative and unique.

The specific part is designed to ask questions about different spending areas. For example, respondents are asked which factors most influence the purchase of a particular good or service in the following four areas: food products and supplements, personal care and health products, technological products, and leisure products/services.

The remaining seven questions concern who makes the decisions when a certain good or service is purchased, who manages the budget for the four spending areas, which products are purchased and where, and whether there is any seasonality in product purchases. Finally, respondents were asked whether they would like to add any additional information or suggestions.

It is important to note that both questions and answers vary across interviews, depending on the choices of both respondents and interviewers. For example, respondents may suggest certain topics, modify the interview structure, or provide additional information in response to the same questions.

As mentioned above, this dataset differs from other document collections. There are three main issues to consider. These issues were determined by both the nature of the interviews and their structure and administration.

They can be summarised as follows: heterogeneity of word distributions both between and within interviews, interdependence between questions and answers, and finally the limited size of the dataset (12 interviews with a median of 1667 words per interview), which is characterised by a domain-specific language (e.g., words such as ASL, which could be translated as Local Health Authority, and INPS, which is the National Social Security Institute) and by the Italian spoken language style.

Regarding the heterogeneity of word distributions, three different levels can be considered. First, the number of words varies between interviews. This variation is usually greater among respondents than among the interviewers.

Within each interview, there is also heterogeneity in the number of words across the various answers. This heterogeneity also exists in the questions, though to a lesser extent. For example, the answers to some questions, especially from certain respondents, consist of a single word, such as “Yes.” Other answers, as in Interview 13, can exceed 400 words (Figs. 1, 2).

Third, the number of exchanges (i.e., question–answer pairs) also varies across interviews.

[1] The paper was produced using data collected within the framework of the project funded by the European Union – Next Generation EU – Project “Age-It - Ageing well in an ageing society” (PE0000015), CUP H73C22000900006, Spoke 6, PNRR – PE8 - Mission 4, C2, Investment 1.3. However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

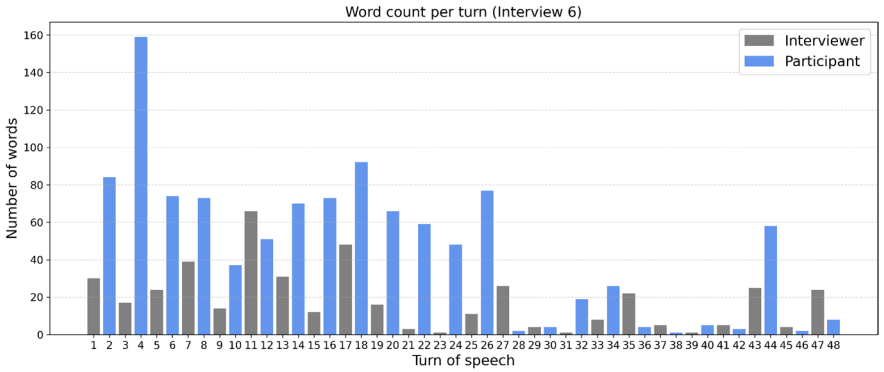

Fig. 1 Number of words per turn (Interview 6)

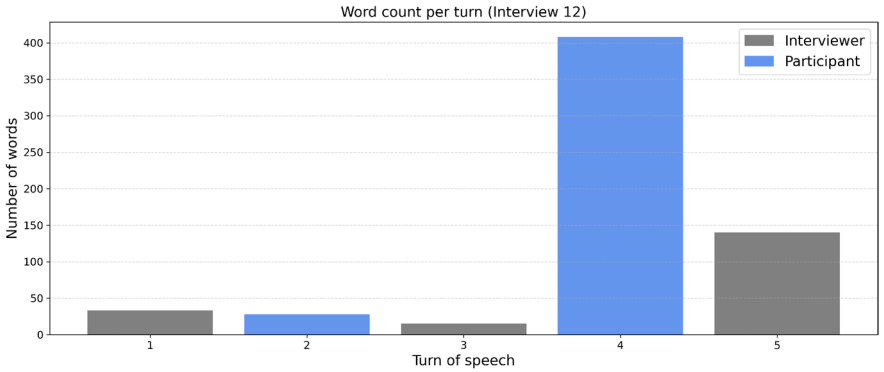

Fig. 2 Number of words per turn (Interview 12)

For example, Interview 12 contains a high number of exchanges (42), while others, such as Interview 1, have only 16. However, despite the high number of exchanges, the interviewee in Interview 12 (about 720 words) spoke a similar number of words as in Interview 13 (about 580 words), which had only three exchanges. These values differ greatly from Interview 1, which has 16 exchanges and about 2890 words spoken by the interviewee.

The ratio of words produced by the interviewee to those produced by the interviewer is also noteworthy: it is around 1 in Interview 12 (720 words each) and exceeds 5 in Interview 1 (2890 words for the interviewee vs. 530 for the interviewer). Heterogeneity between interviews also includes the limitation of Interview 13 to the general section at the interviewee's request, as well as the paper-based format of Interview 6 (Fig. 3).

Such differences in exchange length across interviews can be addressed using a long-form dataset.

A second issue, related to the nature of the interviews, concerns the non-independence of question and answer information. This issue is particularly problematic for very short answers. A frequently occurring answer is "Yes." However, the relationship between turns extends beyond single question–answer pairs. Consider, for instance, the question asked by the interviewer in Interview 11: "Ok, so is it the same for personal care products as well?"

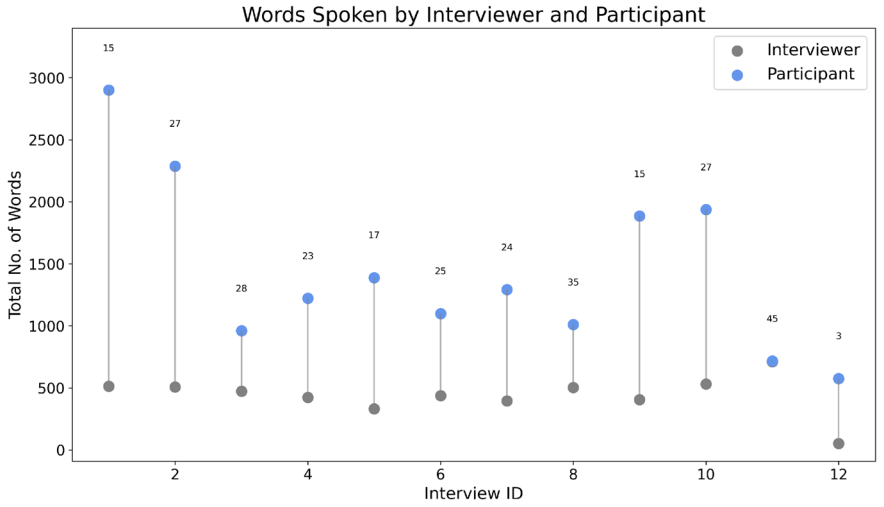

Fig. 3 Number of words per turn (Interview 6)

This question follows the budget question but refers to a different category. Without considering the proximity or dependency between questions, such information might be lost. Similar cases can occur with questions that confirm what has been heard or further explore an answer.

Finally, the dataset size should be considered, as it includes only 12 interviews. The small size of the dataset poses several challenges, including difficulties in identifying key topics in the interviews (topic modeling), grouping interviews by similar themes, and analysing respondents' perceptions of certain subjects (sentiment analysis). Moreover, the text contains many technical terms (e.g., ASL, INPS), is in Italian, and is written in a spoken style. These issues pose challenges both when using pre-trained embedding models and when fine-tuning embeddings on the text.

4 Methodology

4.1 Methodological framework

The article proposes two methods to standardize interview questions. Given the issues discussed above, the analysis focuses on the questions rather than on the entire interviews. The main idea is to match the questions asked in each interview to a set of template questions used by the interviewer to structure the survey. This alignment enables the extraction and comparison of information corresponding to the same question across different interviews. Unlike approaches that aim to discover latent topics in an unsupervised manner, interviews explicitly impose topics through their questions. Consequently, the core of the analysis lies in comparing the answers associated with each question. Moreover, as previously noted, certain specific questions may be excluded in standard topic modeling approaches, whereas the proposed method preserves question-level information. From a methodological perspec-

tive, questions also tend to be more homogeneous than answers, making them more suitable for alignment.

To perform the question mapping, two approaches are considered. The first adopts a bag-of-words representation, while the second relies on sentence embeddings derived from different pre-trained models. In addition, a Latent Dirichlet Allocation (LDA) model (Blei 2012) and a BERTopic model (Grootendorst 2022) are used to analyse differences in the results. In particular, the analysis examines whether the topics identified through topic modeling are reflected in the template questions.

For ease of interpretation, the results originally collected in Italian are translated into English.

4.2 Bag-of-words model

The first approach uses a bag-of-words (BoW) model, which represents each document by counting word occurrences and producing a Document–Term Matrix, where rows correspond to documents and columns to terms. Each document is then represented as a vector of term frequencies.

BoW models are often combined with term frequency–inverse document frequency (TF-IDF), a weighting scheme that reflects the importance of words by combining their frequency within a document with their rarity across the corpus.

The main steps of this approach are as follows: splitting questions into sentences, preprocessing, dimensionality reduction, clustering, and training a supervised model to classify interview questions based on the template questions. As mentioned later, clustering serves two purposes: first, to provide a general idea of how sentences are grouped by an unsupervised approach, which helps label sentences for the supervised model; second, to serve as features included in the model.

First, only the questions were extracted from the interviews and the interview template.

Second, the TF-IDF matrix was computed by treating each sentence as an individual document.

The third step involves clustering. To improve clustering, some metadata variables (number of words and presence of a question mark) were included as features alongside the TF-IDF features. Moreover, before performing the clustering procedure, a dimensionality reduction method was applied. Among the available techniques, Truncated Singular Value Decomposition (TSVD), also referred to as Latent Semantic Analysis (LSA) in the NLP literature, is a popular option since it is well-suited for sparse matrices such as those generated by TF-IDF (Deerwester et al. 1990; Dumais et al. 1996; Jurafsky and Martin 2008).

Improvements in results using LSA may be due to features of the data, such as the degree of similarity among sentences. LSA may squeeze out redundancy and noise (Dumais et al. 1996). This is expected to improve performance due to the presence of similar but not identical sentences.

Fourth, the questions were clustered in order to identify groups of similar questions across different interviews. It should be noted that some questions in the dataset are highly similar. However, this is often the case with semi-structured interviews. The degree of similarity depends on both the interview's topics and the interviewer's attitude toward the respondent.

A popular method for choosing the number of clusters is the silhouette score. This metric measures how closely a data point belongs to its own cluster (cohesion) compared to other

clusters (separation) (Tatsat et al. 2020). Its value ranges from -1 to 1. However, due to factors such as the interview subjects and the interviewer's attitude, it is recommended to inspect the results qualitatively in addition to relying on the silhouette score when determining the number of clusters (Tatsat et al. 2020).

Similarly to the process used for clustering, two supervised models were trained. The first uses only TF-IDF with Truncated SVD and metadata features (number of words and presence of a question mark), while the second adds these variables to the clustering results, treating them as labels obtained from the unsupervised approach described above.

Since the clustering results, which are later used as features, were trained on the entire dataset, they may lead to data leakage issues. However, since the aim is to classify sentences in this dataset without predicting on out-of-sample data and given the limited amount of data, this feature was included in the second model.

In other words, the clustering step has two main purposes. First, it shows how an unsupervised model groups the questions, which is useful for assigning labels for the supervised process. Second, the clustering results may be included as a feature in the supervised model. In this case, it is important to check for correlations between the clustering-based features and the other variables. Furthermore, it is possible to compare models that include the clustering feature with those that do not, to assess whether the supervised model aligns with the clustering process and to adjust the labels accordingly. This step is especially useful in cases involving limited data and peculiar semi-structured interviews that require a customized analytical approach.

Since the list of questions should be known to the interviewer, such template questions also play a crucial role in defining the labels for the supervised model. Indeed, the clustering may incorrectly identify every template question, or it may identify some irrelevant questions.

Subsequently, approximately 20% of the entire dataset was manually validated. Due to the relatively large number of labels (one per template question), manual validation was balanced across them.

To identify a model for classifying unlabeled questions, a stratified 5-fold cross-validation was performed. A stratified 10-fold cross-validation was not feasible, as stratified cross-validation requires each label to have at least as many samples in the training set as there are folds. This issue may arise when there are few examples for certain questions.

Furthermore, in order to capture irrelevant questions, a class labeled 0 was introduced. The idea is to classify questions that differ from the template questions, grouping them into a residual cluster rather than forcing the model to classify each sentence into one of the template question categories. However, since the non-relevant sentences in the validation set vary widely across the data, it should be noted that the models may learn to assign a higher proportion of 0 labels than necessary for non-heterogeneous questions. Moreover, it should be considered that, during the validation phase, the number of instances assigned to label 0 was proportionally higher than that of the other labels. Nevertheless, sentences assigned to label 0 were clearly distinguishable from the template questions.

The models used in the cross-validation are the following:

- Random Forest
- Decision Tree
- Logistic Regression

- Linear SVC
- XGBoost

Once the best model was selected, it was used to predict the labels of previously unlabeled sentences.

Subsequently, the classified sentences were aggregated back to the question level. Since each sentence was assigned a label, the final label for each question was determined by majority vote, selecting the label that appeared most frequently among its constituent sentences. Sentences assigned the label 0 were excluded from this procedure, as this label was reserved for non-relevant content. This choice reflects the fact that the interviewer often repeated or rephrased parts of the same question to provide clarification.

The final dataframe has the same dimensions as the original dataframe, but each question and its corresponding response are now classified. Given this structure, it is possible to extract information on a specific topic for a given question and systematically compare responses across respondents. For example, the different responses to Question 7–concerning the factors that most influence product purchasing decisions–can be analysed and contrasted between participants.

4.3 Embeddings

The second approach uses embeddings to match interview questions to template questions. Embeddings are vector representations that capture the semantic meaning of words. Because embeddings encode semantic information in a continuous vector space, it is possible to measure the distance or similarity between two word representations (Goldberg 2017).

Cosine similarity is a widely used metric for semantic similarity analysis and computes the cosine of the angle between two vectors. It is defined as the ratio of the dot product of two vectors, a and b, to the product of their magnitudes (or norms), as shown in Equation (1). Cosine similarity can be interpreted as a normalized dot product, making it insensitive to differences in vector length. As noted by Curran (2003), the dot product alone "does not account for the length of each vector, and therefore tends to favor longer vectors." The cosine similarity score typically ranges from -1 to 1, although in some applications it is normalized to fall within the interval [0,1].

$$\mathrm{cosine\_similarity}(\mathrm{a}, \mathrm{b}) = \frac{\mathrm{a}^\top \mathrm{b}}{\|\mathrm{a}\| \, \|\mathrm{b}\|} \tag{1}$$

For example, consider the following two questions:

- Do the elderly allocate any budget to technology?
- Do seniors spend any money on smartphones or tablets?

These questions do not share any relevant terms, yet their meanings are closely related. Indeed, the cosine similarity between these two questions–computed using pre-trained embeddings (paraphrase-multilingual-MiniLM-L12-v2) without additional pre-processing–is approximately 0.75, which is relatively high.

The process of matching template questions to interview questions consists of the following steps: question splitting, pre-processing, embedding generation, cosine similarity-based matching, and refinement of the matching using a supervised approach.

With regard to pre-processing, standard techniques were applied, including stop-word removal, lemmatization, and part-of-speech (POS) tagging. As in the previous approach, the stop-word list was manually revised to exclude words such as "where" and "when", which may carry relevant semantic information in this context. Subsequently, questions were transformed into vector embeddings. This procedure was performed at both the question and sentence levels.

The embedding model employed was SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2"), a multilingual pre-trained model.

As with the BoW model, a supervised approach was employed to refine the matching results. Consistent with the previous method, 20% of the dataset was manually validated, and a 5-fold cross-validation was performed to compare the performance of different models.

In this approach, however, the features include the embeddings of interview questions, obtained by splitting them into sentences, the embedding of template interview questions, the cosine similarity between each template question and sentence-level interview questions, and the normalized turn order of the question. The models used in the cross-validation, identical to those used in the BoW-based method, were trained using an 80/20 train-test split.

Unlike the BoW model, which provides a classified original dataframe, the Embedding approach establishes a one-to-one mapping between each template question and interview sentence. However, to achieve one-to-one matching, only the pair with the highest cosine similarity between the interview sentence and the template question is selected.

4.4 Topic modeling

The following section presents a more standard topic modeling approach. Topic modeling is employed to identify latent topics within a collection of documents. In this analysis, both Latent Dirichlet Allocation (LDA) and BERTopic are applied. The results are subsequently compared with those obtained from previous approaches, aiming to assess whether topic modeling can recover the subjects of the template questions and identify latent topics shared across interviews, or subtopics not detectable by focusing solely on the questions.

4.4.1 Latent Dirichlet allocation (LDA)

LDA is an unsupervised probabilistic model that generates topics by assuming that documents are produced by sampling from a set of topics. Each topic is associated with a set of words, each assigned a certain probability. LDA attempts to reconstruct the document-generation process and estimates the probability that a document belongs to a specific topic. The model operates on a term frequency matrix that disregards word order within the corpus. The initial step involves identifying collections of terms related to a given topic, under the assumption that similar words tend to co-occur within the same topics. Subsequently, the model estimates the probability distribution of topics within each document, typically using Gibbs sampling.

Unlike previous approaches, both this method and BERTopic focus on responses rather than questions. Applying the same process to questions alone may not yield meaningful results for comparison. Moreover, responses rather than question-answer pairs were preferred, as information from questions could have been lost due to the high number of terms in the responses. The goal is to determine whether topic modeling can retrieve topics closely related to the subjects of the questions.

LDA was applied at both the sentence and response levels to assess which level of analysis yields more meaningful, comparable results with respect to the classified question responses used in previous approaches. The suitability of each level depends on factors such as sentence and response length, as well as the homogeneity of terms and topics. The final decision was based on a qualitative evaluation of topic coherence and consistency across sentences and responses.

4.4.2 BERTopic

Another topic modeling method considered in this study is BERTopic. BERTopic follows a multi-step pipeline. First, the text is transformed into embeddings, which are generated by default using sentence-transformers models commonly employed to capture semantic similarity between documents. Second, dimensionality reduction is applied to facilitate subsequent clustering. BERTopic uses Uniform Manifold Approximation and Projection (UMAP) by default, but it also allows alternative techniques such as Principal Component Analysis (PCA) and Truncated Singular Value Decomposition (TSVD). Third, the embedded documents are clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Owing to its modular design, BERTopic also supports other clustering algorithms, including K-Means and Agglomerative Clustering. Finally, BERTopic computes a class-based variant of TF-IDF (c-TF-IDF) to identify the most representative terms for each topic.

The pre-trained embedding model used in this analysis is paraphrase-multilingual-MiniLM-L12-v2, a widely adopted multilingual SentenceTransformers model. For BERTopic's modular components, the default methods described above were used.

As for LDA, BERTopic was performed at both the sentence and response levels.

4.5 Keyword extraction

Once the questions are mapped using the BoW and Embedding approaches, it is possible to compare responses to the same questions and analyse whether specific topics are associated with responses from particular respondents. In this analysis, a qualitative approach is used to compare the results. In particular, keywords are extracted using four different keyword extraction methods. The most relevant keywords for a selected question (for the BoW and Embedding approaches) or for a given topic (for the topic modelling approaches) are compared in the Results section.

The following keyword extraction methods are used: TF-IDF, RAKE (Rose et al. 2010), YAKE (Campos et al. 2018), and KeyBERT (Grootendorst 2020).

RAKE (Rapid Automatic Keyword Extraction) is an unsupervised keyword extraction technique that employs three measures: the number of co-occurrences (degree), the term frequency of candidate keywords (non-stop words or word sequences), and the ratio of these

two measures. Keywords extracted by RAKE are usually sequences of words, since co-occurring words are combined. The number of selected keywords is computed as one-third of the unique words in the co-occurrence graph. Because RAKE relies on co-occurrences, it is domain-independent and language-independent.

YAKE (Yet Another Keyword Extractor) is another unsupervised method that uses features such as term frequency, acronym frequency, uppercase term frequency, and term position. Like RAKE, YAKE is domain- and language-independent. For both RAKE and YAKE, no pre-processing was applied, as they operate directly on the raw corpus.

KeyBERT is a keyword extraction method based on BERT pre-trained embeddings. Its main goal is not merely to select words with high frequency relative to other documents, but to extract keywords that best represent each document. KeyBERT involves three main steps: first, the document is embedded using BERT; second, N-grams are extracted from the document; and finally, the cosine similarity between the N-gram embeddings and the document embedding is measured to identify the most representative words or phrases. KeyBERT can also be applied with different pre-trained embeddings. Note that the results of this method depend on sentence length and the choice of embeddings.

The different outputs of the keyword extraction methods are reported in the Result section.

5 Analysis

Due to the issues discussed in the data section, the analysis section applies the methodology to our data while addressing the issues previously outlined.

5.1 Bag-of-words model

The first step involves splitting the questions into sentences. In our data, this process involved 332 questions (including the template questions), which were split into 490 sentences. This was done by transitioning to a long-form dataframe to account for the different lengths of the 12 interviews and by dividing each question into its constituent sentences. Second, standard pre-processing procedures were applied, including stop-word removal, lemmatization, part-of-speech (POS) tagging, and other common text-processing steps. It is worth noting that predefined stop-word lists included words that, in this analysis, carry relevant information. For instance, the words “where” and “who” were removed from the stop-word list. Therefore, the stop-word list was manually revised by removing some terms and adding others.

The TF-IDF matrix was computed by treating each of the 490 sentences as an independent observation, yielding 326 features. These 326 unique terms, along with two metadata variables–the number of words in each sentence and the presence of a question mark–were used for clustering.

In this analysis, Truncated Singular Value Decomposition (TSVD) was used for dimensionality reduction before performing clustering. The number of components was set to 50, explaining approximately 77.9% of the cumulative variance. Figure 4 shows that the increase in cumulative explained variance gradually slows as more components are added. The choice of 50 components represented a trade-off between the elbow of the curve, located

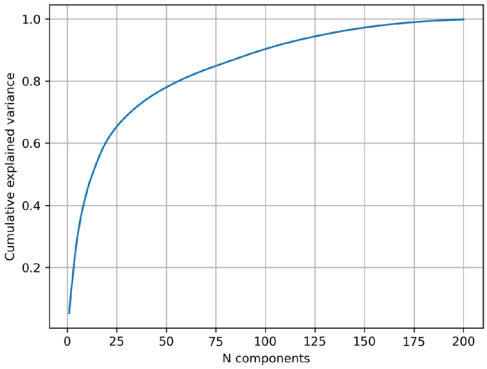

Fig. 4 Cumulative explained variance using Truncated SVD

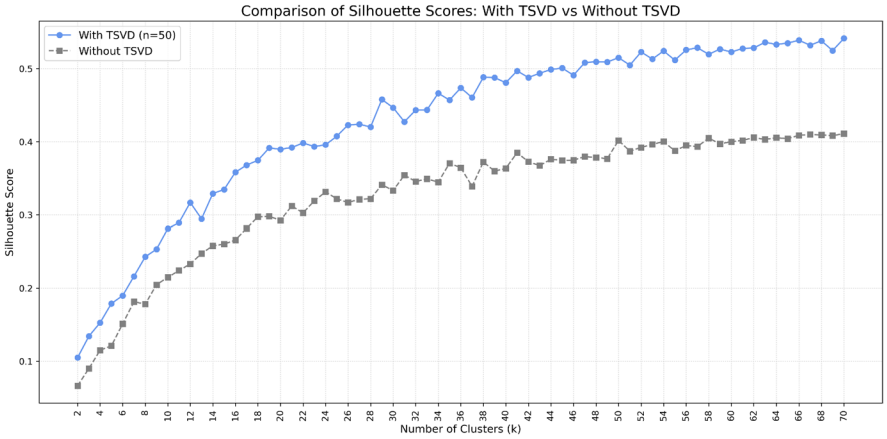

Fig. 5 Comparison of Silhouette scores: with TSVD versus without TSVD (BoW model)

between 25 and 50 components, and the potential loss of information associated with retaining fewer components and therefore a lower cumulative explained variance. Retaining approximately 80% of the cumulative explained variance while reducing the dimensionality to only 50 components was therefore considered a reasonable compromise.

Fourth, the questions were clustered. As shown in Fig.5 , the silhouette score increases with the number of clusters. This behaviour may be explained by the presence of several closely related questions that differ only in minor details. In other words, the model continues to identify better splits. This is unusual, since low values of k (especially $k = 2$) tend to have higher scores. It is evident that the silhouette score begins to level off around $k = 50$.

It should be noted that 50 was also the number of components used in the TSVD version. However, clustering without TSVD achieved similar silhouette scores across the different values of k. Regarding the qualitative assessment of clustering, at around 16 clusters, almost all template questions were successfully identified. As mentioned, in addition to the 14 template questions, the clustering identified one cluster for each of the four spending areas and one for "all categories," for a total of 19. However, after $k = 16$, the model starts splitting clusters of sentences into those referring to the same template question group and those that are non-relevant.

Table 1 provides some examples of the three types of clustering outcomes identified through the clustering process. However, before analyzing the clustering results, it is necessary to recall the structure of the interviews. Specifically, the interview protocol includes six general questions, six specific questions, and a final set of questions related to suggestions. As previously mentioned, the specific questions were repeated for each of the four spending areas. Therefore, if repetition were not expected, one would expect approximately 14 questions. However, since each specific question had to be answered for all spending areas, the total number of questions increased to approximately 30.

The clustering results reflect this structure. The identified groups can be divided into three categories: non-relevant sentences (residual cluster), general questions, and specific questions.

The residual cluster, containing 201 sentences, primarily consists of non-question statements and follow-up prompts. For example, it contains sentences (translated from Italian into English for clarity) such as "By seniors, we mean those over 65" or "Could you explain in more detail how this system works?". In addition, some sentences were used to introduce the structure of the specific questions or to provide other meta-information. For example, the specific section begins with the statement: "So now, for four categories–nutrition and supplements, personal care and health, technology, and leisure–I will ask some questions, always from the perspective of senior citizens." Other sentences are responses to the interviewer or expressions typical of oral interactions. Such sentences often constitute noise, which is why analysis was performed at the sentence level. The presence of repeated sentences resulting from a split question is very limited.

The remaining two types of clusters correspond to general questions and questions related to specific spending areas. Since general questions are not repeated across spending areas, they are easier to identify through clustering. In contrast, the specific questions exhibit greater heterogeneity in their formulation.

This heterogeneity arises because the interviewer did not repeat the full wording of the same question across spending areas. The structure of the specific section had already been explained to the participants.

In practice, participants were asked specific questions in two distinct ways. The first phrasing explicitly mentioned both the main subject of the question and the spending area (e.g., "Which factors most influence the purchase of personal care and health products?"). The second phrasing referred only to the spending area, without restating the main subject (e.g., "And what about technological products?").

Finally, another group of questions consists of confirmation prompts, such as "All categories?", which were used to verify whether a given answer applied to all spending areas.

Table 1 Examples of question cluster types

Category	Example
Residual cluster	By seniors, we mean those over sixty-five
General questions	What is the hierarchy of needs for goods and services among senior citizens?
Spending areas questions	Where do senior citizens obtain information for purchasing basic food products and supplements?
Spending areas questions	When it comes to leisure, which ones do people buy the most?
Spending areas questions	Technological products?
Spending areas questions	All categories?

Table 2 Cross-validation accuracy (BoW approach with and without the clustering feature)

Model	Accuracy (no clustering)	Accuracy (with clustering)
Random Forest	0.816	0.864
Decision Tree	0.696	0.792
Logistic Regression	0.728	0.824
Linear SVC	0.856	0.896
XGBoost	0.792	0.816

Table 3 Accuracy and F1 scores for the LinearSVC model with different Train–Test splits and stratification settings (BoW approach without the clustering feature)

	80/20 split		70/30 split	
Stratification	Non stratified	Stratified	Non stratified	Stratified
Accuracy	0.84	0.80	0.74	0.84
F1-score (weighted)	0.89	0.79	0.77	0.84
F1-score (macro)	0.82	0.80	0.71	0.85

Table 4 Accuracy and F1 scores for the linearSVC model with different Train–Test splits and tratification settings (BoW approach with the clustering feature)

	80/20 split		70/30 split	
Stratification	Non stratified	Stratified	Non stratified	Stratified
Accuracy	0.96	0.88	0.76	0.95
F1-score (weighted)	0.95	0.85	0.78	0.94
F1-score (macro)	0.97	0.83	0.68	0.95

Moving to the supervised approach step, the number of clusters was kept at 16, as in the previous section. As mentioned in the Methodology section, two models are trained. The first includes the clustering results as a feature, while the second does not. Twenty labels were used for training. This number was determined based on the clustering results and the structure of the interview template. Each label was associated with a specific question.

The label set comprises six general questions, seven specific questions, one question related to suggestions and comments, four labels referring exclusively to spending areas, one label indicating that a response applies to all four spending areas, and a final label (label 0) assigned to non-relevant sentences or questions.

The labels were treated as nominal rather than ordinal. This is because, even though the labels followed the order of the questions in the interviews, they also include questions specific to each category, such as “Technological products?”. Furthermore, treating the labels as an ordinal variable with many categories would have introduced several problems.

The models evaluated during cross-validation and their corresponding accuracy results are reported in Table 2.

The best-performing model was LinearSVC based on cross-validation accuracy. However, when the accuracy and F1 scores (weighted and macro) are examined, as shown in Tables 3 and 4, it is evident that the results are unstable. This variability in accuracy and F1 scores is likely due to the high number of labels and the presence of several rare classes, coupled with the limited dataset. This instability is more relevant for the model that includes the clustering feature.

Regarding the two models, the model with the clustering feature is more accurate, likely due to data leakage. However, it does not decrease substantially when the clustering features are removed.

The model that shows the largest decrease is the Decision Tree model, which is highly dependent on the number of observations and does not involve any bagging (as in Random Forest) or boosting (as in XGBoost). Another model whose performance is affected by the use of clustering features is Logistic Regression.

This suggests that the information captured through clustering is also captured by the supervised model.

The correlation matrices for both models trained with and without the clustering features are reported in the Appendix. They both show low inter-feature correlation.

The correlation matrix without the clustering features was expected to show no strong correlations among the features since this model only uses TSVD components and metadata. However, the clustering feature may have been correlated with the TSVD components.

Another issue to consider is the presence of very similar (sometimes almost identical) sentences in the dataset, which may appear in both the training and test sets. This characteristic of the data may inflate the accuracy and F1 scores.

In other words, the model may be recognizing existing examples from the training set rather than generalizing features. For this reason, it should be noted that the semi-structured interview context is typically characterized by similar, if not almost identical, sentences.

The LinearSVC model was selected to predict the labels of previously unlabeled sentences. Subsequently, the 490 sentences were aggregated back into the original set of 332 questions.

Because the validation labels follow the order of the interview template, the expected range is 1–13. As shown in Table 5, the predicted labels generally preserve the original order of the interview questions. In this interview, label 0 was correctly assigned to non-relevant or follow-up sentences. These zero-labeled sentences were subsequently merged with the preceding non-zero labeled question.

The same merging strategy was applied to labels 14 (Food and supplements), 15 (Personal care and health), 16 (Technology), and 17 (Leisure), which refer to specific spending areas associated with a preceding question, as well as to label 19 (All categories), which was used to confirm whether a response applied to all four spending areas. Finally, label 18 (Seasonality) corresponds to a question that was not included in the original interview template but emerged during the interviews.

As illustrated in Table 5, Question 7 (the first question in the specific section) includes both formulations that explicitly restate the main subject–namely, the factors that most influence product purchasing decisions–and formulations that mention only the spending area. This latter formulation occurs particularly frequently in technology-related questions.

5.2 Embeddings

The first steps of the embedding-based approach are similar to the BoW approach. The number of sentences after splitting the questions is the same. Minor changes were made in the pre-processing; in particular, a less aggressive text-cleaning approach was preferred.

Table 5 Example of predicted labels (Interview 2) (BoW model, both with and without the clustering feature)

Turn	Questions	Predicted label
1	By seniors we mean people over sixty-five. I'd like to ask you to introduce yourself; tell us what role you have in the third sector, particularly in relation to seniors	1
2	So, could you describe your experience with providing goods or services to senior citizens?	2
3	Is it managed by you or by the region?	0
4	And do you find this service to be an interesting idea?	0
5	As far as you know, is this something that exists only in Tuscany?	0
6	So, what do they need the most? According to your experience, what is the hierarchy of needs for goods and services among senior citizens?	3
7	So, based on your experience–although you have already answered part of this, but feel free to add anything–what are the biggest difficulties senior citizens face when accessing goods and services intended for them?	4
8	According to your experience, what have been the main changes in recent years regarding the provision of goods and services to senior citizens?	5
9	So, what do you think will be the main future changes?	6
10	Now we'll move to the specific part; it's very brief, don't imagine anything long. I'll ask you various questions about four categories: food and supplements, personal care and health, technology, and leisure. So, in your experience, which factors most influence the purchase of food products and supplements? So which factors most influence the purchase of food products and supplements?	7
11	So which factors most influence the purchase of personal care and health products?	7
12	Technology products?	16
13	And what about products and services for leisure–what factors influence their purchase?	7

Table 6 Template-to-interview question turn alignment across 12 interviews (embedding approach)

Template questions	Int.1	Int.2	Int.3	Int.4	Int.5	Int.6	Int.7	Int.8	Int.9	Int.10	Int.11	Int.12
1	1	1	1	4	1	1	1	1	1	1	3	1
2	2	2	2	2	6	2	2	4	2	2	3	2
3	2	6	3	4	7	3	2	4	2	3	10	2
4	2	7	4	4	8	4	5	5	3	4	10	2
5	7	8	5	7	9	5	6	6	4	5	11	2
6	7	9	6	9	10	6	7	7	4	6	12	2
7	9	10	8	19	11	7	8	8	5	7	15	2
8	9	14	17	19	11	11	12	12	9	9	17	2
9	10	18	17	14	13	15	17	16	10	10	22	2
10	8	20	17	15	14	16	17	24	5	18	34	2
11	12	23	17	19	6	16	17	24	2	2	22	2
12	13	20	21	21	15	21	21	27	2	18	34	2
13	9	14	17	19	11	7	8	24	9	9	15	2
14	15	27	28	23	17	23	23	35	14	27	5	2

Once the embeddings were generated, each question in a given interview was matched to each template question. Table 6 presents the order in which the interview questions were matched to the 14 template questions.

As mentioned, the template questions follow their original order in the template. However, the interview questions also include follow-up questions, which can lengthen the

Table 7 Cross-validation accuracy (Embeddings approach)

Model	Accuracy
RandomForest	0.883
DecisionTree	0.836
LogisticRegression	0.905
LinearSVC	0.907
XGBoost	0.875

Table 8 Accuracy and F1 scores for the LinearSVC using stratified Train–Test splits (embedding approach)

Level	80/20 split	70/30 split
Accuracy	0.91	0.90
F1-score (weighted)	0.89	0.89
F1-score (macro)	0.91	0.90

interviews. Table 6 reports the order of the cosine similarity matches between the template questions and the interview sentences. To ensure accurate matching, a qualitative approach would be preferable. However, the purpose of Table 6 is to investigate whether the sequence of numbers associated with the questions is preserved, as this suggests that the main questions have been correctly identified. In other words, the template questions are asked in the template order, while additional follow-up questions are asked, increasing the number of turns within an interview.

Table 6 shows that some interviews closely follow the template questions, while others include more follow-up questions. For example, in Interview 3, the respondent appears not to have received any follow-up questions in the first part of the interview. Conversely, other interviews show less consistent alignment with the template questions, such as Interview 11. In Interview 12, only three exchanges were recorded due to the respondent's limited participation. In addition, some matches may not be accurate, since an interview sentence (which relates to the subject of template question 3) may contain noise and may be matched to a different template question (question 4, for example).

The best-performing model was LinearSVC. Cross-validation results are reported in Table 7, while Table 8 shows accuracy and F1 scores for LinearSVC with stratified 80/20 and 70/30 training-test splits.

Unlike the previous model, the training process does not appear to be affected by the high number of classes and the relatively small number of observations.

The accuracy and F1 scores are high, as the model was trained on 2617 sentences identified by matching each template question to each sentence and removing matches with a cosine similarity below 0.4. The choice of a 0.4 threshold was motivated by a trade-off between retaining sentences that would otherwise have been omitted with a higher threshold and maintaining an adequate dataset size. In other words, the lower the threshold, the fewer the sentences that were excluded and the larger the number of examples to be classified. Furthermore, lower thresholds yielded sentences that were less similar to the template questions.

As shown in Tables 7 and 8, the model achieves high accuracy and F1 scores, although these evaluation metrics may be inflated due to the use of duplicates (similar sentences among different interviews) and the training process. Indeed, since the training process pairs each sentence with all template questions, this introduces a risk of information leakage

into the evaluation process, leading to overly optimistic scores. For this reason, the manual evaluation was conducted by limiting duplicate entries. However, removing highly similar duplicate entries from the dataset may excessively reduce its size. In other words, the model focuses on identifying patterns of similar sentences rather than generalizing.

After predicting the unlabeled sentences, the cosine similarity score is used to retain the best match between interview sentences and template questions. Finally, the predicted labels can be used to retrieve responses to specific questions. If Question 9 (budget allocation) is of interest, the responses classified as 9 are retrieved.

5.3 Latent Dirichlet allocation (LDA)

LDA was applied at both the response and sentence levels. The response-level analysis benefits from greater semantic similarity across entire answers, while the sentence-level analysis provides a larger number of observations. The dataset comprises 210 responses and 742 sentences. Modeling was performed using single words as well as bigrams and trigrams.

The coherence scores for the four models are shown in Fig. 6. Among the different models, the 11-topic sentence-level model achieved the highest average coherence across the four approaches. Sentence-level models generally exhibited higher coherence scores. The two highest coherence scores for the bigram/trigram sentence-level model are 0.541 (6 topics) and 0.540 (11 topics). For the unigram sentence-level model, the highest value is 0.523 (8 topics). For unigram response-level models, the top coherence scores are 0.444 (2 topics), 0.408 (14 topics), and 0.400 (20 topics). In the bigram/trigram response-level models, high coherence values were observed for 10 and 20 topics (0.419 and 0.409, respectively).

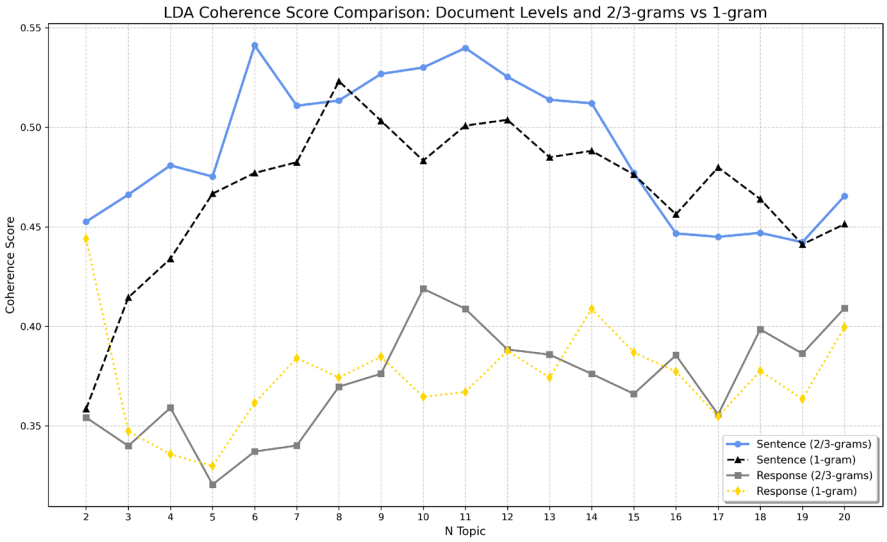

Fig. 6 LDA Coherence Scores

5.4 BERTopic

As with LDA, BERTopic was applied at both the sentence and response levels. As with LDA, the response-level analysis yields fewer observations (210 responses) than the sentence-level analysis (742 sentences). In both cases, the outlier topic (-1) was handled using the reduce outliers function, which reassigns outlier documents to the most appropriate topic.

5.5 Keyword extraction

Regarding keyword extraction, the different methods were reviewed manually using a qualitative approach. The Results section shows the most relevant keywords extracted using the aforementioned methods. As explained below, Question 9 (budget allocation across four spending areas) is considered a case study. Among the information extraction methods, YAKE performs best, while RAKE performs slightly worse. TF-IDF performs poorly because it relies solely on word frequency, while KeyBERT (using paraphrase-multilingual-MiniLM-L12-v2) has limitations due to its reliance on a pre-trained model and language-specific patterns. The results using TF-IDF, RAKE and KeyBERT are available in the Appendix (Tables 19, 20 and 21).

6 Results

The results analysed in this paragraph focus exclusively on Question 9 (budget allocated to four spending areas), which was chosen as an example. The budget represents a key issue in the SENIOR project, which aims specifically at studying the spending behaviours of elderly people. Furthermore, Question 9 is part of the specific section of the interview template, in which each question is repeated for the four spending areas. As shown in the clustering results, the specific section may be more problematic in the classification process. Tables 9 and 10 show the top 3 most relevant key phrases extracted from the responses to Question 9 of the two supervised models. As shown, YAKE correctly focuses on numbers and captures budget percentages. The results for each interview appear to emphasize a specific spending area, likely because in some interviews Question 9 was asked for different areas,

Table 9 Top 3 relevant key phrases using YAKE (BoW model)

Responses	Top 3 relevant key phrases
1	Crtainly technological ten, forty forty anyway, thirty thirty forty
2	Oh God, population, I would throw out numbers
3	Food products, supplements, I think forty percent, food supplements
4	Important share, twenty–thirty percent, I would say food products
5	Food and supplements, personal care, food products
6	Technology fortunately, fortunately as we allocate, price reduced
7	We easily go, easily to seventy, food
8	Food products, let's say sixty, need to eat
9	Budget is managed, the budget comes, it is managed
10	Ten percent technology, twenty percent food, personal care
11	A lot of budget is given, it is given, it is given

Table 10 Top 3 relevant key phrases using YAKE (Embeddings model)

Responses	Top 3 relevant key phrases
1	Stock to keep, keep, third place
2	Definitely technological ten, forty forty roughly, thirty thirty forty
3	Food products supplements, I think forty percent, food supplements
4	Important share, twenty–thirty percent, I would say food products
5	Food and supplements, personal care, food products
6	Technology fortunately, fortunately as we allocate, reduced the price
7	Food products, let's say sixty, need to eat
8	Budget is managed, budget comes, is managed
9	Ten percent technology, twenty percent food, personal care
10	A lot of budget is given, it is given, it is

while in others a single response covered multiple categories. The two models obtained similar results. Starting with the BoW model, 11 responses were assigned label 9. Since one interviewee did not respond to the specific part, only one interview is missing. Moving to the embedding-based model, 10 responses were correctly classified.

The results show clear evidence of budget percentages. Furthermore, the percentages do not seem to be limited to specific spending areas but instead vary across categories.

The food category percentage shows high variability, with values between 20 and 70 percent. Indeed, Response 3 of the BoW model mentions "forty percent" in addition to the keywords "food products," "supplements," and "food supplements." Response 8 highlights the "need to eat" and estimates 60 percent of the budget for food products. A smaller percentage is allocated according to Response 4, in which "20–30 percent" is mentioned. However, Response 4 also mentions that the food category represents an "important share".

Moving to the technology category, the budget is estimated to be around 10 percent. This estimate is supported by the first ("certainly technological ten") and tenth ("ten percent technology") responses of the BoW model. Interestingly, alongside the "certainly technological ten" keywords, the other keywords seem to indicate the percentages of the other spending areas: "forty forty anyway", "thirty thirty forty".

The results also show responses that express uncertainty regarding such a question. For example, the keywords "oh God", "population", and "I would throw out numbers" in Response 2 of the BoW model suggest a general reference to the budget question; however, the keyword extractor does not seem to retrieve any relevant information regarding budget percentages.

Regarding the results of the LDA model, Tables 11 and 12 show the top 5 most relevant terms for the sentence-level and response-level 2- and 3-gram models. The model identifies several interesting topics, although some do not correspond to a template question. For example, a topic coincides with a template question regarding the difficulties faced by seniors. Indeed, both tables mention terms such as "problem" and "difficulties". On the other hand, other topics include terms such as "pathology", "assistance", and "visit". Another topic concerns visits from family members. However, there is no specific question about the relationship between seniors and their relatives. Regarding the four spending areas, they are not easily identifiable. For example, the health category is spread across different topics, while other categories do not seem to belong to a single group, and some may not be present in any topic.

Table 11 Top 5 relevant terms using LDA (2/3-gram sentence level and N topics = 11)

N topics	Counts	Top 5 relevant terms
1	60	To depend, pathology, question, result, point
2	59	Part, healthcare, to say, subject, to manage
3	35	Category, technology, year, to search, way
4	62	Doctor, pharmacy, drug, technological, information
5	28	To exist, initiative, kind, visit, to make
6	45	Model, to see, family member, patient, large
7	43	Society, cooperative, university, municipality, large
8	42	Problematic issue, difficulty, widespread, supplement,..
9	218	Elderly, service, person, type, need
10	40	Small, topic, other, to take, specific
11	110	Cooperative, social, percent, to succeed, assistance

Table 12 Top 5 relevant terms using LDA (2/3-gram response level and N topics = 10)

N topics	Counts	Top 5 relevant terms
1	56	Pharmacy, home, elderly, information, service
2	16	Healthcare, economic, large, public, problem
3	14	Cooperative, facility, pathology, model, social
4	32	Percent, free time, elderly, fund, technology
5	26	Elderly, need, issue, understand, visit
6	26	Service, difficulty, elderly, country, eating
7	10	Integration, occupy, product, senior, means
8	22	Drug, university, society, winter, initiative
9	42	Service, doctor, social, patient, cooperative
10	13	Assistance, population, service, category, healthcare

Regarding the budget question, the term "percent" appears in Topic 11 of the sentence-level model and in Topic 4 of the response-level model. On the sentence level, the topic containing the "percent" term includes 110 observations. This topic is highly heterogeneous in terms of subjects, and it does not identify a subtopic focused solely on budget allocation.

On the other hand, at the response level, the topic includes 33 responses, and the term "percent" has the highest average contribution score in that topic, which represents the average topic probability across documents. Indeed, of the 33 documents, 8 explicitly mention percentages. Some other responses mention percentages or numbers not related to the budget, as well as responses to other questions, such as the hierarchical priority of needs. In other words, the topic captures the importance of the term "percentage" even when "budget" is absent. However, there is no dedicated topic for it, and some responses do not include relevant LDA keywords at the response level. For 17 of the 33 responses in Topic 4, the results are shown in Table 13. As mentioned later, some responses not related to the budget still involve numbers and the concept of purchase.

Moving to BERTopic, Tables 14 and 15 show better results compared to LDA. The four main areas are better identified, and both latent topics and spending areas are captured. For example, a latent topic concerns loneliness and relationships with relatives or children. Regarding the spending areas, they appear to reflect information from the template questions, such as the place where a product is purchased (pharmacy, small market, supermarket) or the type of activity involved (e.g., yoga, gym, and laboratory). There is also a topic

Table 13 Top 3 relevant key phrases using YAKE (LDA - Topic 4 - response level)

Responses	Top 3 relevant key phrases
0	Stock to keep, keep, third place
1	Because I come from, a lot of experience, another user area
2	Food supplements, I think forty percent, food supplements
3	Example, free, all in all
4	Ten percent technology, twenty percent nutrition, personal care
5	I have to divide them, divide them, I have to
6	I would say participation, participation in clubs, clubs and purchase
7	Assistance-related, elderly people, self-sufficient
8	Personal care treatment, physical appearance, inner appearance
9	Elderly people buy more, the question depends, difficult to answer
10	Food products, let's say sixty percent, need to eat
11	Ministerial Decree seventy-seven, Ministerial Decree, Ministerial seventy-seven
12	Hierarchical priority, priority scale, hierarchical
13	Technology, to see
14	Impertinent and very comprehensive, impertinent questions, I would say
15	Mind, comes
16	They buy them, medications that they must, must buy
17	Thirty percent, percent, thirty

Table 14 Top 5 relevant terms using BERTopic (Sentence level)

Topic ID	Counts	Top 5 relevant terms
0	143	Elderly, person, family, son, to depend
1	97	Health, service, hospital, need, assistance
2	76	Cooperative, subject, entity, company, fund
3	61	Technological, to use, television, telephone, minimal
4	38	Beautiful, to tell, perfect, exact, to do it
5	40	Supplement, percent, food product, food product supplement, to believe
6	35	Rest, question, to answer, seasonality, to pose
7	39	Region, territory, system, poor, space
8	34	Pharmacy, drug, pharmacist, pharmacies, to ask
9	28	Free time, cinema, theater, transport, product
10	28	To occupy, director, cooperative, year, healthcare
11	27	Summer, winter, spring, period, to buy
12	32	Technology, third age, style, senior, interest
13	28	Percent, wind, half, to consume, hand
14	22	Doctor, patient, chronic, cultural, to take care
15	14	Shop, supermarket, old, small market, small

related to difficulties (economic and bureaucratic) and seasonality (summer, winter, and spring). Regarding the comparison between the two levels of analysis, the response level provides more general topics and involves longer texts but fewer observations within each topic. However, at the sentence level, the texts are shorter, but each topic has a larger number of observations.

Similarly to LDA, the topics containing the term "percentage" are inspected.

Table 15 Top 5 relevant terms using BERTopic (Response level)

Topic ID	Counts	Top 5 relevant terms
0	43	Service, need, doctor, patient, person
1	28	Technological, use, technological product, online, technology
2	23	Elderly, family member, child, take, person
3	18	Percent, product, food product, part, supplement
4	19	Pharmacy, pharmacist, doctor, depend, part
5	22	Healthcare, social, service, type, deal with
6	16	Product, disorder, gym, nutrition, elderly
7	12	Summer, winter, period, spring, go out
8	14	Drug, leisure time, territory, laboratory, yoga
9	7	Part, category, young, greater attention, pessimistic
10	8	List, economic, difficulty, transport, bureaucratic

Table 16 Top 3 relevant key phrases using YAKE (BERTopic - subtopic of topic 5 - sentence level)

Responses	Top 3 relevant key phrases
1	Fifty to sixty years, supplement category, fifty years
2	Would put last, fourth place, third place
3	Certainly ten technological, forty forty roughly, are used
4	Food supplements products, about forty percent, food supplements
5	2011 there was, there was thirty percent, patients who consume
6	Patients consume eighty, consume eighty percent, percent of resources
7	Thirty-forty percent, personal care, health about thirty-forty
8	Personal care, twenty percent nutrition, ten percent technology
9	Given a lot budget, is given, is given
10	Ninety percent, percent, ninety
11	Personal hygiene products, forty percent, percent including nutrition

At the sentence level, two topics include the term “percentage”, namely Topics 5 and 13. Topic 5 involves 40 observations, and the term “percentage” is associated with the food category. Indeed, the other terms include “supplement”, “food product”, and “food supplement”. However, further analysis reveals that this topic can be divided into two subtopics. The first concerns the food category and includes terms such as “pasta”, “nutrition”, “integration”, and “supermarket”. The second contains the term “percent”, along with other terms such as “free time”, “part”, “patient”, and “technology”. Table 16 shows the terms associated with the subtopic of Topic 5, which involves the term “percent”. The results show that the grouping process appears to focus on numbers, rather than percentages or budgets themselves. Indeed, the topic also includes terms such as "fifty years" and "fourth place", along with budget percentages.

Regarding Topic 13, it includes terms that do not seem to belong to a particular category. Indeed, the term “percent” is not always present in Topic 13. Unlike the term “percent” linked to the food category, in Topic 13 it may either appear in another topic or not appear at all. This lack of stability is a typical issue in methods that aggregate observations.

Moving to the BERTopic results at the response level, 9 of the 18 responses are related to budget allocation. These 9 responses have the highest topic score values and are therefore the most representative documents. However, only 7 responses contain neither numbers nor the term “percentage”. The two remaining responses contain percentage values but are

Table 17 Top 3 relevant key phrases using YAKE (BERTopic - Topic 3 - response level)

Responses	Top 3 relevant key phrases
1	Surely technological ten, forty forty roughly, thirty thirty forty
2	Food products supplements, I think forty percent, food supplements
3	Important share, twenty-thirty percent, I would say food products
4	Ten percent technology, twenty percent nutrition, personal care
5	Given large budget, is given, given
6	Food products, let's say sixty, need to eat
7	I have to split them, split them, I have to
8	I explain what I mean, I mean, I explain
9	Here are forty, will be, twenty
10	We easily reach, around seventy, food
11	Type of nutrition, health type, mainly
12	Hard to answer, answer, difficult
13	Thirty percent, percent, thirty
14	Personal tastes, tastes, personal
15	Person definitely detergents, moisturizing cleansing products, detergents moisturizers
16	Certainly meeting, digital transition, over seventy-five
17	Food products, depends, products
18	Red meat mainly, bread and pasta, vegetables and fruit

Table 18 Top 5 relevant terms identified by BERTopic for Topic 3 (Technology) at the sentence level

Topic ID	Counts	Top 5 relevant terms
-1	7	Part, to use, person, to pay, utility
0	26	Technological, cultural_level, to think, internet, technology
1	19	Television, media, means, information, friend
2	9	Telephone, mobile_phone, smartphone, camera, elderly

not related to the budget question. The most relevant keywords extracted using YAKE are shown in Table 17. As mentioned, BERTopic provides better results than LDA. Indeed, the number of responses on the topic of percentages is smaller and more homogeneous. However, the topic still includes some examples that are not related to the budget. The heterogeneity of the elements within each topic may be due both to the use of key terms such as "budget" or "percentage" in different contexts and to cases where shared terms across topics lead to the merging of different topics. Both cases result in broader and less controlled selections of topic elements.

As shown, the term "percent" is mentioned within the food-category topic. However, it may also be distributed across different spending-area topics. Therefore, the technology topic at the sentence level is analysed using YAKE to extract information related to percentages and budget allocation. For example, there are terms such as "definitely technological ten, forty forty anyway, to be used, thirty thirty forty, technological ten" or "technological is minimal, technological, minimal". However, only 3 of the 61 sentences involve budget allocation. Indeed, as shown in Table 18, no subtopics that exclusively involve budget allocation within the tech category are identified. In other words, this budget-related information is lost within the technology topic.

7 Conclusions

The article provides an exploratory analysis of possible approaches for analyzing semi-structured interviews that present several challenges inherent to their format. Some of these issues include differences in interview length, variations in the number of terms across answers (or questions) within the same interview, and differences in answer length to the same question across participants. Additional challenges involve the dependencies between questions and answers, which may lead to information loss if not properly considered. Furthermore, this case study is constrained by a limited dataset, the use of Italian, and domain-specific terminology.

This analysis attempts to address these issues by mapping interview responses to the template questions originally used by the interviewer to guide the conversation. The results are then compared with those obtained using two standard topic modeling methods, namely LDA and BERTopic.

The results show that the outcomes of the matching process differ from those of the topic modeling methods. Indeed, while topic models identify latent topics, the matching process focuses on identifying the template subjects. In other words, topic modeling is better at uncovering hidden themes, but it may not highlight certain information embedded in the questions. Conversely, the mapping method is designed to retrieve responses to specific questions, though it may fail to capture latent topics.

More specifically, even when topic modeling is able to identify the subject of a question, the resulting topic may also include additional subjects, making it broader than necessary. In other words, the topic may be heterogeneous with respect to the subjects under investigation due to the limited control provided by a model without manual labeling.

This is often due to terms that are unintentionally agglomerated. As shown, budget-related responses also include records containing numerical values. However, some of these numerical values are not related to budget allocation. Such cases may lead to incorrect conclusions, especially because topic modeling does not often take into account the question associated with each response; when it does, the question itself may introduce noise. Another issue concerns the splitting of a specific subject, such as the budget, across different topics. For example, some budget-related responses were split into topics associated with a particular spending area, such as the technology category. In other words, some budget-related responses were embedded in topics related to technology due to the terms used in this category. Furthermore, the results of topic modeling are not stable and may vary across iterations.

On the other hand, the focus on questions in supervised methods allows the content of the responses to be better controlled, resulting in more homogeneous groups of responses.

However, supervised methods require manual labeling, while topic modeling does not. In addition, the problem of information loss is not completely solved, and questions may not be easily identifiable, which makes them difficult to classify.

The effectiveness of supervised approaches and topic modeling is linked to the goal of the analysis. If certain information needs to be extracted, then supervised approaches provide more controlled and homogeneous results. On the other hand, if the goal is to explore interview responses in order to identify which topics appear most frequently, then topic modeling is more appropriate. Furthermore, this article focuses on a limited dataset and

on semi-structured interviews. Therefore, the characteristics of the data must be taken into account and the methodology adjusted accordingly.

This study provides an exploratory analysis, which involves several potential improvements. Indeed, this approach focuses on semi-structured interviews, which typically feature highly similar questions. Furthermore, the management of information from questions not included in the template was not addressed. The limited dataset size represents another constraint, which could be mitigated by expanding it in future studies. Regarding the pre-trained models used for the embeddings, more recent models may be implemented.

8 Appendix

See (Figs.7, 8, 9, 10) and See (Tables 19, 20, 21)

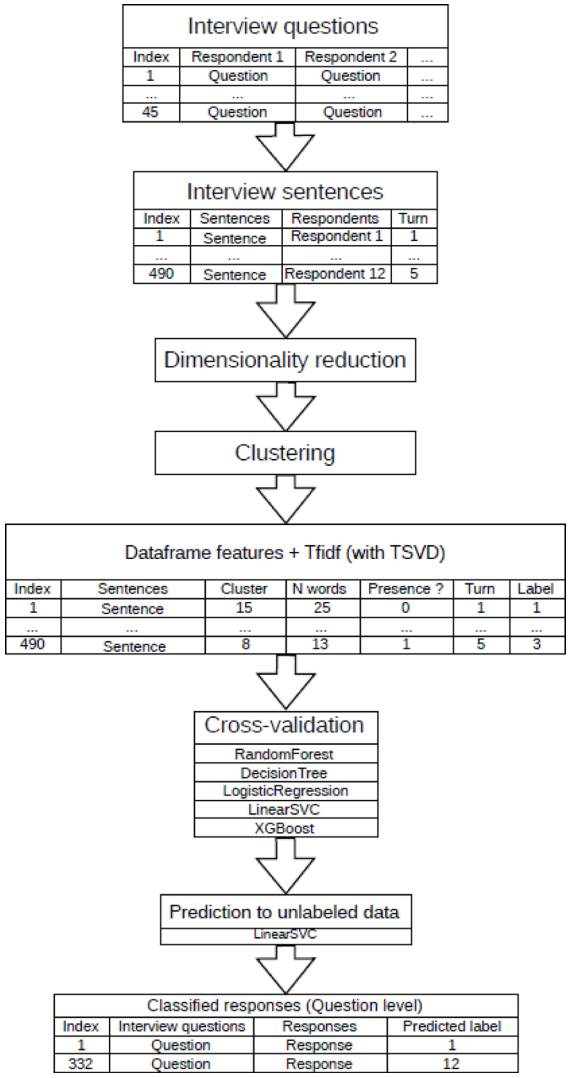

Fig. 7 Flow diagram of the process (BoW approach)

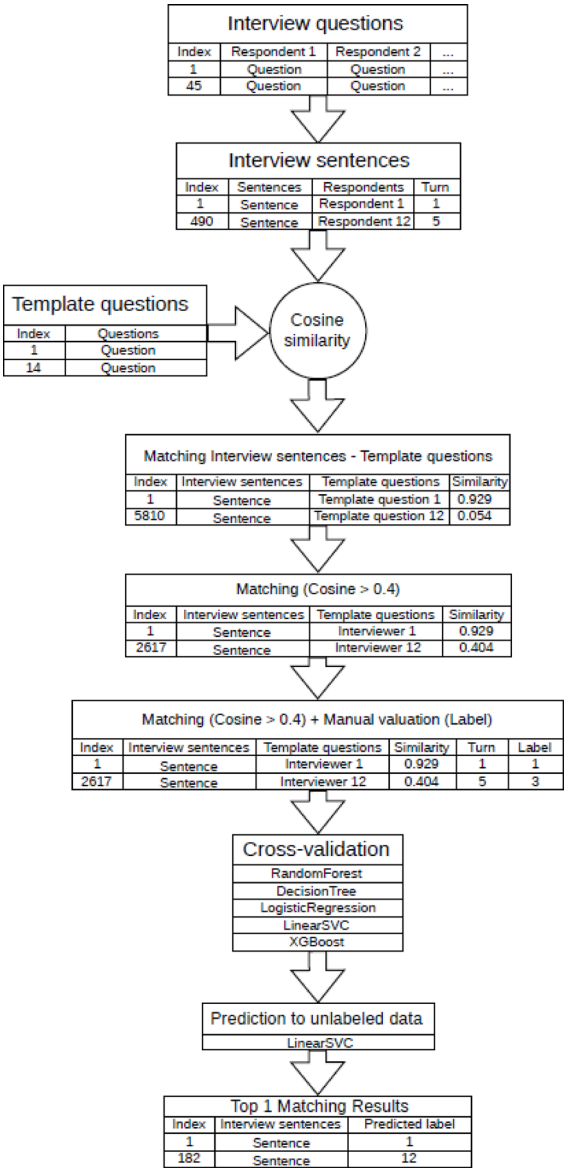

Fig. 8 Flow diagram of the process (Embeddings)

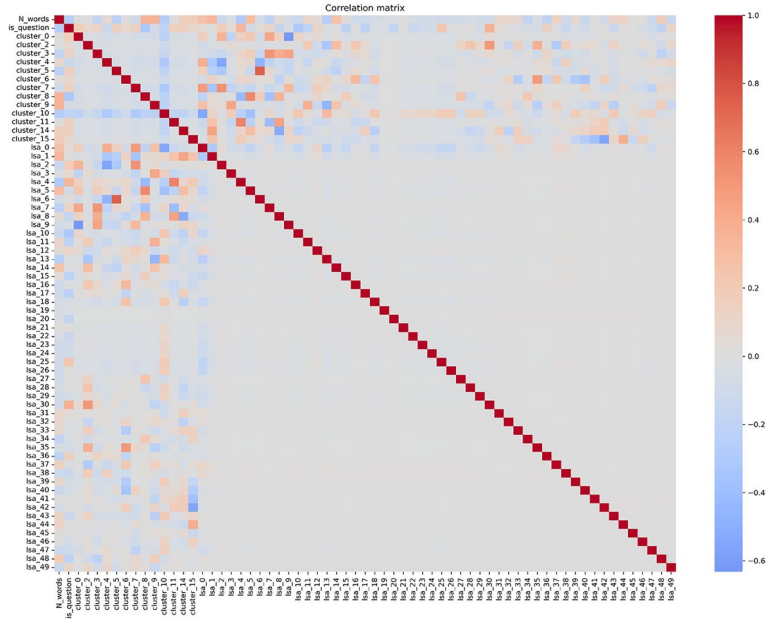

Fig. 9 Correlation matrix with clustering features (BoW approach)

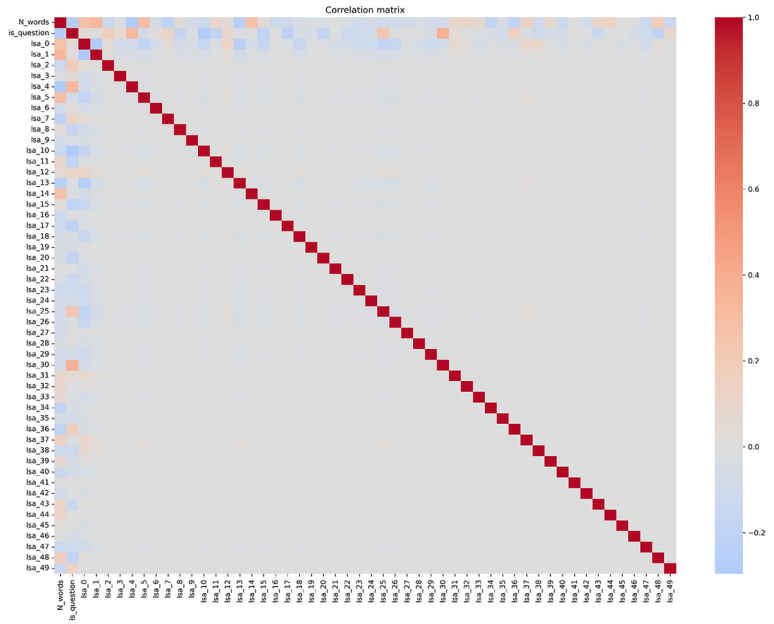

Fig. 10 Correlation matrix without clustering features (BoW approach)

Table 19 Top 5 relevant terms using TF-IDF

Responses	Top 3 relevant key phrases
1	Part, enough, to use, minimum, technological
2	To know, population, really, rice, food
3	Percent, to believe, equal, roughly, to allocate
4	To say, percent, product, remaining, share
5	Priority, to produce, instance, personal, health
6	Prevention, price, inaccessible, to orient, export
7	To go, easily, clothing, to eat, to remain
8	Supplement, food, product, to take, need
9	Product, category, consumption, interest, aging
10	Percent, nutrition, personal, technology, leisure
11	To give, to come, less, home, budget

Table 20 Top 3 relevant key phrases using Rake

Responses	Top 3 relevant key phrases
1	Quite thirty thirty forty forty, anyway quite important, definitely makes one hundred, technological ten, food part in my opinion
2	I would shoot numbers, four hundred euros, four tins
3	Food products supplements, I think forty percent, twenty percent
4	Remaining free time, health about thirty, I would say food products
5	Food products, first instance, then products
6	Prices still very high, free time export, zero cost simply
7	Clothing comes after, rest goes after, something remains
8	Food products yes, supplements however, well, staying there
9	Budget is managed, over sixty-five years, generally pays attention
10	Then twenty percent free time, personal care in my opinion, another twenty percent
11	Comes... eighty food products supplements, given a lot of budget, fifty percent

Table 21 Top 3 relevant key phrases using KeyBERT

Responses	Top 5 relevant terms
1	Buy technology, minimum technology, technology part
2	Rest of the population, population indeed, population awareness
3	Regarding health, percentage product, percentage regarding
4	Food supplement, food product, percentage product
5	Food supplement, food product, personal health
6	Usable healthcare, lesser need, healthcare affected
7	Food share, eating clothing, remaining eating
8	Food supplement, buy supplement, supplement supplement
9	Aging product, successful aging, aging aging
10	Percentage nutrition, nutrition percentage, percentage technology
11	Budget allocation, budget, less at home

Author contributions All authors contributed equally to the elaboration and the interpretation of results.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement. The paper was produced using data collected within the framework of the project funded by the European Union – Next Generation EU – Project "Age-It - Ageing well in an ageing society" (PE0000015), CUP

H73C22000900006, Spoke 6, PNRR – PE8 - Mission 4, C2, Investment 1.3. However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Data availability The datasets generated and/or analyzed during the current study are not publicly available due to ethical, legal, and confidentiality constraints and cannot be shared.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Apports de logiciels d'analyse de données textuelles (ADT) dans les procédures d'analyse de contenu d'entretiens semi-directifs de recherche: Alceste et hyperbase. Bull. de Méthodologie Sociologique. **57**, 25–48 (1997). https://api.semanticscholar.org/CorpusID:62654063

Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: using language models to simulate human samples. Polit. Anal. **31**(3), 337–351 (2023). https://doi.org/10.1017/pan.2023.2

Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Sebastopol (2009)

Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012). https://doi.org/10.1145/2133806.2133826

Bonneau, J., Dister, A.: Logométrie et modélisation des interactions discursives. L'exemple des entretiens semi-directifs, in 10° Journées internationales d'Analyse statistique des Données Textuelles (JADT), S. Bolasco, Ed., Rome, Italy: Università di Roma, Jun (2010). https://univ-cotedazur.hal.science/hal-01362689

Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa

Brinkmann, S., Kvale, S.: Doing Interviews, SAGE Publications (2018)

Brown,T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020). arXiv:https://arxiv.org/abs/2005.14165

Campos, R., Mangaravite, V., Pasquali, A., Mário Jorge, A., Nunes, C., Jatowt, A.: YAKE! collection-independent automatic keyword extractor, In: Advances in Information Retrieval, pp. 806–810, (2018)

Chintagunta, B., Katariya, N., Amatriain, X., Kannan, A.: Medically aware GPT-3 as a data generator for medical dialogue summarization, In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, C. Shivade, R. Gangadharaiah, S. Gella, S. Konam, S. Yuan, Y. Zhang, P. Bhatia, and B. Wallace, Eds., Online: Association for Computational Linguistics, pp. 66–76 (2021). https://doi.org/10.18653/v1/2021.nlpmc-1.9

Curran, J.R.: From distributional to semantic similarity, , Edinburgh (2003)

Dalud-Vincent, M.: Les choix du sociologue avec Alceste - De la forme du corpus aux résultats obtenus. Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique **105**(1), 25–52 (2010). https://doi.org/10.1177/0759106309352584

Danescu-Niculescu-Mizil, C., Lee, L.: Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs, In: Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, F. Keller and D. Reitter, Eds., Portland, Oregon, USA: Association for Computational Linguistics, pp. 76–87 (2011). https://aclanthology.org/W11-0609/

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (NAACL), (2019)

Doyle., G.,. Frank, M.C: Investigating the sources of linguistic alignment in conversation, In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, pp. 526–536 (2016). https://doi.org/10.18653/v1/P16-1050

Dumais, S., Furnas, G., Landauer, T., Deerwester, S., Harshman, R.: Using latent semantic analysis to improve access To textual information, In: Proceedings of CHI, vol. 88 (1996). https://doi.org/10.1145/57167.57214

Goldberg, Y.: Neural network methods for natural language processing. Synth. Lect. Hum. Lang. Technol. **10**, 1–309 (2017). https://doi.org/10.2200/S00762ED1V01Y201703HLT037

Goodfellow, I., Bengio, Y., Courville, A.: MIT Press, Deep Learning (2016)

Grootendorst, M.: KeyBERT: Minimal keyword extraction with BERT, Zenodo. version v0.3.0, 2020

Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure, (2022). arXiv:2203.05794 arXiv preprint

Guetterman, T., Chang, T., DeJonckheere, M., Basu, T., Scruggs-Wodkowski, E., Vydiswaran, V.G.: Augmenting qualitative text analysis with natural language processing: methodological study. J. Med. Internet Res. **20**, e231 (2018). https://doi.org/10.2196/jmir.9702

Hämäläinen, P., Tavast, M., Kunnari, A.: Evaluating large language models in generating synthetic HCI research data: a Case Study, In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), Hamburg, Germany: Association for Computing Machinery 433, 19 (2023). https://doi.org/10.1145/3544548.3580688

Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: Semantic similarity from natural language and ontology analysis, arXiv preprint arXiv:1704.05295, Apr 18 (2017). arXiv:https://arxiv.org/pdf/1704.05295

James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning: with applications in R, Springer (2021)

Jurafsky, D., Martin, J.: Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, (2008). (2nd ed.)

Karimi, A., Ma, A.: NLP's word2vec: negative sampling explained, Mar 18, (2024). https://www.baeldung.com/cs/nlps-word2vec-negative-sampling

Kobayashi, S.: Contextual Augmentation: Data augmentation by words with paradigmatic relations, In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, pp. 452–457 (2018). https://doi.org/10.18653/v1/N18-2072

Kuckartz, U.: Qualitative text analysis: a guide to methods. Softw. Pract. Exp. (2014). https://doi.org/10.4135/9781446288719

Leeson, W., Resnick, A., Alexander, D., Rovers, J.: Natural language processing (NLP) in qualitative public health research: a proof of concept study. Int J Qual Methods **18**, 160940691988702 (2019). https://doi.org/10.1177/1609406919887021

Lison, P., Pilán, I., Sanchez, D., Batet, M., Øvrelid, L.: Anonymisation models for text data: state of the art, challenges and future directions, In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, pp. 4188–4203 (2021). https://doi.org/10.18653/v1/2021.acl-long.323

Merriam, S.B., Tisdell, E.J.: Qualitative Research: A Guide to Design and Implementation, Wiley (2015)

Miles, M., Huberman, M., Saldaña, J.: Qualitative data analysis: a methods sourcebook, (2013)

Nasir, M., Chakravarthula, S., Baucom, B., Atkins, D., Georgiou, P., Narayanan, S.: Modeling interpersonal linguistic coordination in conversations using word Mover's distance, in Interspeech 2019, 2019, pp. 1423–1427, 2019. https://doi.org/10.21437/Interspeech.2019-1900

Parfenova, A.: Automating the Information Extraction from Qualitative Data: a Study of Approaches to Analyze Semi-Structured Interview Transcripts, (2022). (PhD thesis)

Pilehvar, M., Camacho-Collados, J.: Embeddings in Natural Language Processing, Morgan & Claypool Publishers (2021)

Quillivic, R., Payet, C.: Semi-Structured Interview Analysis: A French NLP Approach for Social Sciences, in Mots comptés, textes déchiffrés, Presse Universitaire de Louvain, Bruxelles, Belgium, (2024). https://hal.science/hal-04627806

Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Conference on Empirical Methods in Natural Language Processing (EMNLP), Aug 14, (2019)

Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents, in Text Mining: Applications and Theory, pp. 1–20, (2010)

Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0

Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with Monolingual data, In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, pp. 86–96 (2016). https://doi.org/10.18653/v1/P16-1009

Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc **28**(1), 11–21 (1972). https://doi.org/10.1108/eb026526

Steck, H., C.E.: Is Cosine-Similarity of Embeddings Really About Similarity? In: ACM Web Conference 2024, (2024)

Talmy, S.: Qualitative interviews in applied linguistics: from research instrument to social practice. Annu. Rev. Appl. Linguist. **30**, 128–148 (2010). https://doi.org/10.1017/S0267190510000085

Tatsat, H., Puri, S., Lookabaugh, B.: Machine Learning and Data Science Blueprints for Finance, O'Reilly Media (2020)

Vechtomova, O.: Introduction to information retrieval (review). Comput. Linguist. **35**, 307–309 (2009). https://doi.org/10.1162/coli.2009.35.2.307

Zhou, K., Ellis, K.: Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 401–423 (2022)

Wei., Zou, K.: EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, pp. 6382–6388 (2019). https://doi.org/10.18653/v1/D19-1670

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.