Complex Systems Informatics and Modeling Quarterly (CSIMQ)
eISSN: 2255-9922
Published online by RTU Press, https://csimq-journals.rtu.lv
Article 244, Issue 44, September/October 2025, Pages 52–65
https://doi.org/10.7250/csimq.2025-44.04

Unsupervised Approach for Specialized Vocabulary Creation and Enrichment: A Case Study in the Multidisciplinary Building Sector

Lydia Khelifa Chibout[1] and Manuele Kirsch Pinheiro[2*]

[1] CSTB Scientific and Technical Center for Building, 84 Av Jean Jaurès,
77420 Champs-sur-Marne, France
[2] Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne,
90 rue de Tolbiac, 75013 Paris, France

Lydia.chibout@cstb.fr, Manuele.Kirsch-Pinheiro@univ-paris1.fr

Abstract. The exponential growth of digital information has exposed organizations to unprecedented challenges in managing and structuring their knowledge repositories. In the context of knowledge management, the ability to extract, organize, and use relevant information from large collections of documents has become a critical factor for operational efficiency and informed decision-making. However, identifying necessary knowledge sources and building appropriate knowledge bases represents a significant and time-consuming barrier. In this article, we address these challenges by leveraging advanced Natural Language Processing (NLP) techniques, particularly in combination with Large Language Models (LLMs), to facilitate the selection of more representative keywords for the creation and enrichment of vocabularies for knowledge management purposes. We explore the application of clustering techniques combined with NLP-driven keyword extraction to support the construction of specialized vocabularies that address the multidisciplinary nature of the content at CSTB, a French scientific research center focused on building science. We applied a pipeline with two approaches for keyword extraction: document-based clustering and chunk-based clustering. We provide a detailed overview of the proposed pipeline, present the results of our experiments, and describe the human validation process used to evaluate these results.

Keywords: Keywords Extraction, Clustering, Vocabulary Identification, Knowledge-Based Construction, Knowledge Management.

1 Introduction

Knowledge management becomes a key element for any modern organization. By implementing well-defined systems for storing, retrieving, and indexing documents, organizations can ensure

[*] Corresponding author

Reference: L. K. Chibout and M. K. Pinheiro, “Unsupervised Approach for Specialized Vocabulary Creation and Enrichment: A Case Study in the Multidisciplinary Building Sector,” Complex Systems Informatics and Modeling Quarterly, CSIMQ, no. 44, pp. 52–65, 2025. Available: https://doi.org/10.7250/csimq.2025-44.04

Additional information. Author ORCID iD: L. K. Chibout – https://orcid.org/0000-0002-6448-4580 and M. K. Pinheiro – https://orcid.org/0000-0002-5611-4942. PII S225599222500244X. Received: 29 August 2025. Accepted: 21 October 2025. Available online: 31 October 2025.

that pivotal information is easily available to those who require it, enabling swift and informed decision-making [1], [2]. This streamlined access to information facilitates knowledge sharing and collaboration, fostering a dynamic environment conducive to continuous process improvement [1], [3]. The benefits of organized documentation extend across various sectors, encompassing academia [4], the business world [5], and public administration [6]. In each of these use cases, efficient access to explicit knowledge translates directly into improved performance, enhanced innovation, and increased overall efficiency. The CSTB[†], a French Scientific Center specializing in buildings, is not an exception. The knowledge management strategy at CSTB aims at facilitating the sharing and accessing of knowledge, which is particularly significant given the multidisciplinary nature of its projects, spanning areas such as environmental science, construction materials, and health in buildings.

However, identifying necessary knowledge sources and building appropriate knowledge bases represents a cumbersome and time-consuming task. The capacity to readily identify relevant information allows better decision-making, solving problems more efficiently, and contributes more effectively to organizational goals. A well-structured vocabulary not only facilitates the retrieval of relevant documents but also enhances the overall efficiency of information management and decision-making processes. Keyword extraction is a key step in this process, as it enables the automatic identification of essential terms that encapsulate the core content of the documents. By extracting keywords from multiple documents, it becomes possible to identify terms that are shared across multiple disciplines. These common keywords act as bridges between distinct domains, highlighting areas of conceptual overlap or shared interests. This process fosters interdisciplinary collaboration by providing a clear starting point for joint projects and discussions. For instance, in CSTB, identifying shared terms such as “sustainability” or “data modeling” might connect environmental science and computational research teams, enabling the development of innovative cross-disciplinary solutions. This approach not only streamlines knowledge integration but also encourages synergy between multidisciplinary teams.

These keywords are directly related to the document context and constitute a semantic repository. To achieve this, unsupervised keyword extraction techniques have gained attention, particularly in contexts in which labeled datasets are unavailable or impractical to create. Recent advancements in natural language processing (NLP), including the development of large language models (LLMs) such as BERT [7] and GPT [8], offer promising opportunities for improving the quality and precision of keyword extraction. These models, through their deep contextual understanding, can capture nuanced relationships between words and concepts in text, providing a robust foundation for unsupervised approaches. Studies such as [9] on TextRank and [10] on graph-based methods have established the basis for unsupervised keyword extraction. More recently, BERTopic [11] and other LLM-based techniques have emerged as effective tools for analyzing and summarizing large corpora, making them highly relevant for building semantic repositories in knowledge-intensive environments.

However, as underlined by Churchull and Singh [12] and by Bisht [13], there is no magical approach that would fit every situation. Data set characteristics, such as size and complexity, will interfere with models’ performance, which enhances the necessity of modular solutions that could be easily adapted accordingly. Besides, evaluating obtained results on real-world scenarios can easily become a challenging task, since no “ground truth” data may be available, notably on multidisciplinary domains in which keywords are not necessarily shared or have the same meaning in all the disciplines.

In this work, we propose a real-world case study in a multidisciplinary sector. Through this case study, we aim to leverage unsupervised methods using LLMs to extract keywords from a documentary corpus, contributing to the building of specialized vocabulary to support document retrieval and knowledge sharing. This article proposes a modular pipeline in which two approaches for keyword extraction based on the type of unsupervised clustering have been used. The keyword

[†] Scientific and Technical Center for Building. http://www.cstb.fr/en/

extraction, by using c-TF-IDF[‡], makes finding accurate and relevant information much easier. Clustering documents is essential for researchers engaged in interdisciplinary research across various topics. Our proposed pipeline, including extracting keywords after clustering text documents, significantly enhances the discovery of useful information, addresses issues related to understandability, and improves searchability for users. We apply this pipeline to a set of multidisciplinary CSTB documents, whose identified keywords have been submitted to a panel of domain experts for evaluation purposes.

This article is structured as follows: Section 2 reviews related works on keyword extraction methods. Section 3 details our proposed pipeline, followed by the presentation of the experimentations performed on CSTB (Section 4). Section 5 discusses the results, as well as the evaluation and validation processes. Finally, Section 6 concludes the article.

2 Related Work

The organization of an effective document database offers numerous advantages to organizations, regardless of their sector of activity [4]–[6]. In each of these cases, efficient access to explicit knowledge directly translated into improved performance, increased innovation, and enhanced overall efficiency. To this end, the identification of relevant keywords represents a key step in building these databases.

Various works in the literature address keyword identification and extraction. Examples include traditional methods relying on statistical analyses, such as TF-IDF [14], and linguistic approaches, like lemmatization and nominal phrase extraction [15]. However, these often struggle to capture the semantic richness of complex texts. Unsupervised algorithms, such as TopicRank [16], introduce a graph-based approach to structure keywords around coherent themes. Khelifa et al. [17] propose structuring words into a topic graph while preserving their contextualization within the text through semantic dimensions such as region, time, discipline/field, or language. Abilhoa and De Castro [18] propose a keyword extraction method for tweet collections that represents texts as graphs and applies centrality measures to find relevant vertices (keywords). Hasan et al. [19] propose a system that extracts a specific number of key terms from documents to identify a text's main content. Data is collected from various sources, like books and journals. Various well-known machine learning techniques, such as SVM, logistic regression, or the PAT-tree algorithm, were used for keyword extraction. Bisht [13] evaluated different keyword extraction methods based on spatial distribution and proposed a measure based on frequency, inverse document frequency, variance, and Tsallis entropy; the results highlighted that no single perfect method exists. Ahadh et al. [20] proposed an automated, semi-supervised, domain-independent approach for analyzing accident reports. Given a set of user-defined classification topics and domain literature such as manuals, glossaries, and Wikipedia articles, the method can identify domain-specific keywords and cluster them into topics with minimal expert involvement. These keywords and topics can then be used for various data mining purposes, including classification. However, these methods often require a large number of labeled documents as training examples.

More recent approaches have explored the use of deep learning techniques to improve the accuracy and efficiency of keyword extraction, demonstrating the potential of neural networks to learn complex patterns in texts and identify relevant keywords [21].

Furthermore, document clustering with techniques such as K-means or DBSCAN [22] improves the contextualization of keywords by grouping similar documents, but the evaluation of clustering quality often relies solely on indices like the Silhouette Score [23]. LLMs such as BERT [7] or GPT-4 achieve superior precision and relevance compared to traditional approaches, excelling at modeling complex relationships within text, enabling better keyword extraction [24]. Their use in multi-label classification [25] may also optimize the consistency of results. Zhou et al. [26], for

[‡] https://medium.com/@shashankag14/understanding-tf-idf-and-c-tf-idf-in-topic-modeling-071eb82fa858

instance, have experimented with using ChatGPT for keyword extraction, obtaining a representative and operational set for scientific research.

Through these works, we may observe that combining clustering and LLMs optimizes keyword relevance and reduces noise in the data. This approach is particularly effective for tasks such as intelligent indexing, information retrieval, technological watch, and personalized content recommendation. Also, some supervised approaches can be observed in the literature, where keywords are not always grouped into topics.

Besides, keyword extraction can be related to topic modeling techniques, whose goal is to compress a corpus of thousands of documents into a short summary that captures the most prevalent subjects present in the corpus [12]. For instance, Akarsu and Parmaksiz [27] have applied topic modeling approaches to academic literature about digital leadership, identifying the most frequently used terms related to it. Churchill and Singh [12] propose an extensive survey on this subject and on its evolution across the years. From their analysis, we may observe that there is no miracle approach that overperforms any other on topic modeling, since, according to these authors, not all topic modeling approaches are well-suited for all types of text, their performance varying according to data set scale, document size, and noise. Such variation in algorithms' performance and applicability makes it harder to use them in real-world scenarios, which may contain heterogeneous datasets. This reality reinforces the need for a configurable pipeline in which different algorithms could be used and tested on each step.

Real-world scenarios represent a challenge for any keyword extraction or topic modeling technique, not only because of data set complexity, but also because there is no "ground truth" available, making it harder to evaluate the obtained results by traditional metrics. Unfortunately, one may observe that the validation by human experts is not always emphasized in the literature. However, we consider this expert evaluation essential, particularly in multidisciplinary environments like the CSTB (Scientific and Technical Building Centre).

3 Unsupervised Vocabulary Extraction

In this article, we propose an unsupervised analysis of a large volume of unlabeled documents, combining document clustering and keyword extraction with LLMs. The results were presented to a panel of experts for evaluation and validation. We propose using advanced NLP techniques to uncover meaningful insights from textual data. The process follows a structured pipeline designed to optimize topic modeling and clustering tasks. This proposed pipeline represents an unsupervised and modular handling a large corpus of unlabeled multidisciplinary documents. This work intends to tackle the need for a configurable pipeline mentioned before, as well as the human validation, using a panel of domain experts, who would validate the results of each step.

An overview of the different steps of our pipeline is presented in Figure 1:

1. Extraction and Pre-processing of a document corpus.
2. Text Chunking: input documents are divided into smaller, coherent segments to facilitate accurate representation and analysis.
3. Embedding Generation and Dimensionality reduction: transformer-based models are used to convert text chunks into high-dimensional vectors that encapsulate semantic meaning and to reduce the dimensionality of embeddings while preserving their structure.
4. Aggregation embeddings (mandatory for the clustering documents approach).
5. Clustering: the reduced embeddings are clustered, which efficiently identifies dense regions and separates noise.
6. Keyword Extraction: methods to extract key terms and enhance topic interpretability from clusters by using c-TF-IDF and CountVectorizer.

This pipeline has been experimented with a large base of technical CSTB documents concerning building, civil engineering, health in buildings, acoustics experimentations, and other construction-related domains. Such documents are available to CSTB researchers, who could evaluate the keywords suggested by the proposed approach. It is worth noting that two different approaches for clustering have been considered, a document-based approach and a chunk-based one, both giving different insights about the analyzed documents.

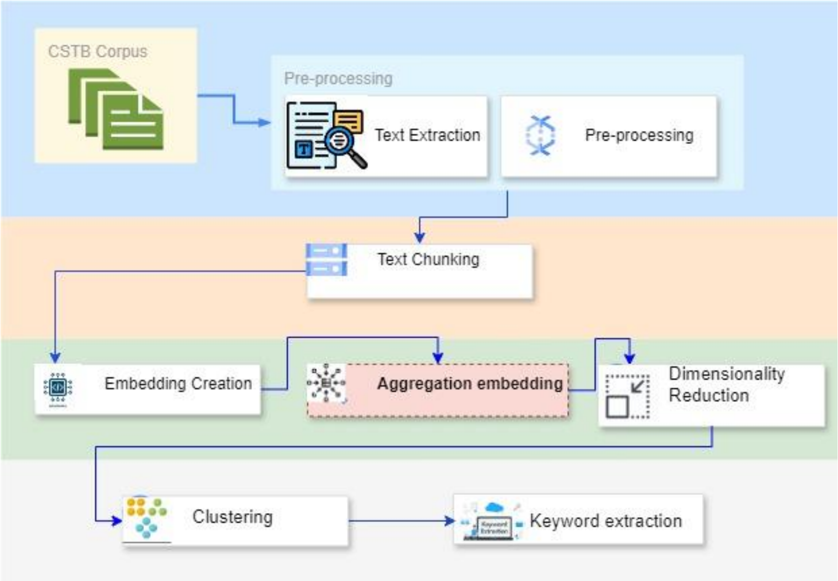


Figure 1. Structured pipeline for topic analysis and clustering

The next sections detail each element of this pipeline before presenting the experiments performed at CSTB (cf. Section 4) and discussing its results (cf. Section 5).

3.1 Text Extraction and Preprocessing Step

The preprocessing step is a necessary step for any data analysis task. Extracting text from PDF documents and preprocessing those are then essential steps in many data processing pipelines. PDFs are widely used for document sharing due to their consistent formatting across devices, but their structure often presents challenges for automated text extraction. This is because PDFs are not inherently designed for text manipulation, and the content can vary widely in complexity, including plain text, tables, images, or even scanned documents containing text as images. Here, the preprocessing of text is carried out in order to ensure the data is clean and ready for further analysis. This preprocessing consists of:

- Text cleaning and normalization: Special characters such as typographic quotes and apostrophes are replaced with their simpler equivalents (e.g., curly quotes replaced with straight quotes). Additionally, newline characters ***\n*** are removed, and excessive spaces are handled by splitting and rejoining the words.
- Removal of numbers: All numeric characters are removed from the text using regular expressions to ensure that the focus remains on the textual content.

Additionally, in order to reduce text noise and to focus the analysis on potentially significant elements, lemmatization and part-of-speech (POS) tagging are applied directly to the keywords, and Bi-gram word extraction. Both steps were requested by the business and expert team who participated in the experimental evaluation discussed in section 5.

- Lemmatization: this process reduces words to their base or root form. For instance, “running” becomes “run,” and “better” becomes “good.” This step is pivotal for ensuring that different forms of a word are treated as the same word during the analysis.
- POS tagging to remove verbs: part-of-speech tagging is used to identify verbs in the extracted keywords, which are then excluded to focus on the nouns and adjectives that are more likely to contribute to topic modeling and key term extraction.
- bi-Gram extraction: for understanding local word dependencies.

3.2 Chunking Process

Text splitting is the process of dividing a large body of text into smaller, more manageable segments, commonly referred to as chunks. The primary objective is to ensure that each chunk remains within the input limits of a model (e.g., token limits in language models) while preserving sufficient context to maintain its meaning.

The method used here is character-based text splitting, in which a long text is divided into chunks based on a specific number of characters. To enhance continuity between chunks, text overlapping has been introduced. This involves including a small portion of the previous chunk at the start of the next one, allowing for better context retention across chunks.

3.3 Embedding and Dimensionality Reduction

Embedding is the process of converting text into numerical vectors that capture its semantic meaning. These vectors form the models on which clustering techniques are applied. Dimensionality reduction is a process used in data analysis and machine learning to reduce the number of features (or dimensions) in a dataset while retaining as much important information as possible [28].

Multiple methods for creating embeddings are proposed in the literature, among them Word2Vec [29], BERT-based sentence embedding [30], and Sentence-BERT [31]. Word2Vec, developed by Google, uses two main approaches: Continuous Bag of Words (CBOW), which predicts a word based on its context, and Skip-gram, which predicts context words from a given word. These models are trained on large text corpora, producing embeddings that capture the relationships between words. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that generates context-sensitive embeddings. For tasks requiring sentence- or document-level embeddings, methods such as Sentence-BERT (SBERT) or Doc2Vec are used. SBERT refines BERT to produce semantically meaningful sentence embeddings that can be compared by cosine similarity. These techniques have various applications in natural language processing, including similarity search, document clustering, and information retrieval.

Considering dimensionality reduction, here again several algorithms are commonly used for this purpose, including UMAP (Uniform Manifold Approximation and Projection), which is particularly well-suited to visualizing high-dimensional data. It preserves the overall structure of the data better than t-SNE, making it effective for both visualization and manifold learning [32]. t-SNE (t-Distributed Stochastic Neighbor Embedding) is another very popular non-linear technique for visualizing complex data. It is based on converting similarities between data into joint probabilities and seeks to minimize the Kullback-Leibler divergence between high- and low-dimensional spaces [33]. Finally, PCA (Principal Component Analysis) is a linear dimensionality reduction method that projects data onto directions that maximize variance. It is widely used to reduce the complexity of datasets while preserving as much variability as possible [34].

The results of these processes are highly dependent on the document's language. Mixing documents in different languages may negatively impact these results. However, documents in different languages are commonly found in nowadays organizations, which should be considered during embedding and dimensionality reduction steps. Thus, during experimentation conducted on CSTB, the used dataset could be split into two subsets, composed of English and French documents, respectively. Each subset employed a distinct model suited to each language.

3.4 Clustering Approaches

During the clustering phase, we evaluated the most suitable type of clustering, considering that the documents are multidisciplinary, and a single article can address cross-cutting issues spanning multiple fields. To identify the optimal approach, we propose two clustering methods: chunk-based and document-based. These two methods were tested to determine the most effective strategy for keyword extraction to be integrated into the vocabulary.

3.4.1 Clustering Document-Based Approach with LLM

Document-based approach, represented in Figure 2, emphasizes clustering entire documents rather than individual chunks. To accomplish this, an aggregation method is employed to merge the embeddings of a document's chunks into a unified representation. This aggregated embedding is then utilized for clustering, enabling the identification of similarities between documents as a whole.

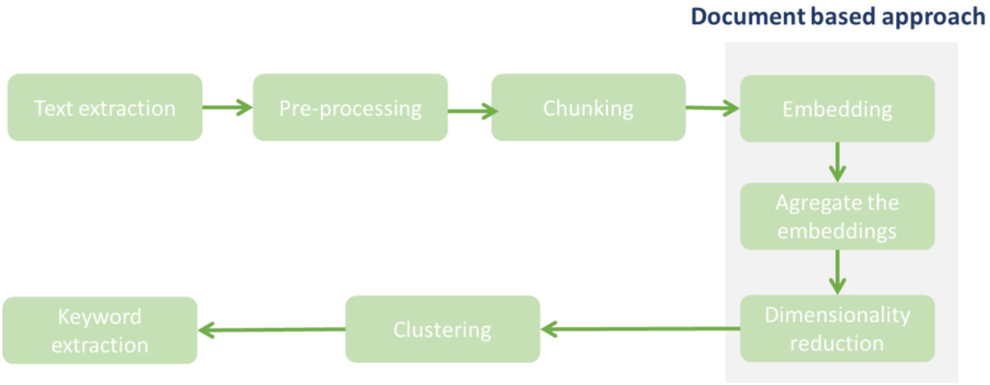


Figure 2. Document-based approach pipeline

Different aggregation methods can be employed to combine the individual embeddings of text chunks into a single representation for each document, such as mean aggregation and sum aggregation. The mean aggregation method calculates the average of the embeddings across all text chunks within a document. Each chunk embedding is a high-dimensional vector, and the mean aggregation creates a single vector by averaging the embeddings. This approach helps create a balanced representation of the document, in which each chunk contributes equally to the final document embedding. Alternatively, sum aggregation combines the embeddings by summing the vectors of all chunks within a document. Both methods allow us to aggregate the chunk-level embeddings into a document-level representation that captures the overall meaning of the document.

3.4.2 Clustering Chunk-Based Approach with LLM

Chunk-based approach, represented in Figure 3, involves clustering smaller segments or "chunks" of a document rather than the entire document.

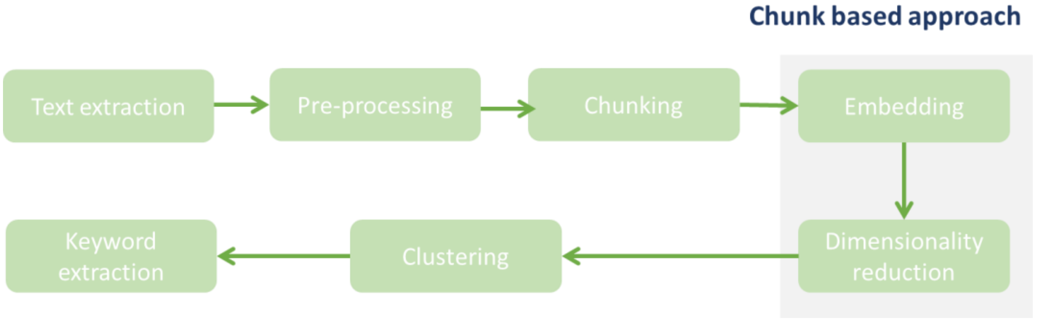


Figure 3. Chunk-based approach pipeline

By dividing the document into chunks and clustering them individually, this method captures more localized patterns and thematic nuances within the text.

3.5 Keyword and Topic Extraction

To build a specialized vocabulary in a multidisciplinary organization such as the CSTB, we need keywords that effectively characterize documents, thereby facilitating access to the knowledge they contain. This keyword and topic extraction is carried out in two steps:

- *Vectorization*: the text of each document is transformed into a term frequency matrix (per document). Each entry in this matrix indicates how many times a specific word appears in each document.
- *Calculation of c-TF-IDF* (class-based TF-IDF): once the clusters are formed, c-TF-IDF [35] is calculated for each cluster. This provides a weighting of terms based on their relative importance within that cluster.

4 Description of Experiments

We have tested our unsupervised and modular handling of a large corpus of unlabeled multidisciplinary documents using a bilingual document corpus, which was divided into two parts based on language (French and English). Specific models for embeddings and natural language processing techniques were applied to each part. The aim of this experimentation is threefold: (1) to perform unsupervised clustering of documents and text chunks, (2) to extract keywords following the two clustering-based approaches, and (3) to evaluate the results to identify the best approach, keywords, and topics for building the CSTB vocabulary.

For English documents, we used all-MiniLM-L6-v2 [36] to generate embeddings that facilitate the grouping of similar texts, thereby simplifying the classification and clustering of documents. On the other hand, for French documents, we employed paraphrase-multilingual-MiniLM-L12-v2, a multilingual model able to generate high-quality embeddings for French and other languages. We use specific algorithms for each step of the process. For embedding, we selected Sentence Transformers [31] due to their ability to capture rich semantic representations. For dimensionality reduction, we opted for UMAP [32], which is fast and preserves both local and global data structures. For clustering, we used HDBSCAN [38], a robust algorithm that does not require specifying the number of clusters in advance and handles variable densities well. Keyword extraction was performed using CountVectorizer combined with c-TF-IDF [35], emphasizing the importance specific to the clusters. Considering the clustering-based approach, we applied a mean aggregation on both languages in order to merge chunks of each document. Table 1 summarizes experimentation choices applied for each language.

Table 1. Adapting models based on the language of the documents

English Documents	**French Documents**
• *Model*: all-MiniLM-L6-v2 • *Description*: A lightweight and efficient model optimized for generating sentence embeddings in English. • *Dimensions*: 384	• *Model:* paraphrase-multilingual- MiniLM-L12-v2 • *Description:* A multilingual model capable of generating high-quality embeddings for French and other languages. • *Dimensions:* 384

5 Results and Evaluation

5.1 Documentary Corpus Description

Most of the available documents at CSTB are research reports published from 2000 to 2024, in PDF format, which cover different scientific domains such as civil engineering, fire safety domain, and health in buildings. This corpus contains a total of 6,627 files, of which 3,279 are in French, 3,055 are in English, and 293 documents were deemed unexploitable. The primary causes of unusability are attributed to various issues: many PDFs are scanned documents without OCR processing, making their content non-searchable; others are poorly encoded, causing issues with readability and accessibility; and some PDFs are completely empty, offering no usable data. For the French-language documents, the analysis identified 2,808 unique files and 419 groups of duplicates, where each group contains identical files reduced to a single representative. Similarly, for the English-language documents, there are 2,265 unique files and 716 groups of duplicates. This detailed categorization and grouping of duplicate files improves the efficiency of document management and helps streamline further analysis by ensuring that redundant data does not clutter the dataset. The findings highlight the importance of preprocessing and enhancing document quality to ensure optimal usability in any knowledge management or data analysis effort.

5.2 The Evaluation and Validation Process

Given the importance of the extracted candidate keywords for the construction of specialized vocabulary and the significance of this vocabulary in the knowledge management strategy at CSTB, we opted for a human validation process of the results. Indeed, this evaluation and validation process has followed a cyclical and iterative process, depicted in Figure 4, carried out with a panel of domain experts. For this purpose, a committee of CSTB experts from different areas of expertise (civil engineering, building health, etc.) was formed. Its role is to identify the best clustering methods, NLP techniques, and topic naming to identify scientific domains.

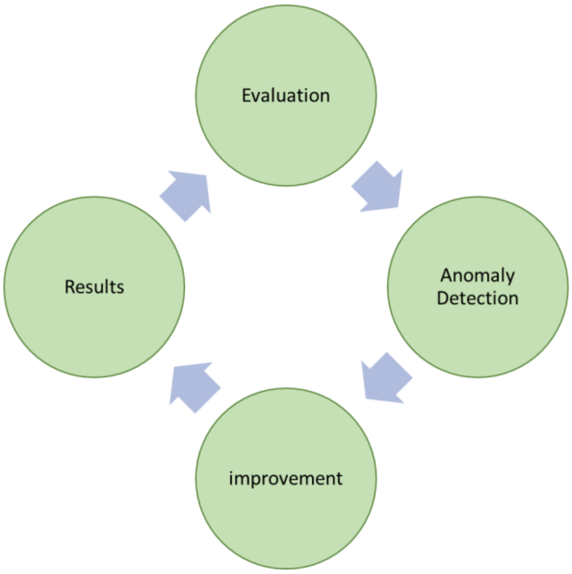


Figure 4. Validation and evaluation process

This panel of experts consists of: (i) A civil engineer who use to contribute to document indexing (15 years of experience at CSTB); (ii) A researcher specializing in building health and comfort (40 years at CSTB); (iii) Two expert documentalists (28 and 20 years of experience each in technical document indexing and content management) and occasionally an information technology watcher (14 years of experience at CSTB).

The committee's role was to guide, based on the results, the choice of the best clustering methods and NLP techniques, and to name the topics corresponding to the resulting clusters to identify the scientific fields. This committee, which is already in charge of leading workshops to build the CSTB's semantic repository, applied an analysis approach based on: (1) the thematic relevance of keywords for their future integration into the semantic repository, (2) their suitability for the organization's multidisciplinary challenges, and (3) their potential for interoperability with existing systems.

The validation process was organized into four working meetings (each lasting approximately two hours). During these meetings, the results of the clustering and extraction were presented, and requirements/recommendations were established for the keywords. Among those, the committee established three selection criteria: (1) keywords must not be verbs, (2) they must not contain numbers, city names, or country names, and (3) they must consist of one-grams or two-grams. The experts have also defined, by consensus, a key criterion for evaluating the results, namely the number of chunks or documents per topic according to the applied clustering approach. They have shown their commitment to this metric, particularly in order to identify documents that may be underrepresented but are of interest to research teams.

5.3 The Results and Discussion

The objective of these experiments is to observe the extracted keywords from the CSTB corpus using both proposed approaches and to identify the best approach that extracts keywords covering the maximum number of domains and documents at CSTB.

In our document-based clustering experiment, conducted on 2,265 English documents, we have identified six distinct topics. Each topic represents a cluster, which is characterized by the extracted keywords and the number of documents it contains. Figure 5 illustrates topics obtained with a document-based approach. We may observe that identified topics seem coherent considering their composing keywords. For instance, topic 0 seems related to building modeling, topic 1 groups acoustic-related keywords, while topic 2 concerns wind-related ones. As illustrated in Figure 6, we can see that by using bi-gram extraction, the topic number 3 contains bi-gram “wind speed” and “wind tunnel” that the validation committee identifies as important keywords, which refer to experiments conducted with the Jules Verne wind tunnel, a research facility for technical research and evaluations at CSTB. The committee found that bi-gram words provide more meaningful insights compared to simple keywords. We can see that most of the documents contain “building”, “constructions”, and related words, which is normal due to the context of all documents.

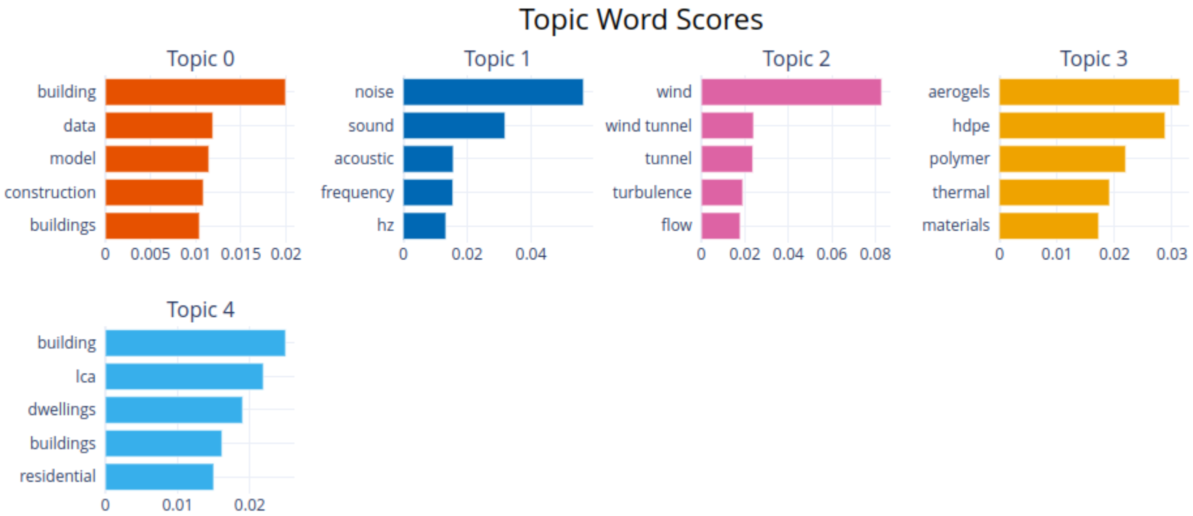
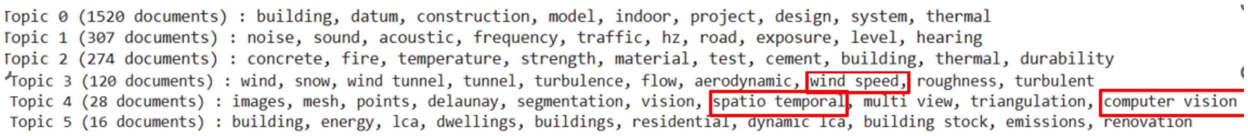


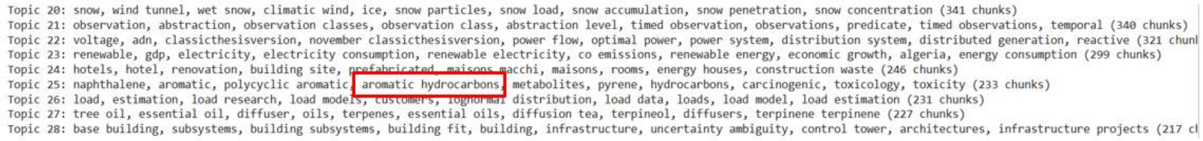
Figure 5. Keywords and topics distribution obtained from document-based clustering



```
Topic 0 (1520 documents) : building, datum, construction, model, indoor, project, design, system, thermal
Topic 1 (307 documents) : noise, sound, acoustic, frequency, traffic, hz, road, exposure, level, hearing
Topic 2 (274 documents) : concrete, fire, temperature, strength, material, test, cement, building, thermal, durability
Topic 3 (120 documents) : wind, snow, wind tunnel, tunnel, turbulence, flow, aerodynamic, wind speed, roughness, turbulent
 Topic 4 (28 documents) : images, mesh, points, delaunay, segmentation, vision, spatio temporal, multi view, triangulation, computer vision
 Topic 5 (16 documents) : building, energy, lca, dwellings, buildings, residential, dynamic lca, building stock, emissions, renovation
```

Figure 6. Results from document-based clustering extraction with bi-gram

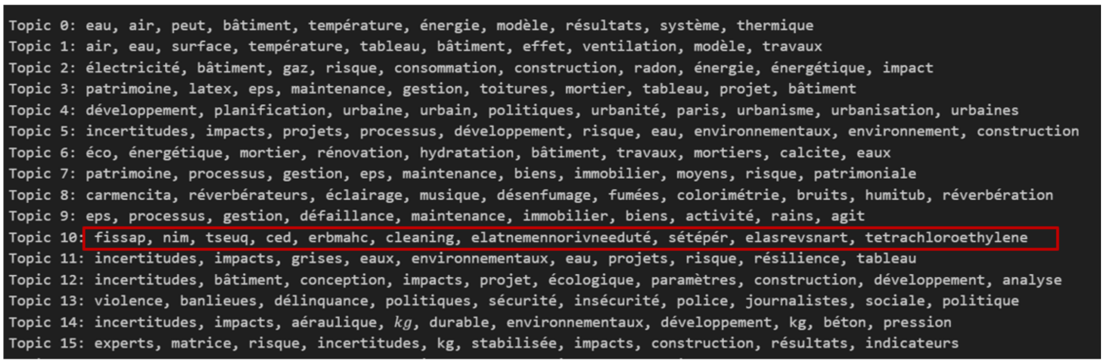
The same English corpus of documents was tested with the chunk-based approach. As depicted in Figure 7, the chunk-based approach offers more topics and keywords. The fields are more comprehensively covered and broader, ensuring that no areas addressed by CSTB were neglected. The extracted bigrams also provide more semantic richness, according to the experts, similar to the first approach. The committee indeed considered such results richer and closer to the nature of the content of CSTB documents.



```
Topic 20: snow, wind tunnel, wet snow, climatic wind, ice, snow particles, snow load, snow accumulation, snow penetration, snow concentration (341 chunks)
Topic 21: observation, abstraction, observation classes, observation class, abstraction level, timed observation, observations, predicate, timed observations, temporal (340 chunks)
Topic 22: voltage, adn, classicthesisversion, november classicthesisversion, power flow, optimal power, power system, distribution system, distributed generation, reactive (321 chunl
Topic 23: renewable, gdp, electricity, electricity consumption, renewable electricity, co emissions, renewable energy, economic growth, algeria, energy consumption (299 chunks)
Topic 24: hotels, hotel, renovation, building site, prefabricated, maisons macchi, maisons, rooms, energy houses, construction waste (246 chunks)
Topic 25: naphthalene, aromatic, polycyclic aromatic, aromatic hydrocarbons, metabolites, pyrene, hydrocarbons, carcinogenic, toxicology, toxicity (233 chunks)
Topic 26: load, estimation, load research, load models, customers, lognormal distribution, load data, loads, load model, load estimation (231 chunks)
Topic 27: tree oil, essential oil, diffuser, oils, terpenes, essential oils, diffusion tea, terpineol, diffusers, terpinene terpinene (227 chunks)
Topic 28: base building, subsystems, building subsystems, building fit, building, infrastructure, uncertainty ambiguity, control tower, architectures, infrastructure projects (217 cl
```

Figure 7. Results from chunk-based clustering extraction with bi-gram

However, when testing the corpus in French, we encountered many inconsistent results for both the chunk-based and document-based approaches, as illustrated in Figure 8. For instance, topic 10 in Figure 8 concentrates several words that could not be interpreted by the experts, such as “fissap”, which are probably a consequence of the text overlapping used during the chunking step. Experts felt uncomfortable about this phenomenon. Although other identified topics were considered relevant by the experts, this phenomenon demonstrates some limitations of our approach, indicating that there are still improvements to be made, such as incorporating slang words.



```
Topic 0: eau, air, peut, bâtiment, température, énergie, modèle, résultats, système, thermique
Topic 1: air, eau, surface, température, tableau, bâtiment, effet, ventilation, modèle, travaux
Topic 2: électricité, bâtiment, gaz, risque, consommation, construction, radon, énergie, énergétique, impact
Topic 3: patrimoine, latex, eps, maintenance, gestion, toitures, mortier, tableau, projet, bâtiment
Topic 4: développement, planification, urbaine, urbain, politiques, urbanité, paris, urbanisme, urbanisation, urbaines
Topic 5: incertitudes, impacts, projets, processus, développement, risque, eau, environnementaux, environnement, construction
Topic 6: éco, énergétique, mortier, rénovation, hydratation, bâtiment, travaux, mortiers, calcite, eaux
Topic 7: patrimoine, processus, gestion, eps, maintenance, biens, immobilier, moyens, risque, patrimoniale
Topic 8: carmencita, réverbérateurs, éclairage, musique, désenfumage, fumées, colorimétrie, bruits, humitub, réverbération
Topic 9: eps, processus, gestion, défaillance, maintenance, immobilier, biens, activité, rains, agit
Topic 10: fissap, nim, tseuq, ced, erbmahc, cleaning, elatnemennorivneedutè, sétépér, elasrevsnart, tetrachloroethylene
Topic 11: incertitudes, impacts, grises, eaux, environnementaux, eau, projets, risque, résilience, tableau
Topic 12: incertitudes, bâtiment, conception, impacts, projet, écologique, paramètres, construction, développement, analyse
Topic 13: violence, banlieues, délinquance, politiques, sécurité, insécurité, police, journalistes, sociale, politique
Topic 14: incertitudes, impacts, aéraulique, kg, durable, environnementaux, développement, kg, béton, pression
Topic 15: experts, matrice, risque, incertitudes, kg, stabilisée, impacts, construction, résultats, indicateurs
```

Figure 8. Inconsistent results from the chunk-based approach extraction for the French corpus

Setting aside the processing of French documents for now, the results retained by the committee after four working meetings are that the best clustering method is chunk-based.

6 Conclusion and Perspectives

In this article, we present an unsupervised and modular handling a large corpus of unlabeled multidisciplinary documents to facilitate the creation and enrichment of vocabularies for knowledge management purposes. We explore the application of clustering techniques combined with NLP-driven keyword extraction to support the construction of specialized vocabularies that address the multidisciplinary nature of the content at CSTB, a French scientific research center specializing in buildings. We experimented with two approaches for keyword extraction: document-based clustering and chunk-based clustering. This article provides a detailed overview of the proposed methodologies, presents the results of our experiments, and describes the human validation process used to evaluate these results. Our comprehensive analysis of keyword extraction methods focuses on the chunk-based approach. According to the validation committee, for texts in English, this approach is preferred due to its superior granularity in keyword extraction.

The results indicate that the chunk-based method not only identifies a diverse range of keywords but also significantly reduces noise, leading to more coherent thematic clusters.

In future works, we aim at adding either a new layer of algorithms or a new layer of a language model (LLM) to enhance representation by employing MMR [39] in order to balance relevance and diversity, thereby avoiding redundancy in topic keywords. The anomalies detected in the French corpus should also be the target of future works, particularly considering the use of French-dedicated models, such as CamemBERT [40] and CamemBERT 2.0 [41]. Finally, the expert committee intends to assign thematic labels to the identified clusters (“Acoustics”, “Energy”, etc.), allowing the use of those clusters for enriching CSTB’s semantical referential.

References

[1] P. Maharjan, “Knowledge Management Enablers for Knowledge Creation Combination in Nepalese Hospitality Industry,” *Journal of Balkumari College*, vol. 9, no. 1, 2020, pp. 25–33. Available: https://doi.org/10.3126/jbkc.v9i1.30064

[2] R. Morse, “Management in the 21st Century Knowledge Management Systems: Using Technology to Enhance Organizational Learning,” *Proceedings of the 2000 Information Resources Management Association International Conference on Challenges of Information Technology Management in the 21st Century*, pp. 426–429, 2000.

[3] R. Y. Narazaki, M. Silveira Chaves and C. Drebes Pedron, “A project knowledge management framework grounded in design science research,” *Knowledge and Process Management*, vol. 27, no. 3, pp. 197–210, 2020. Available: https://doi.org/10.1002/kpm.1627

[4] J. Priti, “An Empirical Study of Knowledge Management in University Libraries in SADC Countries,” *New Research on Knowledge Management Applications and Lesson Learned*, pp. 137–154, 2012. Available: https://doi.org/10.5772/36309

[5] L. Yao-Sheng, “The effects of knowledge management strategy and organization structure on innovation,” *International Journal of Management*, vol. 24, no. 1, pp. 53–60, 2007.

[6] H. Laihonen, A. A. Kork, and L. M. Sinervo, “Advancing public sector knowledge management: towards an understanding of knowledge formation in public administration,” *Knowledge Management Research & Practice*, vol. 22, no. 3, pp. 223–233, 2024. Available: https://doi.org/10.1080/14778238.2023.2187719

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and A. Neelakantan et al., “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, vol. 33, pp. 1877–1901, 2020.

[9] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Texts,” *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, 2004.

[10] X. Wan and J. Xiao, “Single Document Keyphrase Extraction Using Neighborhood Knowledge,” *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 855–860, 2008.

[11] M. Grootendorst, “BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling,” *arXiv:2203.05794*, 2022. Available: https://doi.org/10.48550/arXiv.2203.05794

[12] R. Churchill and L. Singh. “The Evolution of Topic Modeling,” *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–35, 2022. Available: https://doi.org/10.1145/3507900

[13] R. K. Bisht, “A Comparative Evaluation of Different Keyword Extraction Techniques,” *International Journal of Information Retrieval Research (IJIRR)*, vol. 12, no. 1, pp. 1–17, 2022. Available: https://doi.org/10.4018/IJIRR.289573

[14] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988. Available: https://doi.org/10.1016/0306-4573(88)90021-0

[15] A. Delamaire, M. Beigbeder, and M. Juganaru-Mathieu, “Exploitation de syntagmes dans la découverte de thèmes,” *Actes de la conférence CORIA (Conférence en Recherche d'Information et Applications)*, 2019 (in French).

[16] A. Bougouin, F. Boudin, and B. Daille, “TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction,” *International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 543–551, 2013.

[17] L. N. Khelifa, N. Lammari, J. Akoka, and T. Bouabana-Tebibel, “Building Contextualized Topic Maps,” *19th IBIMA (International Business Information Management Association) Conference on Innovation Vision 2020*: *Sustainable Growth, Entrepreneurship, Real Estate and Economic Development*, 2012.

[18] W. D. Abilhoa and L. N. De Castro, “TKG: A graph-based approach to extract keywords from tweets,” *Distributed Computing and Artificial Intelligence, 11th International Conference*, Springer, pp. 425–432, 2014. Available: https://doi.org/10.1007/978-3-319-07593-8_49

[19] H. M. M. Hasan, F. Sanyal, and D. Chaki, “A novel approach to extract important keywords from documents applying latent semantic analysis,” *10th International Conference on Knowledge and Smart Technology (KST)*, pp. 117–122, 2018. Available: https://doi.org/10.1109/KST.2018.8426144

[20] A. Ahadh, G. V. Binish, and R. Srinivasan, “Text mining of accident reports using semi-supervised keyword extraction and topic modeling,”. *Process Safety and Environmental Protection*, vol. 155, pp. 455–465, 2021. Available: https://doi.org/10.1016/j.psep.2021.09.022

[21] M. Umair, A. Khan, F. Ullah, A. Masmoudi, and M. Faheem, “Global and Local Context Fusion in Heterogeneous Graph Neural Network for Summarizing Lengthy Scientific Documents,” *IEEE Access*, vol. 13, pp. 53433–53447, 2025. Available: https://doi.org/10.1109/ACCESS.2025.3553755

[22] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *KDD’96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

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

[24] N. B. Mansour, H. Rahimi, and M. Alrahabi, “How Well Do Large Language Models Extract Keywords? A Systematic Evaluation on Scientific Corpora,” *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, pp. 13–21, 2025. Available: https://doi.org/10.18653/v1/2025.aisd-main.2

[25] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview,” *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007. Available: https://doi.org/10.4018/jdwm.2007070101

[26] J. Zhou, Y. Jia, Y. Qiu, and L. Lin, “The potential of applying ChatGPT to extract keywords of medical literature in plastic surgery,” *Aesthetic Surgery Journal*, vol. 43, no. 9, pp. NP720–NP723, 2023. Available: https://doi.org/10.1093/asj/sjad158

[27] O. Akarsu and H. Parmaksiz. “Anatomy of Digital Leadership Studies: An Analysis with Topic Modeling Approaches,” *Business and Economics Research Journal*, vol. 16, no. 2, pp. 179–205, 2025. Available: https://doi.org/10.20409/berj.2025.463

[28] F. Anowar, S. Sadaoui, and B. Selim, “Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE),” *Computer Science Review*, vol. 40, 2021. Available: https://doi.org/10.1016/j.cosrev.2021.100378

[29] K. Church, “Word2Vec,” *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017. Available: https://doi.org/10.1017/S1351324916000334

[30] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic BERT sentence embedding,” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, 2022. Available: https://doi.org/10.18653/v1/2022.acl-long.62

[31] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. Available: https://doi.org/10.18653/v1/D19-1410

[32] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *arXiv:1802.03426*, 2018. Available: https://doi.org/10.48550/arXiv.1802.03426

[33] L. Van der Maaten, and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[34] H. Abdi and L. J. Williams, “Principal component analysis,”. *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. Available: https://doi.org/10.1002/wics.101

[35] D. D. Xu and S. B. Wu, “An improved TFIDF algorithm in text classification,” *Applied Mechanics and Materials*, vol. 651–653, pp. 2258–2261, 2014. Available: https://doi.org/10.4028/www.scientific.net/AMM.651-653.2258

[36] W. Wenhui, W. Furu, D. Li, B. Hangbo, Y. Nan, and Z. Ming, “MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers,” *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS’20)*, pp. 5776–5788, 2020.

[37] M. Ciancone, I. Kerboua, M. Schaeffer, and W. Siblini, “MTEB-French: Resources for French Sentence Embedding Evaluation and Analysis,” *arXiv:2405.20468*, 2024. Available: https://doi.org/10.48550/arXiv.2405.20468

[38] C. Malzer and M. Baum, “A hybrid approach to hierarchical density-based cluster selection,” *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 223–228, 2020. Available: https://doi.org/10.1109/MFI49285.2020.9235263

[39] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information Retrieval*, ACM, pp. 335–336, 1998. Available: https://doi.org/10.1145/290941.291025

[40] L. Martin, B. Muller, P. Javier, O. Suárez, Y. Dupont, L. Romary, E. De la Clergerie, D. Seddah, and B. Sagot, “CamemBERT: a tasty French language model,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, 2020. Available: https://doi.org/10.18653/v1/2020.acl-main.645

[41] W. Antoun, F. Kulumba, R. Touchent, E. De la Clergerie, B. Sagot, and D. Seddah, “CamemBERT 2.0: A Smarter French Language Model Aged to Perfection,” *arXiv:2411.08868*, 2024. Available: https://doi.org/10.48550/arXiv.2411.08868