Semantic Analysis Experiments for French Citizens' Contribution : Combinations of Language Models and Community Detection Algorithms

Sami Guembour[1] **Catherine Dominguès**[1] **Sabine Ploux**[2]

[1]Univ Gustave Eiffel, ENSG, IGN, LASTIG,
F-77420 Champs-sur-Marne, France
firstname.lastname@ign.fr

[2]Centre d'Analyse et de Mathématique Sociales CNRS-UMR 8557,
Ecole des Hautes Etudes en Sciences Sociales, Paris, France
firstname.lastname@ehess.fr

Abstract

Following the Yellow Vest crisis that occurred in France in 2018, the French government launched the *Grand Débat National*, which gathered citizens' contributions. This paper presents a semantic analysis of these contributions by segmenting them into sentences and identifying the topics addressed using clustering techniques. The study tests several combinations of French language models and community detection algorithms, aiming to identify the most effective pairing for grouping sentences based on thematic similarity. Performance is evaluated using the number of clusters generated and standard clustering metrics. Principal Component Analysis (PCA) is employed to assess the impact of dimensionality reduction on sentence embeddings and clustering quality. Cluster merging methods are also developed to reduce redundancy and improve the relevance of the identified topics. Finally, the results help refine semantic analysis and shed light on the main concerns expressed by citizens.

Keywords: Semantic analysis . Language models . Community detection . Clustering . Dimensionality reduction . Cahiers Citoyens

1 Introduction

As an answer to the Yellow Vest crisis, in January 2019, the French government launched the *Grand Débat National*[1] [in English, Large National Debate] (GDN) offering both a dematerialized digital platform and physical supports, called *Cahiers Citoyens* [Citizens' Notebooks], leaved in various public places (town halls, roundabouts, hospitals, prisons, etc.). These notebooks enabled citizens to freely express their views on topics of their choice, choosing the format (letters, paragraphs, emails, bullet lists, petitions) and length (ranging from a few words to several pages) that suited them. At the GDN close, in mid-March 2019, *Cahiers Citoyens* gathered 225,224 contributions located to the place where each one had been written or deposited. Among the 34,970 municipalities in France in 2019, 17,014 proposed at least one notebook.

A team has been formed to conduct a semantic analysis of the content of *Cahiers Citoyens*, and this paper is part of the project's framework. The analysis is based on both the text of the contributions and their location.[2] Due to the volume of contributions, the adopted approach consisted in identifying the topics they addressed (Guembour, 2024). To achieve this, clustering was applied to the texts of the contributions using community detection algorithms. However, the first clustering results varied widely in number of clusters and of unclassified citizens' contributions, making necessary to explore various combinations of parameters (algorithms, hyperparameters, language models, etc.) and post-treatments. After a presentation of related works in Section 2 and the corpus of *Cahiers Citoyens* in Section 3, this article describes the end-to-end process implemented to identify the contributions' semantic organization, combining text representation models and community detection algorithms. Section 4 presents the tested combinations and Section 5 evaluates them through different indexes. Section 6 introduces post-treatments intended to enhance clustering performance, while Section 7 provides a detailed assessment of these methods along with

[1]https://granddebat.fr/

[2]One hypothesis, supported by numerous previous sociological studies, is that citizen expression depends on the location where it is produced.

Proceedings of the 16th International Conference on Computational Semantics, pages 241–251

the final results. Finally, Section 8 presents the main conclusions of this study and discusses the perspectives opened by this work.

2 Related Works

Community detection in graphs constructed from textual data has emerged as a widely adopted technique in text mining, enabling the unsupervised discovery of latent thematic structures. Among the most commonly used algorithms, Louvain (Blondel et al., 2008) and the Label Propagation Algorithm (LPA) (Zhu and Ghahramani, 2003) are particularly prominent for their ability to identify coherent clusters within semantic graphs.

Several studies have employed the Louvain algorithm specifically for topic modeling across large textual corpora. For example, (Marco et al., 2024) used Louvain to improve the semi-supervised clustering of customer reviews in the domain of customer services. (Monnet and Loïc, 2024) combined doc2vec representations with Louvain, k-means, and spectral clustering to enhance topic classification across a broad document collection. (Chowdhury et al., 2023) reformulated the topic modeling task as a community detection problem in a word co-occurrence graph generated from a text corpus. Similarly, (Wang et al., 2021) applied Louvain to cluster COVID-19-related articles by thematic similarity, following an automatic summarization process. In all these cases, Louvain demonstrated strong capabilities in uncovering semantically meaningful clusters from unstructured textual data. (Boutalbi et al., 2022) introduced an innovative method, IEcons (Implicit and Explicit Consensus), which combines multiple textual representations —including TF-IDF, Word2Vec, and BERT embeddings— to improve the robustness of clustering. Their approach uses a dual consensus strategy: explicit consensus through the aggregation of clustering results obtained from each representation independently, and implicit consensus through the fusion of similarity matrices into a unified similarity tensor. For the final clustering step, several algorithms are evaluated, including Louvain, which is particularly effective in extracting dense communities from the resulting weighted graph. In a different application, (Abdine et al., 2022) leveraged the Louvain algorithm to detect political communities from a user graph built from French tweets, where edges are defined by retweet behavior. Although the graph structure is based on user interaction rather than content similarity, this work reflects a growing interest in combining community detection techniques with language models such as RoBERTa and CamemBERT, which the authors use for offensive language detection.

In parallel, the Label Propagation Algorithm (LPA) has also received attention due to its simplicity and computational efficiency on large graphs. (Tang et al., 2022) proposed a classification framework for scientific and technical documents (e.g., patents and academic papers) using Word2Vec embeddings and a consensus clustering approach based on LPA. (Pawar et al., 2018) developed an LPA-based method for weakly supervised text classification, where documents are modeled as nodes in a similarity graph, and labels are propagated through the network. (Han et al., 2016) focused on improving LPA itself, introducing a modified version, LPAf, that enhances the quality of detected communities in large-scale networks. These contributions illustrate LPA's suitability for fast, scalable classification and clustering tasks over vast document sets.

Beyond Louvain and LPA, other methods have been proposed that integrate semantic information directly into the graph structure. For instance, a community detection method was developed by (Ruan et al., 2013), incorporating both network connectivity and TF-IDF scores of textual content, demonstrating improved thematic coherence in the resulting communities. Similarly, (Gao et al., 2023) proposed a sentiment-aware community detection framework, where TF-IDF vectors and sentiment scores are jointly used to construct a weighted graph reflecting both topical and emotional affinities between users in social networks. This approach enhances the identification of semantically and emotionally coherent communities.

To facilitate the application of community detection algorithms, several studies have introduced dimensionality reduction techniques, particularly when working with high-dimensional vector representations of textual data. These methods aim to project the original graph or embedding space into a lower-dimensional representation while preserving the essential topological or semantic properties of the data. For instance, (Aman et al., 2021) employ structural embedding methods such as DeepWalk and Node2Vec to learn low-dimensional node embeddings, enabling more efficient community detection. Unlike semantic-based approaches,

these methods focus exclusively on the structural properties of the network.

While most studies focus on general-purpose corpora or domains such as customer reviews or social media, few works have addressed the analysis of deliberative citizen-generated content. Yet, this type of corpus —as exemplified by the *Cahiers Citoyens* or participatory platforms— raises important challenges due to its thematic diversity, variability in writing quality, and lack of structure.

A government-commissioned report by the Roland Berger firm (Berger and Bluenove, 2019), in collaboration with the agency Cognito Consulting[3], served as a starting point for the analysis of the *Cahiers Citoyens*. The approach relied on semantic mapping to cluster textual contributions into eight major themes: democracy and citizenship (144,071 ideas), ecological transition (89,103), taxation and public spending (138,667), state organization and public services (62,597), economy and employment (26,686), education and training (9,638), purchasing power (75,652), and health, solidarity, and integration (63,574). However, the methodology has been criticized for its lack of transparency, particularly regarding the definition of what constitutes an "expressed idea", the algorithmic procedures used, and the rapidity with which the results were delivered — all of which raise questions about the robustness and interpretability of the findings.

(Ray, 2023) explored topic extraction methods such as BERTopic (Grootendorst, 2022) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to analyze the contributions from the *Cahiers Citoyens* corpus. The objective was to compare the extracted thematic structures with those identified in the Roland Berger firm report. This approach provides a renewed perspective on the thematic diversity present in citizen contributions. The most salient clusters uncovered by the analysis relate to issues such as pension increases, rural life, education, ecology and agriculture, electoral processes including recognition of blank votes, speed limitations on highways, and taxation.

(Monnier, 2023) conducted an in-depth study on the theme of wind power based on the *Cahiers Citoyens* corpus. Her work adopts a cross-disciplinary perspective in the social sciences, combining linguistic and geographical approaches. The analysis focused on three departments where wind-related concerns were particularly salient, allowing for a territorialized interpretation of contributions based on the natural and social characteristics specific to each region. Spatialized text extractions were visualized through map-based representations.

While previous research has explored a wide range of textual representations — from traditional models such as TF-IDF and Word2Vec to more advanced embeddings from BERT— none, to the best of our knowledge, have leveraged pretrained Transformer-based language models specifically designed for French (such as CamemBERT) to represent texts, followed by dimensionality reduction of these vectors to construct an optimized semantic graph, on which community detection algorithms are applied. This is the approach proposed here in order to reveal latent topics in large-scale citizen-generated content from the *Cahiers Citoyens*.

3 Corpus of *Cahiers Citoyens*

The notebooks of *Cahiers Citoyens* were made up of handwritten and/or typed texts, emails sent directly to the city councils, files sometimes including attachments, as well as collective petitions, and reflect a diversity of citizen concerns. A textometric and spatialized analysis of the corpus is presented in detail in (Dominguès and Jolivet, 2024).

3.1 Contribution Segmentation into Sentences

Previous topic modeling analyses of *Cahiers Citoyens* (Ray, 2023) revealed that a single contribution often addresses multiple topics. Consequently, the contribution could not serve as the unit of semantic analysis. Therefore, the contributions were segmented into sentences, resulting in a corpus of 4,200,831 sentences, hereafter referred to as CC. This finer granularity was intended to facilitate the identification of meaning units and their grouping into clusters. This segmentation was performed using the Spacy model (Honnibal et al., 2020) based on transformers.[4] This choice was based on two main criteria:

- Model performance: This model achieved the highest performance among those available in Spacy, with a sentence segmentation accuracy of 0.92;
- Adaptability to the specificities of the corpus: the contributions of *CC* come from citizens

[3] https://www.cognito.fr/

[4] *fr_dep_news_trf* : https://spacy.io/models/fr#fr_dep_news_trf

with varied profiles, leading to typographical variations, such as the absence of capital letters at the beginning of sentences and the lack of strong punctuation marks at the end of sentences in numerous contributions. This model has proven capable of segmenting sentences effectively, even when these typographic markers are absent.

3.2 Sentence Preprocessing

The segmentation of contributions into sentences revealed that a number of them are frozen or fixed, in the sense they contain no information about the themes and topics raised by citizens in their contributions. These sentences include elements such as contribution dates, contributor names, recipients, polite expressions, etc. (e.g., *Je vous prie d'agréer, Monsieur, mes sincères salutations* [Please accept, Sir, my sincere greetings], *Mardi 18 décembre 2018* [Tuesday, December 18, 2018]). In order to prevent them from affecting the semantic analysis, they were removed during a preprocessing phase. This filtering step helps reduce memory consumption and speeds up clustering computations by eliminating non-informative sentences for topic modeling. The method used is based on two complementary approaches:

- Detection based on syntactic patterns: identification of specific linguistic structures such as dates, email addresses, phone numbers, etc.
- Clustering by semantic similarity: use of the Fast Clustering algorithm to automatically identify formatted sentences by grouping them according to their similarity (for example, one cluster for dates and another for polite expressions).

Through this preprocessing phase, the total number of sentences was reduced to 2,789,465, resulting in a refined corpus referred to as *filtered-CC*. However, due to limited computational resources (in terms of both memory and processing time), it was not feasible to process the entire corpus. Therefore, a random sample of 50,000 sentences[5] was selected from *filtered-CC* to conduct the study. Table 1 presents the different versions of the corpus.

[5]Statistical tests were conducted to assess the representativeness of the sample with respect to the full corpus. Results indicate that the sample is representative in terms of sentence length and morphosyntactic distribution.

Table 1: Table detailing the different corpus versions

Corpus	Unit	Count
Cahiers Citoyens	Contribution	225,224
CC	Sentence	4,200,831
filtered-CC	Sentence	2,789,465
Sample of *filtered-CC*	Sentence	50,000

4 Combinations of Language Models and Community Detection Algorithms

As stated in Section 1, the proposed method consists in representing sentences as embeddings (vectors) using language models, then applying clustering algorithms to these embeddings to obtain clusters. The objective is to compare different combinations of language models and clustering algorithms in order to identify the one that provides the best clustering of sentences and, consequently, the most accurate identification of topics. In addition, each combination is tested both with and without dimensionality reduction using Principal Component Analysis (PCA) (Hotelling, 1933). The purpose of applying PCA is to evaluate the impact of dimensionality reduction on sentence embeddings and clustering quality. For sentence vector representations, three language models were selected due to their high performance in French: CamemBERT-large (Reimers and Gurevych, 2019; Martin et al., 2020), Solon-large[6], and Distil-CamemBERT (Delestre and Amar, 2022). Regarding clustering, since the exact number of topics (clusters) discussed in *Cahiers Citoyens* is unknown, we opted for community detection algorithms, which are designed to uncover structure in graphs without requiring a predefined number of clusters. To do this, a graph has been constructed where each sentence embedding represents a node, and an edge is established between two nodes if their cosine similarity exceeds the threshold of 0.68 (this threshold was chosen based on a comparative analysis conducted by (Guembour, 2024), which examined various pairs of sentences).[7] Community detection algorithms are then applied to the graph to obtain clusters. The algorithms have been selected from related works (in Section 2): LPA (Label Propagation Algorithm) and the Louvain algorithm.
With three models, mixed or not with dimension-

[6]https://huggingface.co/OrdalieTech/Solon-embeddings-large-0.1

[7]The study showed that, in the corpus, some expressions appear either in their full form or as acronyms, and that a threshold of 0.68 effectively groups together these variations when they occur in similar contexts.

ality reduction, and two algorithms, we obtain 12 combinations to compare. Algorithm 1, presented below, describes the process of applying these combinations, while Table 2 shows that PCA enables to substantially reduce the number of embedding dimensions while retaining a large part of the inertia (90%).

Algorithm 1 Application of Language Model and Community Detection Algorithm Combinations

Input: Sample of 50,000 sentences from *filtered-CC*

Output: Sentence clusters

1: Select a language model m
2: Compute sentence embeddings using model m
3: Apply PCA to the sentence embeddings or not, retaining 90% of the inertia
4: Construct a graph G where each node represents a sentence embedding
5: Connect two nodes (i, j) if $similarity(i, j) \geq 0.68$
6: Select a community detection algorithm a
7: Apply a on G to detect communities
8: Evaluate the quality of the obtained clustering

Table 2: Number of embedding dimensions before and after PCA

Model	Initial Dimensions	With PCA
CamemBERT-large	1024	382
Solon-large	1024	334
Distil-CamemBERT	768	165

5 Evaluation and Interpretations of the Combinations

Evaluation: The performance of each combination (model, algorithm) is measured through the quality of the clustering. Several metrics adapted to unsupervised clustering and community detection have been selected: the Calinski-Harabasz index (CHI) (Caliński and Harabasz, 1974), the Davies-Bouldin index (DBI) (Davies and Bouldin, 1979), and Modularity (Newman, 2006). These metrics assess the internal cohesion of groups, the separation between clusters, and the structure of communities within the resulting graph. The CHI and the DBI assess clustering quality by measuring the separation and compactness of clusters, with a high value being desirable for the former and a low value for the latter. Modularity, on the other hand, measures the density of connections within communities in a graph, with a high value indicating well-defined communities. Table 3 provides the index values for each of the 12 combinations. In our case, since the objective is to identify the largest number of addressed topics, we consider that a good clustering is characterized by a high number of classified sentences and optimal evaluation metrics, particularly the CHI, the DBI, and modularity. The number of classified sentences corresponds to the number of nodes in the graph, as each sentence is represented by an embedding and becomes a node of the graph built for community detection. Thus, the more similar (in the sense of the semantic similarity measure) embeddings the model generates, the more nodes will be connected in the graph. Conversely, sentences that do not exhibit any link with others are not included in the graph.

Interpretations: Table 3 shows that the language model classifying the highest number of sentences is CamemBERT-large. The Distil-CamemBERT model classifies slightly fewer sentences than the CamemBERT-large model, indicating that it retains a good ability to capture similarities between sentence embeddings. In contrast, Solon-large classifies significantly fewer sentences, suggesting that its embeddings are less homogeneous and less effective at linking sentences within the graph, thereby reducing the number of nodes. For all three models, using embeddings without dimensionality reduction allows for a higher number of classified sentences compared to when PCA is applied. This means that dimensionality reduction via PCA decreases the model's ability to capture similarities between embeddings. Applying the Louvain algorithm to CamemBERT-large embeddings produces the smallest number of clusters, demonstrating a better ability to group similar elements than the LPA algorithm. Louvain generates approximately 1,000 fewer clusters with CamemBERT-large and Distil-CamemBERT, and approximately 500 fewer with Solon-large, suggesting that it better captures global thematic structures. In contrast, using PCA generally results in an increase in the number of clusters, as it reduces the similarity between embeddings, preventing them from being grouped together. This means that initially similar sentences may be considered dissimilar after reduction.

The CHI shows that CamemBERT-large delivers the best performance among the tested mod-

Table 3: Table detailing the performance of each combination. **w/o PCA** = without PCA; **w PCA** = with PCA. **Bold values** represent the best performances without PCA. Underlined values represent the best performances with PCA.

Metric	CamemBERT-large + Louvain		CamemBERT-large + LPA		Distil-CamemBERT + Louvain		Distil-CamemBERT + LPA		Solon-large + Louvain		Solon-large + LPA	
	w/o PCA	w PCA	w/o PCA	w PCA	w/o PCA	w PCA	w/o PCA	w PCA	w/o PCA	w PCA	w/o PCA	w PCA
# Classified Sentences	**23,375**	22,050	**23,375**	22,050	23,155	22,189	23,155	22,189	17,721	14,231	17,721	14,231
# Clusters	**2,398**	2,449	3,300	3,279	2,425	2,591	3,560	3,647	2,571	2,551	3,163	2,909
CHI	**7.25**	7.76	7.08	7.48	6.34	6.59	6.09	6.25	5.19	5.31	4.98	5.07
DBI	1.26	1.21	1.19	1.13	1.28	1.28	1.23	1.17	1.21	0.99	**1.14**	0.91
Modularity	0.88	0.89	0.86	0.87	0.90	0.92	0.87	0.90	**0.92**	0.92	0.90	0.90

els. This model produces well-separated and compact clusters, as reflected by its high CHI values. Although Distil-CamemBERT performs well, its results are slightly lower, suggesting that CamemBERT-large captures finer semantic relationships and thus provides better vector representations. Solon-large, with even lower CHI scores, appears less suited for this clustering task.

Regarding community detection algorithms, Louvain outperforms LPA across all three language models tested. Louvain generates better-separated and more homogeneous clusters, confirming its ability to identify distinct communities by minimizing cuts between them and producing more coherent clusters. The impact of PCA on clustering quality is also significant. All combinations with PCA show higher CHI than those without dimension reduction. PCA reduces dimensionality while enhancing cluster separation and compactness, although it can sometimes decrease the similarity between embeddings, preventing their clustering.

As mentioned before, a low DBI value indicates better clustering with more compact clusters, where cluster points are closer to their centroid. The obtained values show that the Solon-large model produces more compact clusters than CamemBERT-large and Distil-CamemBERT, at the cost of a smaller number of classified sentences, which reduces intra-cluster dispersion.

Concerning the algorithms,The LPA generates clusters with a lower DBI than Louvain for all models, indicating more cohesive and less dispersed groups. For all combinations, the use of PCA systematically reduces DBI values, confirming that dimensionality reduction improves cluster cohesion by limiting their internal dispersion.

In terms of modularity, the results show that Solon-large achieves slightly higher modularity because it classifies fewer sentences, reducing the density of the graph. Clustering with Louvain ensures better modularity than LPA, meaning that the detected communities are better defined. PCA has a very limited impact on the modularity of the CamemBERT-large and Distil-CamemBERT models. Indeed, although applying PCA to the embeddings of these two models reduces the number of nodes in the graph, it nevertheless remains dense.

In summary, the most suitable combination for our corpus appears to be CamemBERT-large paired with the Louvain algorithm. This configuration maximizes similarity between embeddings, groups more sentences into fewer clusters, and has the best CHI value. Although its DBI is not the lowest, it remains close to the values obtained with other combinations. Similarly, while some configurations show slightly better modularity, the difference remains marginal. Finally, PCA improves CHI, DBI, and modularity scores. However, it reduces the number of classified sentences, as it decreases the model's ability to capture similarities between sentence embeddings, thereby limiting the formation of clusters grouping semantically close sentences.

6 Post-Treatments to Merge Redundant Clusters

In Section 5, we identified the most effective combination for clustering, namely the use of CamemBERT-large for sentence embeddings and the Louvain algorithm for community detection. This configuration enables classifying the largest number of sentences while ensuring good cluster cohesion. However, the number of clusters obtained remains very high (2,394 without PCA and 2,446 with PCA). Yet, some clusters could be redundant in the sense they might address similar topics. So, optimizing clustering results could mean reducing the number of clusters and obtaining more populated clusters while improving performance according to CHI, DBI, and modularity.

To achieve this, we developed three approaches

designed to merge redundant clusters:

- Merging clusters sharing the three most frequent stems;
- Merging clusters with identical DBI values;
- Merging clusters using Hierarchical Clustering (HC) (Johnson, 1967).

The evaluation of these approaches, as well as the presentation of the results of the best-performing approach, are detailed in section 7.

6.1 Merging Clusters Sharing the Three Most Frequent Stems

The first approach to grouping redundant clusters is based on the analysis of the three most frequent stems. For this, the sentences of each cluster were tokenized, and the stems of the words were extracted before being sorted by descending frequency. Clusters sharing the three most frequent stems were merged, after removing stop words, whose stems were not considered in this operation.

6.2 Merging Clusters with Identical DBI Values

The second approach is based on the DBI. In Section 5, we used this measure to evaluate the quality of the clustering and observed that some clusters with the same DBI value deal with similar topics. Based on this, we hypothesized that if two clusters have exactly the same DBI value, they are likely to be close in terms of thematic content. Indeed, the DBI of a cluster reflects its proximity to the most similar cluster. Therefore, all clusters sharing an identical DBI value were merged.

6.3 Merging Clusters Using Hierarchical Clustering

The last proposed approach aims to reduce the number of clusters by applying HC to the clusters detected by the Louvain algorithm. This approach allows for merging clusters whose Euclidean distance is less than or equal to 0.32. This threshold was chosen to align with the cosine similarity of 0.68 used when constructing the initial clusters (with the Euclidean distance approximately equal to 1 - cosine similarity).

In this approach, each cluster is represented by its centroid, defined as the node closest to the other cluster nodes, according to the closeness centrality measure (Bavelas, 1950; Sabidussi, 1966).[8] This central node is therefore the one that is, on average, closest to the other elements of the cluster, making it a good representative of its structure. HC was then applied to these centroids, allowing the identification and merging of clusters deemed sufficiently close by the algorithm.

7 Semantic Analysis of *Cahiers Citoyens* through the Clusters

Evaluation of the Clustering after Merging Clusters:

The CHI, the DBI, and modularity must be recalculated after applying merging methods. Table 4 presents the new values as well as the number of clusters obtained after merging.

The merging of clusters sharing the three most frequent stems slightly improves the CHI as well as the DBI. However, this approach results in a minimal reduction in the number of clusters, decreasing to 13 clusters without PCA and 19 clusters with PCA. Modularity remains unchanged, both with and without PCA. This stability is explained by the slight reduction in the number of clusters, which limits the impact on the overall clustering structure. In summary, although this approach slightly enhances some quality indices, it does not lead to a significant reduction in the number of clusters.

The merging of clusters sharing an identical DBI also results in a modest reduction in the number of clusters, with 18 clusters without PCA and 14 with PCA. Although this approach slightly decreases the number of clusters, the quality of cluster separation deteriorates, as evidenced by the decline in the CHI in both cases (with and without PCA). Furthermore, the values of the DBI and modularity remain unchanged, indicating that this method does not significantly improve cluster compactness or modularity.

The best-performing approach is to merge clusters using HC, which reduces the number of clusters by approximately 38 when dimensionality is not reduced, and by 37 when PCA is applied. This

[8]The closeness centrality of a node is the inverse of the sum of the shortest path distances from this node to all other nodes in the network, indicating how close a node is to all others. A higher closeness centrality means the node is more central within the cluster.

Table 4: Table detailing the performance of each merging approach. **w/o PCA** = without PCA; **w PCA** = with PCA. **Bold values** represent the best performances without PCA. Underlined values represent the best performances with PCA.

Metric	CamemBERT-large + Louvain (before merging)		Merging clusters sharing the three most frequent stems		Merging clusters sharing an identical DBI value		Merging clusters using HC	
	w/o PCA	w PCA	w/o PCA	w PCA	w/o PCA	w PCA	w/o PCA	w PCA
# Clusters	2,398	2,449	2,386	2,430	2,380	2,435	2,360	2,412
CHI	7.25	7.76	7.29	7.77	7.06	7.49	**7.35**	7.83
DBI	1.26	1.22	**1.25**	1.21	1.26	1.22	**1.25**	1.21
Modularity	0.88	0.89	0.88	0.89	0.88	0.89	0.88	0.89

reduction improves the CHI, reflecting better cluster separation. Additionally, the merging leads to a decrease in the DBI, indicating that the clusters are now more compact, with points closer to their centroid and reduced intra-cluster dispersion. Although odularity remains largely unchanged, likely due to the limited reduction in cluster count, suggesting a stable overall graph structure. In summary, this approach enhances cluster cohesion while maintaining adequate separation.

Results of the Semantic Analysis:

The semantic analysis of the corpus was performed using the most effective combination, namely CamemBERT-large for sentence vector representation and Louvain for community detection, optimized by the most efficient merging approach: HC. Table 5 presents the 10 most compact clusters, which achieve the highest individual CHI values corresponding to each cluster. It is important to note that the individual CHI values are weighted by the number of sentences in each cluster, thereby highlighting clusters that are both compact and large in size. In this table, the topic of each cluster is identified through its central sentence, determined using the closeness centrality measure. The t-SNE (Maaten and Hinton, 2008) projection of these clusters is shown in Figure 1.

The analysis of the results in Table 5 reveals a strong concentration of discussions around key topics, with variations in the size and coherence of the groups, as reflected in their individual CHI.

Clusters 779 and 681 address the revaluation of retirement pensions and the reinstatement of the ISF (Solidarity Wealth Tax). They stand out due to their large size and high CHI, at 548.26 and 380.46, respectively. These results indicate a strong homogeneity within these clusters, with sentences closely related to widely shared economic

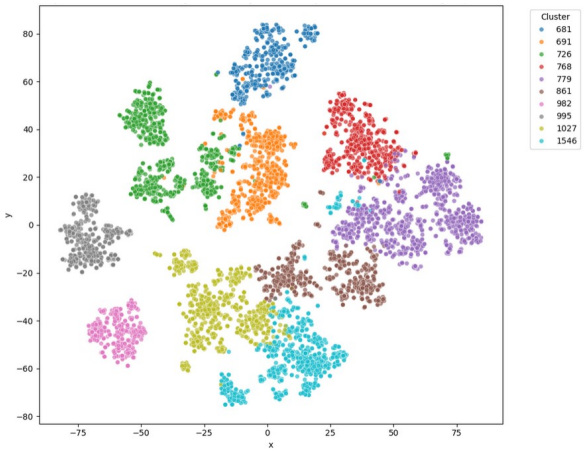

Figure 1: t-SNE Projection of CamemBERT-large Embeddings Reduced by PCA – Louvain Clustering (Top 10 Clusters)

and social concerns. The importance of pension and tax-related issues is reflected in the high number of sentences (1,066 for cluster 779 and 533 for cluster 681).

Cluster 1027, which focuses on reducing the number of deputies, also has a high CHI (330.88) and comprises 807 sentences. This topic, related to institutional reforms, demonstrates citizens' concerns about political representation and the functioning of institutions.

Other clusters, such as those addressing the removal of the CSG (General Social Contribution) tax on retirement pensions (cluster 768) and mandatory voting (cluster 982), focus on specific fiscal and democratic issues. Their respective sizes (588 and 378 sentences) reflect significant interest in these reforms, though to a lesser extent than broader economic and institutional topics.

Additional concerns also emerge, though in a more diverse manner. Tax reduction (cluster 726) and taxation of all incomes (cluster 691) highlight a broader debate on fiscal policies. The abolition

Table 5: Top 10 clusters with their CHI and central phrases

Cluster Index	# of Sentences	Individual CHI	Central Phrase (Translated)
779	1,066	548.26	*Revaluation of retirement pensions*
681	533	380.46	*Reinstatement of the ISF* (N/A: ISF is a Wealth Tax)
1027	807	330.88	*Reduction in the number of deputies*
768	588	297.72	*Remove the CSG on retirement pensions* (N/A: CSG is a Social Tax)
982	378	280.86	*Mandatory voting*
726	869	251.74	*Lower taxes*
691	708	246.43	*TAXATION Taxes on all incomes*
1546	796	236.83	*Abolition of privileges for politicians*
995	474	230.63	*Citizen consultation by referendum (RIC)*
861	635	212.52	*Salary increase*

of political privileges (cluster 1546) and citizen consultation via referendum (cluster 995) reflect a desire for systemic transformation and greater democratic participation. Finally, salary increases (cluster 861), though present, generate a more heterogeneous and less structured discussion.

In summary, these clusters illustrate main citizen concerns, with a predominance of economic and fiscal issues, followed closely by institutional reforms and citizen participation. The CHI, which remains relatively high in most clusters, indicates a clear separation between groups, confirming that concerns are structured around well-defined domains.

As discussed in section 2, the Roland Berger report (Berger and Bluenove, 2019) categorized citizen concerns into eight thematic areas. Our findings partially confirm these broader categories, while offering more granular insights into specific concerns raised by citizens. For instance, issues related to purchasing power and taxation emerge in our results through distinct clusters focusing on pension revaluation, the reinstatement of the ISF, or the removal of the CSG tax. Similarly, the demand for institutional reforms, present in Berger's category Democracy and Citizenship, is reflected in our clusters through topics such as reducing the number of deputies or implementing mandatory voting. Unlike the Berger synthesis, wwhich relied on opaque methods and broad predefined themes, our graph-based clustering approach reveals more specific, bottom-up topics that better capture the fine structure of citizens' discourse.

8 Conclusions and Prospects

This study presented a semantic analysis of a real-world corpus collected during a period of social unrest, aiming to understand citizens' concerns through the comparison of several combinations of language models with community detection algorithms. We found that the most effective combination for this purpose was CamemBERT-large for sentence representation paired with the Louvain algorithm for community detection. PCA played a beneficial role by enhancing cluster separation and reducing intra-cluster dispersion, as shown by the decrease in the DBI and the increase in the CHI. Given the large number of redundant clusters, a merging strategy was attempted: HC proved to be the most effective, grouping clusters on similar themes while improving compactness and homogeneity, thus strengthening clustering quality.

The cluster analysis revealed that citizen concerns focus mainly on economic, fiscal, and political issues. Recurring topics include pension reform (pension revaluation, removal of the CSG tax), taxation (restoration of the ISF, tax reduction), and institutional reforms (reduction in the number of deputies, removal of political privileges). Citizen participation, notably through citizens' initiative referendums (RIC) and compulsory voting, is also a major concern. Wage increases constitute another point of interest, though more diverse.

Looking forward, a key perspective is to extend this analysis to the full *CC* corpus using supercomputers to overcome computational limitations. This would provide a more precise overview of French citizens' concerns during the Yellow Vest crisis and just before the COVID lockdown. Moreover, a detailed analysis of raw texts within each cluster would refine semantic interpretations and improve understanding of the underlying themes.

References

Hadi Abdine, Yanzhu Guo, Virgile Rennard, and Michalis Vazirgiannis. 2022. Political communities on twitter: Case study of the 2022 french presidential election. *arXiv preprint*, arXiv:2210.05121.

Barot Aman, Bhamidi Shankar, and Souvik Dhara. 2021. Community detection using low-dimensional network embedding algorithms. *arXiv preprint*, arXiv:2106.10715.

Alex Bavelas. 1950. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Roland Berger and Cognito Bluenove. 2019. Analyse des contributions libres : Cahiers citoyens, courriers et emails, comptes-rendus des réunions d'initiative locale. Technical report.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Rafika Boutalbi, Mira Ait-Saada, Anastasiia Iurshina, Steffen Staab, and Mohamed Nadif. 2022. Tensor-based graph modularity for text data clustering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1924–1928, Madrid, Spain. ACM.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.

Mahfuzur Rahman Chowdhury, Intesur Ahmed, Farig Sadeque, and Muhammad Nur Yanhaona. 2023. Topic modeling using community detection on a word association graph. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Cyrile Delestre and Abibatou Amar. 2022. DistilCamemBERT : une distillation du modèle français CamemBERT. In *CAp (Conférence sur l'Apprentissage automatique)*, Vannes, France.

Catherine Dominguès and Laurence Jolivet. 2024. Analyse textométrique et spatialisée des Cahiers citoyens. In *JADT 2024Mots comptés, textes déchiffrésTome 1*, pages 309–318, Bruxelles, Belgique, Belgium. Presses universitaires de Louvain.

Jie Gao, Junping Du, Yingxia Shao, Ang Li, and Zeli Guan. 2023. Social network community detection based on textual content similarity and sentimental tendency. In *Artificial Intelligence – Third CAAI International Conference, CICAI 2023, Revised Selected Papers, Part II*, Fuzhou, China. Springer.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Sami Guembour. 2024. Analyse sémantique du corpus des cahiers citoyens. In *Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 17–27, Toulouse, France. ATALA and AFPC.

Jihui Han, Wei Li, Zhu Su, Longfeng Zhao, and Weibing Deng. 2016. Community detection by label propagation with compression of flow. *European Physical Journal B*, 89:193.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Journal of Open Source Software*, 5(51):2456.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *J. Ed. Psych.*, 24:417–441.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika*, 32:241–254.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Ortu Marco, Maurizio Romano, and Andrea Carta. 2024. Semi-supervised topic representation through sentiment analysis and semantic networks. *Big Data Research*, 37:100474.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Nathan Monnet and Maréchal Loïc. 2024. Clustering doc2vec output for topic-dimensionality reduction: A MITRE ATT&CK calibration. *arXiv preprint*, arXiv:2410.11573.

Matilde Monnier. 2023. L'analyse spatiale des cahiers citoyens appliquée au thème de l'écologie.

Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Sachin Pawar, Nitin Ramrakhiyani, Swapnil Hingmire, and Girish K Palshikar. 2018. Topics and label propagation: Best of both worlds for weakly supervised

text classification. In *European Conference on Information Retrieval (ECIR)*, pages 396–408, Cham. Springer.

Marjolaine Ray. 2023. Analyse sémantique et spatialisée des sentiments exprimés dans les Cahiers citoyens.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *https://arxiv.org/abs/1908.10084*.

Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy. 2013. Efficient community detection in large networks using content and links. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 519–528, New York, USA. ACM.

Gert Sabidussi. 1966. The centrality index of a graph. *Psychometrika*, 31:581–603.

Yuqi Tang, Wenyan Song, Caibo Zhou, Yue Zhu, Zheng Jianing, and Rong Wan. 2022. A consensus clustering-based label propagation method for classification of science & technology resources. In *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Kuala Lumpur, Malaysia.

Xiangpeng Wang, Michael Lucic, Hakim Ghazzai, and Yehia Massoud. 2021. Topic modeling and progression of american digital news media during the onset of the covid-19 pandemic. *IEEE Transactions on Technology and Society*.

Xiaojin Zhu and Zoubin Ghahramani. 2003. Learning from labeled and unlabeled data with label propagation.