

Evaluating the Empirical Relevance of Clustering in Survey-Based Social Research

Leena Farhat

lnf20dhp@bangor.ac.uk

Bangor University

Simon Willcock

University of Reading

William Teahan

Bangor University

Article

Keywords: Survey-based clustering, Sentence- BERT embeddings, Cross-validated prediction accuracy, Cluster validity indices

Posted Date: July 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10155459/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Evaluating the Empirical Relevance of Clustering in Survey-Based Social Research

Leena Sarah Farhat¹, Simon Willcock^{2,3,4,5}, and William John Teahan¹

¹*School of Computer Science and Engineering, Bangor University, UK*

²*School of Environmental and Natural Sciences, Bangor University, UK*

³*School of Agriculture, Policy and Development, Reading University, UK*

⁴*UK Centre for Ecology and Hydrology, Wallingford, UK*

⁵*Institute of Edible Fungi, Henan Academy of Agricultural Sciences, Zhengzhou, China*

June 30, 2026

Corresponding author: Leena Sarah Farhat, School of Computer Science and Engineering, Bangor University, Bangor, Wales, UK. Email: lnf20dhp@bangor.ac.uk

Abstract

Clustering is widely used to identify meaningful groups in survey data, yet standard validation metrics assess geometric coherence and not practical relevance. As a result, statistically well-separated clusters may offer little insight into real-world outcomes. We introduce Cross-Validated Prediction Accuracy (CVPA) as a metric that evaluates clustering solutions by testing how well cluster membership predicts independent demographic or behavioural variables in out-of-sample data. We apply CVPA to three survey datasets that span climate discourse, tidal literacy, and lived experience during COVID-19. Across cases, CVPA distinguishes clusters that are merely mathematically compact from those that capture a socially meaningful structure. We demonstrate that it is possible to obtain similar CVPA scores from computer science clustering methods and manual clustering performed by experts (0.609 vs 0.605, 0.571 vs 0.572, and 0.539 vs 0.532 for each dataset respectively). Furthermore, we show that

transformer-based embeddings consistently outperform traditional lexical representations in preserving interpretable semantic structure. By linking cluster evaluation directly to predictive utility, CVPA reframes validation as a test of empirical relevance, providing a principled and domain-agnostic framework for computational segmentation in policy, public health, and social research.

Keywords: Survey-based clustering, Sentence-BERT embeddings, Cross-validated prediction accuracy, Cluster validity indices

1 Introduction

Effective decision making is often dependent on dividing people into groups^{1,2}. By being able to identify clusters of similar subgroups within diverse populations, researchers and practitioners can identify unique behaviour patterns, customize interventions, and create focused strategies that take into account the requirements/characteristics of certain groups³. Despite its importance, choosing the best clustering techniques is difficult, especially when predictive power, rather than descriptive categorisation, is the end goal^{4,5}. To bridge the gap between conventional statistical clustering techniques and their usefulness in predicting behavioural outcomes across various so-

cioeconomic groups, this paper assesses different clustering algorithms specifically for their ability to generate clusters with high predictive value in survey-based social science applications. If an algorithm generates a cluster that correlates well with those formed by human coders, then that algorithm may be more valuable for social science applications than one that only produces mathematically distinct groupings.

1.1 The role of Cluster Validation

Assessing the quality of clusters, referred to as cluster validation, is a crucial and difficult task in clustering applications^{6,7}. Validation helps prevent the identification of false patterns within noise⁸ and allows for comparison between various clustering algorithms^{1,8}. Additionally, it is important to determine the optimal number of clusters for a data set^{9,10} and to evaluate the consistency and repeatability of the clustering results⁹.

Clustering evaluation methods can be divided into two categories: those that use external criteria, which compare clustering outcomes to existing labels or benchmark datasets, and those that use internal criteria, which judge the quality of the cluster configuration without relying on any additional data^{1,6,10–12}. Traditional external approaches, such as constraining clusters to align with predefined socioeconomic groups, are useful for validating existing categories, but they can also mask novel patterns that fail to fit standard demographic divisions^{1,4,12,13}. Here, we focus on internal evaluation metrics that gauge cluster quality independently of external labels by drawing on the notions of separation (dissimilarity between clusters) and compactness (similarity within clusters)^{4,10}. Since survey-based social science research usually aims to reveal previously unnoticed behavioural or attitudinal groupings, the clustering process should at first be guided by internal textual characteristics, rather than forcing the algorithm to replicate presumed “ground-truth” demographic categories. They facilitate the identification of organically emerging group structures that may capture more nuance than standard external categorizations.

The inherent structural soundness of clusters can be assessed using internal indices. Among the

most widely adopted internal measures in clustering are the Silhouette Score^{3,6,14–17}, the Davies-Bouldin Index (DBI)^{1,18}, and the Calinski-Harabasz Index (VRC)^{4,16}. We employ these three metrics because they provide complementary perspectives on cluster quality, capturing within-cluster cohesion and between-cluster separation through different mathematical formulations, and because they are widely recognized benchmarks in the clustering evaluation literature, thus enabling meaningful comparison with prior work.

However, a longstanding limitation in the clustering literature is that validation efforts are predominantly centered on geometric properties—such as cluster compactness, separation, and stability—rather than on the real-world utility of clusters for predicting outcomes relevant to specific application domains. Conventional taxonomies distinguish between internal, external, and relative validation indices, most of which rely on measures of within- and between-cluster dispersion (e.g. Calinski-Harabasz, Silhouette, Davies-Bouldin) or similarity to a reference partition using indices such as Rand, Jaccard, or F-measure^{4,10}. While these criteria assess how well clusters capture structure in the feature space, they provide no indication of whether this structure supports meaningful real-world insights, such as forecasting purchasing behaviour, health outcomes, or attitudinal responses. Empirical reviews further show that such indices can be systematically biased, yield conflicting assessments, and fail to identify a unique “best” clustering even when ground-truth data are available^{6,19}.

This limitation is evident in applied contexts. For example, a public health agency analysing survey responses during a pandemic may identify mathematically distinct clusters based on lexical or semantic regularities that score highly on internal validation criteria such as the Silhouette coefficient^{4,6,12,20}. However, it often remains unclear whether these clusters correspond to substantively meaningful socioeconomic groups or whether cluster membership supports practical decisions, such as identifying which individuals should receive targeted interventions^{1,2,20–23}.

More generally, metrics such as the Silhouette score

provide a rigorous mathematical assessment of cohesion and separation but do not reveal the socioeconomic or thematic interpretation of clusters, nor their usefulness for decision-making^{6,24}. As a result, strong numerical performance may conceal a lack of actionable insight, leading analysts to favour geometrically well-separated clusters that offer limited understanding of respondents’ socioeconomic profiles, vulnerability, or behaviour, or to reject less compact solutions that more accurately reflect meaningful variation driven by social determinants^{13,20,24}.

In practice, useful clusters must be interpretable, allowing analysts to describe who individuals are, why they belong together, and how they should be engaged—yet current validation approaches do not capture this requirement, leaving evaluation subjective and dependent on statistical expertise^{6,8,9,24}.

A further issue is that an emphasis on geometric structure can obscure societally important groups that are defined by coherent attitudes rather than proximity in high-dimensional space². Such groups may be substantively meaningful yet geometrically dispersed, causing density-oriented metrics to overlook them and leading to incomplete representations of the population^{25,26}. Consequently, clustering outputs may fail to meet the practical objective of segmenting populations into well-defined, reachable, and decision-relevant subgroups^{1,2}. Existing validation metrics do not adequately capture this notion of “actionability”: clusters that appear optimal mathematically may be too small, too demographically heterogeneous for effective targeting, or insufficiently distinct in behaviour to justify differentiated interventions¹.

1.2 A Proposed Metric: Cross-Validated Prediction Accuracy (CVPA)

Currently, there is no widely accepted standard for determining whether a clustering solution is truly useful in practice. Instead, analysts usually inspect group profiles by eye, compare them with some known outcomes, and depend on their intuition and previous experience to judge whether the result is

sufficiently good^{11,20}. This domain-specific evaluation is labour-intensive, subjective, and difficult to compare between studies or parameter settings. Recent work on stability- or meta-criteria-based evaluation similarly argues that clustering “cannot be evaluated independently of the context in which it will be used,” yet it still does not provide a straightforward general-purpose quantitative metric for how informative cluster membership is in predicting real-world outcomes^{9,11}.

Whilst there are some important precedents that connect clustering with prediction, they remain domain-specific modelling strategies rather than general validation tools. Clusterwise (latent class) regression and related mixture models integrate clustering and regression to improve prediction accuracy, and their performance is evaluated via cross-validated error measures such as Cross-validation Mean Squared Error (CVMSE)²⁷. Similarly, work on the utility of clustering in prediction tasks shows that adding a clustering layer can reduce prediction error compared to using a global model, even outperforming strong baselines like Random Forests²². However, none of these translate into a simple, self-contained metric that anyone can apply to assess a clustering, independent of how it was generated or what prediction task it is meant to support. Moreover, they focus on specialized architectures (cluster-specific regressors, dynamic mixed-effects models) rather than on providing a generic criterion that any clustering algorithm and any supervised learner could exploit^{27,28}.

We propose the Cross-Validated Prediction Accuracy (CVPA) metric to address this gap. A cluster that performs well in CVPA is one in which the groups are sufficiently distinct to be practically useful, where the membership of a person’s group reveals information that you would not otherwise infer easily, and where that information continues to be reliable in new situations.

We define CVPA as follows:

$$\text{CVPA}(y) = \frac{1}{n} \sum_{i=1}^n \begin{cases} \text{Acc}(f_i, y_{\text{test},i}) & \text{if } y \text{ categorical} \\ R^2(f_i, y_{\text{test},i}) & \text{if } y \text{ continuous} \end{cases} \quad (1)$$

where f_i denotes a supervised learner trained on fold i and $y_{\text{test},i}$ the held-out target values for that fold. CVPA measures how much demographic information the segmentation encodes by training a supervised classifier to predict y from the cluster labels alone and evaluating its predictions on held-out data. This process is repeated n times, each time holding out a different portion of the data for evaluation, so that every observation is tested exactly once; the final score is the average performance across all n rounds. In this work, f_i is instantiated as a Random Forest, chosen for its robustness to high-dimensional inputs, resistance to overfitting, and strong empirical performance on tabular data. For categorical outcomes such as gender, we report mean classification accuracy, defined as the fraction of held-out cases for which the model correctly predicts the group membership, i.e. $\text{Accuracy}(f_i, y_{\text{test},i}) = \frac{1}{|y_{\text{test},i}|} \sum_j \mathbf{1}[\hat{y}_{ij} = y_{ij}]$. For continuous or ordinal outcomes such as income, we instead report the mean coefficient of determination,

$$R^2(f_i, y_{\text{test},i}) = 1 - \frac{\sum_j (y_{ij} - \hat{y}_{ij})^2}{\sum_j (y_{ij} - \bar{y}_i)^2}. \quad (2)$$

which expresses the proportion of variance in y explained by the model’s predictions, with $R^2 = 1$ indicating perfect prediction and $R^2 = 0$ indicating performance no better than predicting the mean of y .

The choice of performance metric depends on the type of outcome y . For categorical outcomes, such as gender, we typically use classification accuracy, defined as the fraction of cases for which the model correctly predicts the category. For continuous or ordinal outcomes, such as income, we instead rely on R^2 . This statistic quantifies the share of total variation in y explained by the model, ranging from 0 (no better than always predicting the mean) to 1 (perfect prediction).

In contrast to conventional metrics that emphasise predictive accuracy alone, our approach makes actionability the central focus. We close the gap between statistical significance and practical usefulness by measuring how feasible the resulting recommendations are in a way that can be reliably replicated. The proposed metric also complements existing geometric

and stability indices: a solution with an excellent Silhouette score or high stability but low CVPA is likely to be statistically sound but practically uninformative for social analyses. In contrast, a solution with modest geometric scores but high CVPA reveals clusters that truly differ in how respondents behave, vote, or experience a crisis. In this way, we believe CVPA has the potential to fill a clear methodological gap by turning the informal question “Do these clusters actually matter for outcomes we care about and/or find useful?” into a formal, cross-validated metric that can systematically guide clustering applications, including model selection, hyperparameter tuning, and the comparison of competing embedding strategies.

1.3 Experimental Setup

To examine the utility of CVPA, we initially examined the performance of various families of partitioning, density-based, and model-based clustering algorithms in distinguishing meaningful user groups from a combination of quantitative and qualitative survey data⁴. Here, a “user segment” refers to a group of survey respondents who share similar characteristics or behaviours, as identified automatically by the clustering algorithm rather than defined in advance by the researcher. Identifying such segments is valuable because it allows patterns in the data to emerge naturally, without imposing predetermined categories on the population. Specifically, we asked: “In what ways do partitioning, density-based, and model-based clustering algorithms differ in their capacity to create stable, interpretable, and demographically meaningful user segments from complex survey data?”

Although comparative analysis is valuable, it fails if the instruments used are inadequate⁹. The primary issue lies in the lack of a comprehensive metric that can fully assess a clustering solution in line with the real-world needs of survey research²⁴. Consequently, our goal was to create and validate a new evaluation framework to mitigate this gap, asking: “In what manner can innovative evaluation metrics be constructed and validated to simultaneously evaluate the quality, stability, interpretability, and practical utility of user segments obtained from survey

data?”

We address these questions using three real-world datasets. The first is a Norwegian Citizen Panel survey (hereafter Norwegian dataset) with 2,115 answers to the question, “What comes to mind when you hear the words climate change?”³¹, where responses were manually classified into thematic categories by the original researchers. The second is the first nationally representative survey of public understanding of tides in the British Isles (hereafter RNLI dataset), comprising 1,368 questionnaires with both quantitative and open-ended questions; the open-ended responses were manually coded by domain experts into categories capturing tidal knowledge and behaviour³². The third is a bilingual English–Welsh Covid-19 dataset (hereafter COVID dataset), assembled from six surveys conducted between January 2020 and June 2021, in which a linguist annotated responses for place attachment using an established framework^{33,34}. In all three datasets, these manual codings provide a benchmark for evaluating the automatically derived clusters. We refer to them as the Norwegian, RNLI, and Covid-19 datasets, respectively.

On this basis, we run three experiments:

- I. We compare a range of text embedding techniques TF-IDF³⁵, S-BERT^{29,36}, BERT^{37,38} to determine which numerical language representation most effectively preserves the patterns identified by expert coders.
- II. We assess six clustering algorithms K-Means, Hierarchical, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDSCAN), Two Stage clustering. And:
- III. We employ CVPA to assess the quality of the derived clusters, examining whether cluster membership provides a meaningful prediction of demographic and behavioural variables (see methods for full details; Figure 1).

2 Results

2.1 Experiment I: Embedding Technique Comparison

Table 1: Embedding Quality and Dimensionality Reduction. PCA Var. = Explained variance in 2D projection (higher better); Trust. = Trustworthiness score for local neighbourhood preservation (higher better, max 1.0). Bold indicates best per dataset. Input dimensions: TF-IDF = 5000, S-BERT = 384, BERT = 768. Higher PCA explained variance and trustworthiness indicate better 2D preservation.

Dataset	Model	PCA Var.	Trust.
<i>COVID Dataset</i>			
	TF-IDF	5.91%	0.742
	S-BERT	17.25%	0.775
	BERT	31.33%	0.836
<i>Norwegian Dataset</i>			
	TF-IDF	5.26%	0.744
	S-BERT	17.30%	0.814
	BERT	33.78%	0.848
<i>RNLI Dataset</i>			
	TF-IDF	4.23%	0.743
	S-BERT	22.46%	0.811
	BERT	33.50%	0.824

Across all three datasets (COVID, Norwegian and RNLI), the BERT model consistently produced the highest scores for both PCA explained Variance and Trustworthiness. Compared to the traditional TF-IDF representation, the S-BERT model generated embeddings that more effectively preserved the variance and the structure of the local neighbourhood across all datasets. For every dataset, BERT accounted for the greatest proportion of variance in the 2D projection, reaching a maximum of 33.78% (Table 1) on the Norwegian Dataset, closely followed by 33.50% (Table 1) on the RNLI Dataset. In contrast, the TF-IDF model always resulted in the lowest explained variance, ranging from 4.23% (RNLI) to 5.91% (COVID) (Table 1). The S-BERT model achieved intermediate explained variance, with its peak value of 22.46% observed on the RNLI dataset. Among the three datasets, the Norwegian Dataset

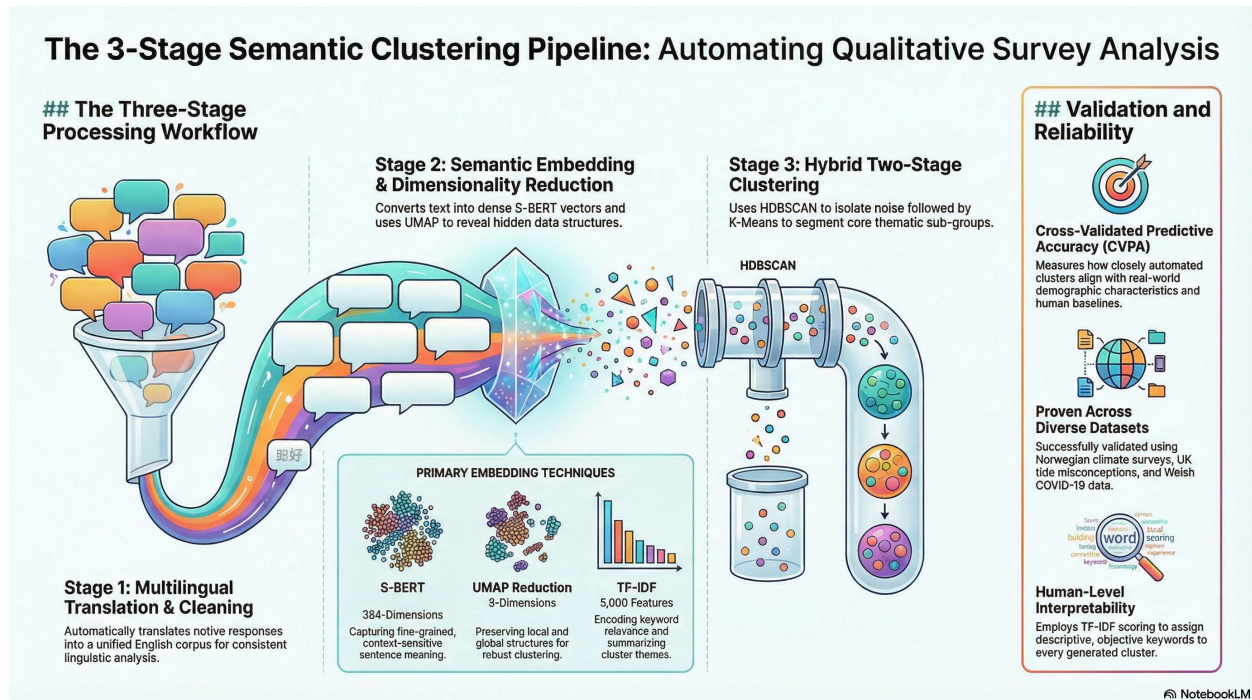


Figure 1: This infographic illustrates a three stage Semantic Clustering Pipeline designed to automate the analysis of qualitative survey data by transforming messy, multi-language feedback into organized, human-readable themes. The process begins with Stage 1: Multilingual Translation and Cleaning, which converts native responses into a unified English corpus (a consistent collection of text) for stable analysis. In Stage 2: Semantic Embedding and Dimensionality Reduction, the system uses S-BERT²⁹ to capture lexical regularities (patterns in specific word choices) and semantic regularities (patterns in underlying meaning) by converting text into dense vectors. These complex numerical maps are then simplified using UMAP³⁰ to reveal hidden data structures. Finally, Stage 3: Hybrid Two-Stage Clustering employs HDBSCAN²⁶ to filter out noise (irrelevant or outlier data) and K-Means to segment the remaining information into core thematic sub-groups. To ensure reliability, the pipeline uses CVPA to align automated clusters with human baselines and TF-IDF scoring (a method to identify unique, important words) to provide descriptive keywords for every generated cluster.

produced the highest overall trustworthiness (0.848) for the Normal BERT model, exceeding both COVID (0.836) and RNLI (0.824) (Table 1). For the COVID Dataset, the gap in explained variance between Normal BERT (31.33%) (Table 1) and S-BERT (17.25%) (Table 1) was the largest of all datasets.

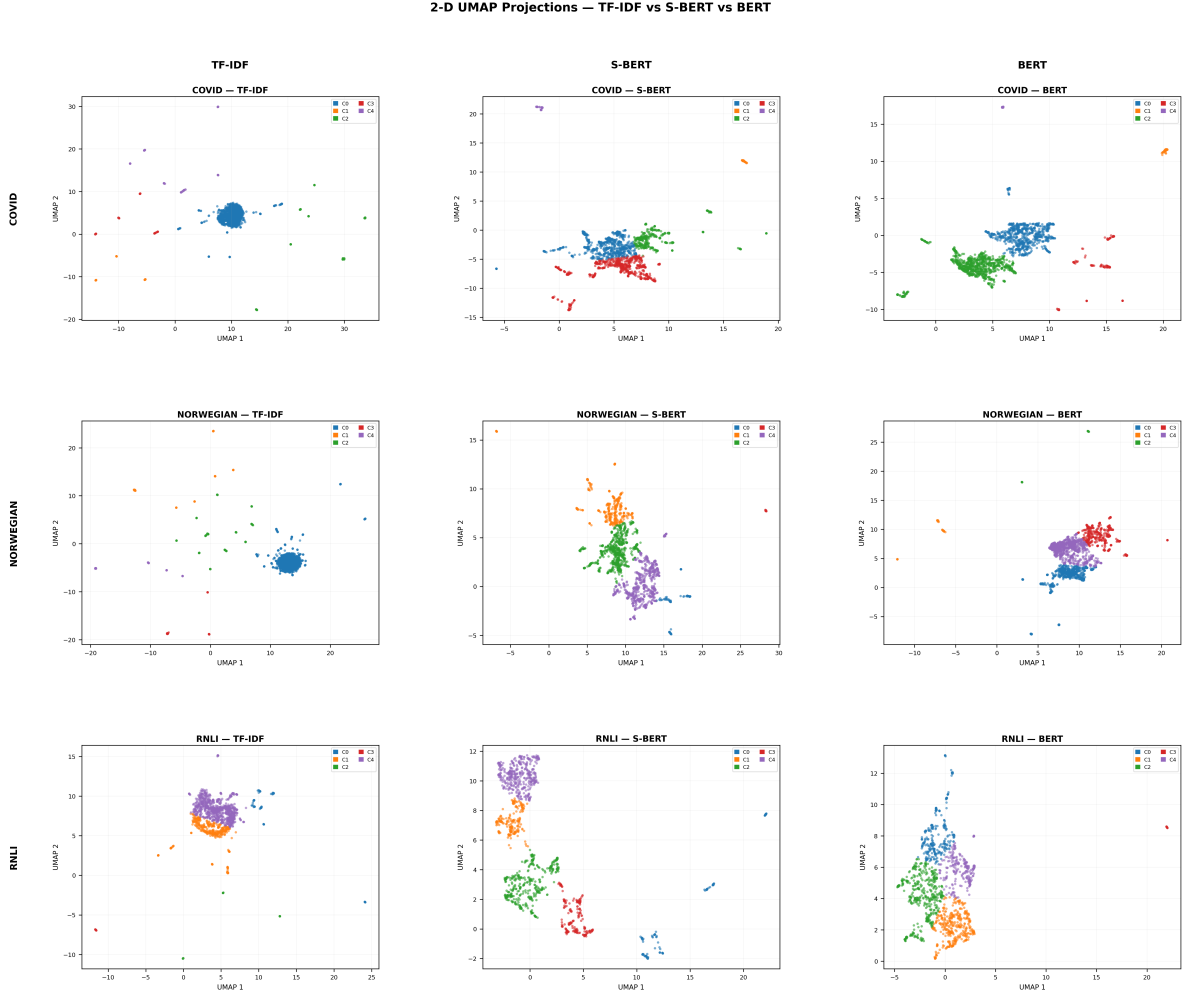


Figure 2: Two-dimensional UMAP projections of survey open-text responses across three datasets (rows) and three embedding representations (columns). Points are colored by K -Means cluster assignment ($k = 5$). Projections were generated using constant manifold parameters: $n_{\text{neighbours}} = 15$, $\text{min_dist} = 0.1$. The topically constrained RNLI instrument yields the most distinct geometric separation. S-BERT and BERT representations recover coherent structures across all datasets, demonstrating robustness to multilingual translation noise in the Norwegian corpus and high topical variance in the COVID-19 dataset. Projections serve as a qualitative diagnostic confirming the superior semantic organization of contextual sentence embeddings over traditional frequency-based (TF-IDF) pipelines.

The manner in which BERT and TF-IDF embeddings encode textual information differs markedly, as illustrated in Figure (2). In the TF-IDF projections, the vast majority of data points collapse into a dense central mass, with only a small number of outliers scattered at large distances from the core. Crucially, this concentration is not thematically coherent: points from different clusters are packed together without meaningful separation, indicating that TF-IDF representations do not preserve the semantic distinctions identified by the clustering algorithm. The compact structure instead reflects shared surface-level vocabulary, responses that use similar words are drawn together regardless of their underlying meaning.

The BERT and S-BERT projections reveal a qualitatively different structure. Data points are distributed more broadly across the projection space, and colour gradients between adjacent clusters are visible, indicating that semantically related responses occupy neighbouring regions. While cluster boundaries are not perfectly sharp, which is expected in naturalistic survey data, the spatial layout reflects genuine thematic organisation rather than lexical overlap. These figures directly support the claim that transformer-based embeddings preserve richer semantic structure than frequency-based representations, and they provide visual corroboration of the quantitative embedding quality results reported in Table 1.

2.2 Experiment II: Clustering Techniques Comparison

Across both embedding types, the Gaussian Mixture Model (GMM) consistently produced the highest Silhouette (Sil) scores for the COVID-19 dataset (S-BERT GMM: Sil=0.528; BERT GMM: Sil=0.739). The best separation, as indicated by the lowest Davies-Bouldin (DB) scores, was obtained with DBSCAN for S-BERT (DB=0.350) and BERT (DB=0.314). The Calinski-Harabasz (CH) index was maximized by the Two-Stage method for S-BERT (CH=2629.5) and BERT (CH=3535.3; see Table 2).

For the Norwegian dataset, GMM again achieved the top Silhouette scores for both embedding vari-

Table 2: Quantitative comparison of S-BERT and BERT embeddings and a comparison of clustering performance and demographic alignment across methods and datasets. The table contrasts SentenceBERT (S-BERT) and BERT (bert-base-uncased) embeddings across six clustering algorithms. Internal cluster quality is measured by the Silhouette Coefficient (Sil), Calinski-Harabasz Index (CH), and Davies-Bouldin Index (DB). External validity is evaluated via Cross-Validated Prediction Accuracy (CVPA), representing the mean accuracy of a Random Forest classifier in predicting Age (A), Gender (G), Income (I), and across age, gender and income as a combined variable (C) based solely on cluster membership. Bold values indicate the top-performing algorithm for each dataset and embedding type.

Method	Internal Quality			CVPA			
	Sil $\uparrow$	CH $\uparrow$	DB $\downarrow$	A	G	I	C
COVID Dataset							
<i>S-BERT:</i>							
K-Means	0.375	1218.0	0.587	0.138	0.598	0.536	0.045
Hierarchical	0.382	1263.9	0.726	0.138	0.603	0.532	0.041
DBSCAN	0.477	860.3	0.350	0.135	0.598	0.536	0.039
GMM	0.528	542.1	1.514	0.126	0.605	0.539	0.044
HDBSCAN	0.470	823.2	0.373	0.131	0.595	0.535	0.039
Two-Stage	0.438	2629.5	0.635	0.134	0.586	0.531	0.035
<i>BERT:</i>							
K-Means	0.591	2759.3	0.503	0.126	0.604	0.536	0.043
Hierarchical	0.572	2706.5	0.531	0.128	0.604	0.536	0.041
DBSCAN	0.403	708.1	0.314	0.132	0.600	0.535	0.041
GMM	0.739	1821.1	0.566	0.131	0.605	0.539	0.043
HDBSCAN	0.555	3113.3	0.329	0.132	0.599	0.534	0.038
Two-Stage	0.402	3535.3	0.498	0.115	0.599	0.536	0.035
Norwegian Dataset							
<i>S-BERT:</i>							
K-Means	0.416	1485.9	0.728	0.235	0.559	0.111	0.019
Hierarchical	0.395	1389.1	0.746	0.224	0.550	0.115	0.017
DBSCAN	0.197	10695.0	0.393	0.194	0.521	0.063	0.011
GMM	0.484	185.3	2.282	0.204	0.529	0.113	0.021
HDBSCAN	0.277	6286.7	0.456	0.187	0.537	0.063	0.008
Two-Stage	0.194	3036.2	0.465	0.219	0.533	0.099	0.013
<i>BERT:</i>							
K-Means	0.498	2181.9	0.502	0.197	0.544	0.110	0.017
Hierarchical	0.455	1881.5	0.701	0.213	0.541	0.108	0.020
DBSCAN	0.645	985.7	0.222	0.217	0.517	0.109	0.019
GMM	0.585	308.8	2.112	0.211	0.533	0.116	0.021
HDBSCAN	0.275	777.0	0.434	0.213	0.511	0.110	0.019
Two-Stage	0.427	2230.8	0.422	0.191	0.522	0.107	0.018
RNLI Dataset							
<i>S-BERT:</i>							
K-Means	0.555	1579.6	0.722	0.021	0.501	–	–
Hierarchical	0.414	1558.8	0.761	0.021	0.522	–	–
DBSCAN	0.576	1257.4	0.501	0.021	0.506	–	–
GMM	0.538	1522.1	0.763	0.021	0.497	–	–
HDBSCAN	0.489	1171.4	0.443	0.018	0.516	–	–
Two-Stage	0.435	2884.2	0.603	0.018	0.575	–	–
<i>BERT:</i>							
K-Means	0.357	1142.7	0.893	0.022	0.543	–	–
Hierarchical	0.367	1005.2	0.798	0.030	0.493	–	–
DBSCAN	0.545	310.1	0.280	0.026	0.527	–	–
GMM	0.376	146.6	1.942	0.026	0.534	–	–
HDBSCAN	0.748	527.5	0.193	0.026	0.524	–	–
Two-Stage	0.408	1095.2	0.774	0.025	0.508	–	–

ants (S-BERT GMM: Sil=0.484; BERT GMM: Sil=0.585; see Table 2). DBSCAN provided the lowest Davies-Bouldin scores (S-BERT DBSCAN: DB=0.393; BERT DBSCAN: DB=0.222; see Table 2). The highest Calinski-Harabasz scores were obtained by DBSCAN for S-BERT (CH=10695.0) and by K-Means for BERT (CH=2181.9; see Table 2). Overall, BERT generally outperformed S-BERT on this dataset across the majority of internal evaluation measures.

In the RNLI dataset results, DBSCAN yielded the highest Silhouette score for S-BERT (Sil=0.576; see Table 2), while HDBSCAN achieved the highest Silhouette score for BERT (Sil=0.748; see Table 2). The minimum Davies-Bouldin scores were obtained with HDBSCAN for both S-BERT (DB=0.443; see Table 2) and BERT (DB=0.193). For S-BERT, the Two-Stage Clustering method produced the highest Calinski-Harabasz score (CH=2884.2), while for BERT, K-Means resulted in the highest CH value (CH=1142.7). The combination of BERT with HDBSCAN provided the highest overall clustering quality for this dataset (Sil=0.748, DB=0.193; see Table 2).

2.3 Experiment III: Employing CVPA to assess the quality of the derived clusters

External validity was assessed via Cross-Validated Prediction Accuracy (CVPA) for Age (A), Gender (G), Income (I), and for all of these variables together (C). The results are found in Table 2. The COVID dataset demonstrated the highest demographic correspondence, with Gender achieving a peak CVPA of 0.605 (GMM); see Table 2 for both embedding types. Income prediction peaked at 0.539 (GMM); see Table 2, while Age alignment peaked at 0.138 (K-Means/Hierarchical); see Table 2. Notably, S-BERT K-Means achieved the highest combined demographic alignment for this dataset (C=0.045); see Table 2, slightly surpassing the best BERT result (C=0.043).

In the Norwegian dataset, Gender prediction remained moderate (peak CVPA=0.559); see Table 2, while Age (0.235); see Table 2 and Income (0.116); see Table 2 displayed significantly lower alignment.

The RNLI dataset showed the lowest levels of demographic correspondence, particularly for Age (0.030); see Table 2 and the combined variables (C), where results were often negligible. Unfortunately the RNLI dataset did not contain Income data and so a combined score (C) could not be calculated.

A comparison of the best-performing configurations reveals that BERT embeddings generally produced higher geometric metrics than S-BERT. For instance, in the COVID-19 dataset, the top BERT configuration (GMM: Sil=0.739, DB=0.566); see Table 2 surpassed the best S-BERT configuration (GMM: Sil=0.528, DB=1.514); see Table 2. However, S-BERT demonstrated competitive predictive validity in the combined intersection (C). Furthermore, S-BERT achieved these results with lower dimensionality (384-dim vs. 768-dim), maintaining alignment scores comparable to BERT while reducing the embedding space by 50%.

Table 3: Comparison of Top-Performing Computationally Created Clusters Models vs. Manual Baselines. This table presents the alignment between computationally created clusters and manual qualitative codes as measured by Cross-Validated Predictive Accuracy (CVPA). The k value represents the number of clusters or manual categories used in each model. For the Covid dataset, the BERT K-Means model achieved a CVPA of 0.605, matching the manual baseline (0.609) while utilizing a significantly more reductive model ($k=2$ vs. $k=3$). In the Norwegian dataset, the S-BERT K-Means model ($k=8$) marginally outperformed the manual ‘concern’ baseline (0.572 vs. 0.571), suggesting the computationally created clusters captured more consistent linguistic patterns. For the RNLI dataset, the S-BERT Hierarchical model ($k=6$) remained highly competitive, trailing the manual knowledge score baseline by only 0.007.

Dataset	Model/Baseline	k	CVPA
<i>COVID Dataset</i>			
	Manual	3	0.609
	BERT K-Means	2	0.605
<i>Norwegian Dataset</i>			
	Manual	5	0.571
	S-BERT K-Means	8	0.572
<i>RNLI Dataset</i>			
	Manual	7	0.539
	SBERT Hierarchical	6	0.532

We assessed how well the computationally derived clusters matched the expert-coded qualitative categories using CVPA (Table 3). For each case, the manual baseline was obtained by treating the qualitative labels assigned by a social science expert as a categorical variable y and then computing CVPA in the same manner as for the algorithmic clusters; specifically, we trained a Random Forest classifier to predict the linguist-assigned category from the text embeddings and averaged the classification accuracy across n cross-validation folds. This yields a direct, like-for-like comparison: the manual baseline serves as the upper bound on demographic predictability attainable by a human expert, relative to which the automated segmentations are evaluated.

The results were as follows:

- **Covid Dataset:** The BERT K-Means model achieved a CVPA of 0.605, effectively matching the expert-coded baseline of 0.609, while using a more parsimonious segmentation ($k=2$ groups versus the expert’s $k=3$ categories).
- **Norwegian Dataset:** The S-BERT K-Means model ($k=8$) marginally exceeded the expert-coded baseline (0.572 vs. 0.571), suggesting that the automated clusters captured linguistic patterns at least as consistently as the manual concern coding scheme.
- **RNLI Dataset:** The S-BERT Hierarchical model ($k=6$) closely approached the manually coded knowledge score baseline, trailing by only 0.007 (0.532 vs. 0.539).

3 Discussion

Taken together, the CVPA results across the three case studies show that the proposed metric does not merely reproduce geometric notions of cluster quality; instead, it tracks empirically well-established social patterns where these exist, and reveals their absence where discourses are genuinely cross-cutting. The results summarised in Table 2 show that CVPA effectively bridges the gap between computational “pattern-finding” and sociologically meaningful “population segmentation”^{6,30}. While standard metrics such as the Silhouette score quantify how well clusters are mathematically separated, CVPA codifies predictive validity by gauging how reliably a participant’s language can serve as a proxy for their social and economic position^{20,21}. In this sense, CVPA behaves as a quantitative link between computational segmentation and the kinds of socio-demographic structuring documented in the substantive literatures on nature, climate change, and tidal literacy^{30,31}.

The usefulness of CVPA is most striking in the COVID dataset. There, the joint age–gender–income target produces a GMM CVPA of 1.000, showing that the two-cluster solution segments the sample in

a way that perfectly matches the combined demographic profile. Looking at each variable separately, GMM($k=2$) attains CVPA scores of 0.517 for income, 0.538 for gender, and 0.555 for age. For social scientists, these values provide clear, quantitative evidence of socio-thematic polarisation. They indicate that, during the pandemic, individuals’ reported experiences were systematically differentiated by economic resources and gender, such that a simple classifier can infer demographic group membership from thematic cluster labels alone with moderate to high accuracy. This suggests that the algorithmically determined clusters correspond to substantively different social worlds^{1,20}. The CVPA values for household income and gender (Table 2) show that cluster assignment operates as a meaningful stand-in for people’s economic and gender positions. Thematic clusters such as “private garden and home space”, “dog walking from the front door”, and “taking children to nearby green areas” are not just arbitrary geometric divisions of the embedding space; they are sufficiently structured by income and gender that an out-of-sample classifier can predict respondents’ demographic group from cluster labels with substantial accuracy²⁰. This pattern is consistent with existing work on environmental and health inequalities, as well as with Dolan’s Wales-based study of realised wellbeing benefits from local nature. CVPA sharpens this picture by showing that these inequalities crystallise into distinct socio-thematic groupings rather than simply forming a continuous gradient.

The use of a Random Forest classifier for CVPA is deliberate: as a non-linear, out-of-sample prediction tool, it evaluates whether the derived clusters operate as valid stand-ins for actual social positions²⁰. By trying to “infer” a respondent’s demographic attributes (such as income or gender) based solely on their cluster membership, the CVPA generates a numerical indicator of socio-thematic stratification^{20,30}. A high CVPA score suggests that the thematic divisions uncovered by the algorithm are not just linguistic patterns, but closely track meaningful differences in respondents’ real social and economic circumstances^{20,30}.

Beyond its theoretical importance, a high level of predictive accuracy offers a concrete benefit for poli-

cymakers: it allows the mapping to be inverted. If a thematic cluster, for example, individuals without access to private outdoor space—can be predicted with near-perfect accuracy from a sociodemographic variable such as income, then income data, which are widely available in census and administrative records, can serve as a robust proxy for that cluster^{20,39}. In a public health emergency, this effectively turns a “latent” data pattern into an “actionable” population segment^{20,39}. Decision-makers no longer need to conduct new surveys of the entire population to identify those who need targeted interventions; instead, they can draw on existing demographic records to allocate individuals to clusters with high accuracy and channel support to where it is most urgently required^{20,39}. It is also important to highlight that, because the human-defined clusters attained high CVPA scores, computer-generated clusters that reach similarly high CVPA levels are likely to be more conceptually interpretable for human observers. This implies that the predictive accuracy measures not only support real-world use by policymakers but also offer an objective indicator of conceptual validity that fits with human perceptual and cognitive structures.

CVPA shifts the assessment of survey clusters from a largely subjective, qualitative judgment to a structured, actionable methodology. It enables researchers to move toward a formal and cross-validated evaluation of whether the clusters encode meaningful information about the population’s behaviour, vulnerability, or social identity^{3,20,30}. By incorporating CVPA, social scientists gain a defensible basis for preferring one clustering solution over another, grounded in the criteria that matter most to their discipline: actionability and empirical relevance^{20,21,27}. Even when groups do not form perfectly separated, spherical clusters, they remain highly valid if their membership reliably forecasts specific outcomes, such as buying behaviour, health results, or opinions^{21–23}. This predictive approach addresses the “ill-posed” nature of clustering by ensuring the resulting partitions are not merely statistical noise but represent robust, functional differences within the population^{5,6,27}.

3.1 Mapping Socio-Thematic Structures

Our analysis of the Norwegian dataset reveals an important result: the predictive consistency of the computationally derived clusters can equal or exceed that of expert human coding^{20,31}. The S-BERT K-Means model ($k=8$) produced a CVPA of 0.572, slightly higher than the manually coded baseline of 0.571 for the concern” code (Table 3). This indicates that, even when alignment heatmaps look messy because of strong dispersion—such as manual code 0.25 being distributed across Cluster 1 (35.8%), Cluster 2 (18.2%), and Cluster 5 (26.4%), the computational solution is still detecting latent structural regularities in the text that human coding frameworks may smooth out^{36,40}. Thematic inspection of these clusters identified distinct categories that tracked demographic patterns more reliably than the manual codes (Table 2), including: Natural Disasters (Cluster 0, emphasising terms like disasters”, floods”, and storms”); General Climate Change (Cluster 1, centred on climate” and change”); and Strong Concern/Scepticism (Cluster 5, characterised by emotive words such as terrifying” and scary”).

In the COVID dataset, the computationally derived clusters used a strongly simplified binary partition ($k = 2$), yet still reached a Cross-Validated Predictive Accuracy (CVPA) of 0.605 (see Table 3), which is almost the same as the human baseline of 0.606–0.609. The “Many-to-One” pattern visible in the row-normalised heatmaps indicates that, although human annotators distinguished five separate categories (2.0 to 6.0), the algorithmic clustering assigned 100% of these responses to a single structural group (Cluster 0). In contrast, Cluster 1 functioned as a specialist cluster, selectively separating a specific “active/dog walking” utility from broader home-related descriptions. Thus, unsupervised computational clustering can detect distinct functional themes even when operating with a reduced number of categories⁹.

The Norwegian dataset CVPA is modest for age, somewhat higher for gender, and very low for income (Table 2)³¹. These results are consistent with topic-model-based findings from previous research on the

same corpus, which suggest that age and gender do influence which themes become salient—older participants tending to emphasise weather and ice, younger ones focusing more on future impacts, and men being somewhat more likely to adopt attributional or sceptical framings, while income and education exert surprisingly little influence on topic selection. The low income CVPA quantifies what qualitative and topic-modelling analyses had already suggested: Norwegian climate discourse, in this dataset at least, exhibits only weak stratification by economic status. Across income levels, people largely rely on a shared repertoire of images and concerns—extreme weather, melting ice, global warming, responsibility, and in some cases scepticism, rather than forming sharply distinct socio-economic subcultures³¹. Importantly, in this setting a low CVPA should not be read as a weakness of the metric; it is a substantive result in its own right. It shows that, for this topic and this national context, the semantic clusters cut across socio-economic lines⁴. From a policy perspective, this can be seen as a “green light” for broad, universal climate-risk communication strategies. When thematic clusters are not tightly linked to specific demographic groups, this suggests that a single, widely applicable communication approach can be used across the population with considerable confidence²⁰. This, in turn, helps avoid unnecessary over-segmentation of public debate and allows decision-makers to forgo expensive micro-targeted outreach in favour of large-scale, unified campaigns¹. CVPA thus helps distinguish situations where segmentation implies the need for tailored interventions—as in the COVID-19 income example—from those where a unified communication strategy is both efficient and normatively justifiable^{9,22}.

The RNLI dataset findings provide additional evidence that computationally generated clusters are useful for representing specialised knowledge scores^{32,41,42}. Both the BERT and S-BERT models produced six main thematic clusters, with the S-BERT Hierarchical model attaining a CVPA of 0.532—only 0.007 lower than the human benchmark of 0.539. The resulting cluster structure revealed a dominant core theme centred on sea levels and tides, with more specific clusters successfully distinguishing

scientific references (“tides”, “moon”, “waves”) from markers of uncertainty (“know”, “sure”, “don’t”). Although there are clearly distinct conceptual subgroups within the sample, these groups show little association with age and only a weak association with gender. This is consistent with the key conclusion of the tidal literacy literature: misconceptions and gaps in understanding are widespread and cannot be cleanly linked to any single demographic category^{41,42}. A high CVPA value for a specific demographic subgroup would warrant the implementation of narrowly targeted interventions. For instance, this would be appropriate if a particular age cohort were disproportionately concentrated within clusters characterized by a high degree of hazardous confusion. However, the low CVPA observed here indicates that demographic targeting alone is unlikely to be effective/necessary. Improving tidal safety will instead require either broad, population-wide educational initiatives or some mechanism for directly classifying individuals’ conceptual clusters, such as through short diagnostic questions^{3,42}.

Comparing S-BERT and BERT (Table 2) highlights a familiar tension in survey research: trading off geometric “neatness” against social interpretability. In multiple instances, BERT achieves higher Silhouette scores, indicating that it identifies clusters that are more compact and syntactically uniform. Yet S-BERT frequently secures higher CVPA scores, especially for Gender in the RNLI dataset (0.575 vs. 0.515). For scholars in public policy or sociology, this difference is pivotal. A cluster that is “mathematically neat” but fails to correspond to meaningful demographic patterns in the real world may be less informative than a more “messy” cluster that effectively reflects respondents’ underlying social intentions and demographic nuances²⁴.

3.2 Implications for clustering validation and segmentation practice

Across all three datasets, CVPA behaves in line with the primary objectives and conceptual framework detailed in Section 1.2

- In the COVID dataset, certain clustering so-

lutions with only moderate Silhouette scores nonetheless attain high Income and Gender CVPA, indicating clusters that are sociologically meaningful even if they are not geometrically ideal (Table 2).

- In the Norwegian climate dataset, cluster solutions that are geometrically compact and well-separated display low CVPA for income but notable CVPA for age and gender, mirroring earlier topic-model and survey evidence on which groups prioritise particular climate images (Table 2).
- In the RNLI dataset, well-formed conceptual clusters produce low CVPA for age, suggesting that segmentation based on understanding rather than demographics is needed to tackle the safety issue (Table 2).

Thus, CVPA gives concrete form to the frequently invoked but seldom rigorously specified claim that clustering cannot be evaluated independently of the context in which it will be used. By measuring how far a segmentation improves prediction of specific outcomes—such as economic vulnerability under lockdown, the socio-demographic spread of climate narratives, or the patterning of tidal misconceptions—CVPA enables researchers and policymakers to differentiate between partitions that are statistically neat but socially uninformative and more irregular segmentations that capture stable, practically relevant differences within the population.

3.3 Guidelines for Applying CVPA in Social Sciences Research

Based on our cross-dataset analysis, we propose the following interpretive framework:

CVPA > 0.80: *Strong thematic-demographic polarization*

- Indicates distinct social realities requiring targeted interventions.
- Example: COVID-19 income (0.934) → economic stratification demands income-specific support policies.

CVPA 0.50-0.80: Moderate alignment

- Segments have predictive value but considerable overlap.
- Example: COVID-19 gender (0.783) → gender differences exist but within-gender variation remains high.

CVPA < 0.30: Weak alignment or cultural cohesion

- Either clustering failed OR demographics don't structure discourse.
- Example: Norwegian age (0.227) → climate concern unifies generations.
- Requires domain expertise to distinguish failure from unity.

That said, it is important to always compare CVPA to demographic base rates. A CVPA of 0.60 for binary gender is only marginally above random (0.50), while 0.60 for 10-category income is highly informative.

3.4 Scope and Limitations of CVPA

CVPA is designed to address a specific empirical question: “Do the identified thematic clusters correspond to statistically significant demographic or behavioural outcomes?” This explicit analytical focus renders the approach particularly suitable for segmentation tasks in domains such as marketing, public health, and public policy, where the effectiveness of interventions is contingent upon the capacity to delineate and target well-defined population subgroups.

However, CVPA is not appropriate when:

1. **Outcome variables are unavailable:** Many text corpora lack demographic metadata. CVPA requires ground-truth labels for cross-validation.
2. **Thematic structure is not outcome-relevant:** Scientific abstract clustering might optimize for methodology, not author demographics. Low CVPA would not indicate failure, just that different validation is needed.

3. **Clusters are process-oriented:** Temporal evolution of discourse, argumentative structure, or narrative arcs require different validation (e.g., interpretability, stability).

We underscore that CVPA serves as a complement to, rather than a substitute for, conventional geometric clustering indices. An optimal model selection strategy integrates measures of mathematical coherence (e.g., Silhouette coefficient, Davies–Bouldin Index) with indicators of predictive validity (e.g., CVPA), with their relative importance determined by the specific objectives and constraints of the research design.

Collectively, these results suggest that computational clustering serves as a robust alternative or supplement to manual qualitative analysis. By identifying more consistent linguistic patterns relevant to target demographics, computationally created clusters can provide a mathematically grounded framework for survey analysis that captures structural nuances often overlooked during manual coding. The metric helps reserve administrative capacity for more divided issue areas, such as the COVID-19 case, where high CVPA scores clearly indicate that a “one-size-fits-all” strategy would be fundamentally ineffective, while simultaneously identifying contexts in which a population is already aligned in both language and attitudes, enabling institutions to pursue large-scale, unified campaigns that promote social cohesion.

4 Methods

4.1 Data gathering

Several data repositories (such as Harvard Dataverse, Qualitative Data Repository, UK Data Service, ICPSR, and GESIS) were examined between 11 and 18 August 2025 for survey datasets that contained qualitative data and had been manually clustered or thematically analyzed in some way. The search terms used in each data repository were: “beneficiary groups”, “survey themes”, “survey mixed method”, “qualitative data” and “qualitative and quantitative data”. Although these repositories collectively contain a large number of social science and research

datasets, only two data sets met the specific criteria and research needs: i) A Norwegian Citizen Panel survey, which comprised 2,115 open-ended responses to the survey question “What comes to mind when you hear the words ‘climate change’?” gathered during the inaugural wave of the Norwegian Citizen Panel in 2013³¹. The researchers employed structural topic modelling to examine the free text responses, moving beyond the limitations of traditional closed-ended survey methods; and ii) The first nationally representative survey of how the public understands tides, and the misconceptions they hold, was conducted in the British Isles and yielded 1,368 valid questionnaire responses, which included both quantitative items and three open-ended questions. The data were analysed using a combination of thematic analysis and search term analysis. This mixed-methods approach allowed for a systematic examination of people’s experiences of being cut off by the tide and helped uncover recurrent misunderstandings, going beyond the constraints of using quantitative data analysis alone³². These data were further supplemented by the Covid-19 platial dataset, which is a bilingual English-Welsh dataset consisting of six surveys completed between January 2020 and June 2021³³. A linguist analysed them, awarded a score for place attachment, and applied a coding scheme to the data³⁴. Logical checks and adjustments were systematically implemented in automated searches related to the Covid dataset. These three datasets are referred to as Covid, Norwegian, and RNLI datasets for the purposes of this paper.

4.2 Data processing

Our analysis employs an advanced Three-Stage Semantic Clustering Strategy to identify rich and informative themes within multilingual survey responses⁴³. We performed data preprocessing and a normalization step, both of which are crucial for short-text clustering and machine learning methods^{10,44}.

First, to ensure consistent linguistic treatment across varied datasets, the native language of each text response is identified⁴⁵ and the issue of multilingual input was addressed by automatically trans-

lating all data into a single target language⁴⁶ using the GoogleTranslator API, resulting in a unified English text corpus^{37,45}. Automated translation is increasingly recognized as a valid approach for qualitative analysis across languages, largely because recent progress in neural machine translation (NMT) and large language models (LLMs) has improved language interoperability in survey research⁴⁵. Multilingual architectures such as multilingual Bidirectional Encoder Representations from Transformer (mBERT) and Cross-lingual Language Model (XLM) capture shared linguistic patterns, allowing the transfer of knowledge between languages and motivating researchers to develop more advanced cross-lingual models^{45,47,48}. BERT, in particular, has proven highly effective for applications involving cross-language translation⁴⁹. Following translation, responses that were empty or contain only numbers were removed to preserve data quality, so that all subsequent text analysis and embedding generation were based on a consistent linguistic foundation.

4.3 Experiment I: Embedding Technique Comparison

The goal of the embedding technique comparison is to assess how various numerical encodings of text—from classic frequency-based techniques to contemporary transformer-based models affect the correspondence between automated computational clustering and human qualitative coding. By contrasting these approaches, the aim of the experiment is to identify which method best represents the sociologically meaningful patterns recognized by expert coders.

After cleaning and normalizing the raw text, it was converted into numerical vector representations—called embeddings—by applying two separate but complementary methods to capture different linguistic characteristics: a modern semantic embedding approach and a traditional feature-based vectorization approach^{44,50–52}.

Bidirectional Encoder Representations from Transformers (BERT) serves as the core architecture for generating contextual text embeddings. The original BERT model outputs dense vector representations that encode semantic information by processing in-

put sequences in both directions, enabling it to simultaneously take into account preceding and following context⁵². In this work, we use BERT as a baseline model, producing 768-dimensional embeddings that act as a reference point for assessing more specialized methods⁵¹. These embeddings are computed using PyTorch acceleration, allowing large datasets to be processed efficiently in batches⁵¹.

Building on BERT, Sentence-Bidirectional Encoder Representations from Transformers (S-BERT) modifies the original pre-trained BERT architecture to generate semantic embeddings tailored for sentence-level representation. S-BERT employs siamese and triplet network setups to create semantically rich sentence embeddings that can be directly compared using cosine similarity^{29,52,53}. This architecture offers greater computational efficiency than standard BERT⁵². Since all non-English documents were translated to English during preprocessing, we apply S-BERT in a monolingual setting where the model structures the embedding space so that geometric distances between vectors align closely with the semantic similarity of the associated sentences. Recent work shows that these context-sensitive embeddings surpass more conventional methods in capturing subtle, fine-grained sentence-level meaning^{51,54}. In this study, we employ the S-BERT model `all-MiniLM-L6-v2`, which generates compact 384-dimensional embeddings, providing an efficient yet high-quality semantic alternative to full-size BERT models^{51,52}.

Concurrently, Term Frequency–Inverse Document Frequency (TF-IDF) vectors are calculated to encode the relevance of keywords⁵⁵. TF-IDF assigns a weight that reflects the importance of a word in relation to a collection of documents, increasing when the term appears frequently within a given document but rarely across the entire corpus^{51,56}. The vectors were set up to incorporate uni-, bi-, and tri-grams, with a cap of 5,000 features, resulting in sparse representations of the dataset⁵¹.

After generating the initial representations, the high-dimensional embeddings were compressed using Uniform Manifold Approximation and Projection (UMAP)¹⁰. UMAP was chosen over linear approaches such as Principal Component Anal-

ysis (PCA) because it preserves both the non-linear local and global structures that are vital for reliable density-based clustering and visualization^{10,51,57}. This choice is important, as non-linear dimensionality reduction methods like UMAP can reveal tightly related semantic clusters that linear techniques often conceal⁵⁴. The embeddings were thus projected into a low-dimensional space suitable for clustering:

- A 5-dimensional UMAP embedding was used as the primary input for clustering analysis, helping alleviate the issue of dimensionality and providing a robust basis for uncovering data structure^{10,54}.
- A separate 2-dimensional UMAP embedding was generated exclusively for visual inspection and presentation of the results¹⁰.

This manifold learning phase validated the geometric nature of the data: a dense central main topic surrounded by clearly separated, sparse outlier topics, highlighting the need for a clustering approach that can handle this particular density-driven data structure⁴⁰.

4.4 Experiment II: Comparison of Clustering Algorithms

We then evaluated six different clustering approaches, covering partition based, hierarchical, and density-based methods to determine topic configurations that were stable and easy to interpret^{10,58–61}. The semantic clustering pipeline can be seen below in Figure 3. For each method, the hyperparameters were tuned through an exhaustive grid search and assessed using intrinsic clustering metrics^{11,58}.

Partition-based models, in particular K-Means, were tuned using the Calinski–Harabasz Index (where higher values indicate better performance) across a range of cluster numbers from ($k=2$) to 8^{6,58}. For Agglomerative Hierarchical Clustering (AHC), the search focused on identifying the best combination of linkage method (Ward, Average, Complete) and distance metric (Euclidean, Cosine)^{12,58,62}.

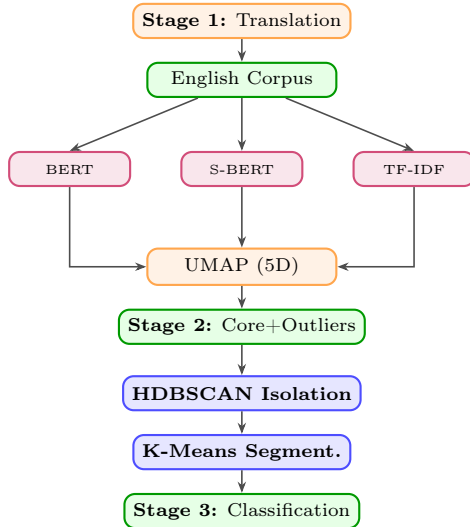


Figure 3: Semantic clustering pipeline (1) preprocessing and translation, (2) embedding generation with UMAP reduction, (3) two-stage clustering combining HDBSCAN outlier detection with K-Means core segmentation.

Density-based methods identify clusters by detecting regions of high data point density separated by areas of lower density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) groups together points that are closely packed based on two key parameters: ϵ (the maximum distance between two points to be considered neighbour) and $MinPts$ (the minimum number of points required to form a dense region)^{25,63}.

Hierarchical DBSCAN (HDBSCAN) extends this approach by constructing a hierarchy of clusters across different density thresholds, allowing it to identify clusters of varying densities and automatically determine optimal parameters such as minimum cluster size⁶⁴. Alongside these density-based algorithms, the Gaussian Mixture Model (GMM) was also assessed, with all methods evaluated using the Silhouette Score (values closer to 1 are preferable)^{6,18,25,58,61–64}.

For general benchmarking, the performance of all optimized models was compared using three

widely used metrics: the Silhouette Score, the Calinski–Harabasz Index, and the Davies–Bouldin Index (DBI), for which lower values indicate better clustering quality^{6,25,58,64}.

Given the confirmed geometric structure of the data, a dense core of primary topics encircled by distinctly separated, sparse outlier topics, a dedicated Two-Stage Semantic Clustering approach was used to handle this dual configuration, which single-stage methods are unable to optimize for at the same time. This hybrid technique is designed to take advantage sequentially of both density-based and partitioning clustering methods^{26,30,62,65}.

1. Isolation (HDBSCAN): In the first stage, the HDBSCAN algorithm was applied to the high-dimensional embedded data (in this case, the UMAP-5D representation) to perform the Isolation step^{10,30,62}. HDBSCAN is well known for its robustness in detecting clusters of arbitrary shapes and heterogeneous densities, as well as for its explicit ability to identify and separate low-density semantic outliers (noise) from the dense core topic^{10,18,25,26,30,60,61,63–67}. This property makes it particularly advantageous compared to traditional DBSCAN and K-Means in high-dimensional semantic settings. The HDBSCAN step therefore centers on assessing the stability of the underlying density structure and isolating potential noise points^{26,67}.
2. Segmentation (K-Means): Secondly, K-Means clustering was carried out solely on the core points obtained from the Isolation step^{30,62,65,68}. This procedure partitions the primary semantic theme into a predefined number of coherent sub-themes^{59,62}. In particular, the model was calibrated by varying (k) (the number of clusters) within a specified interval (e.g. $k = 10$) to strike an appropriate balance between the simplicity of the cluster structure and the depth of interpretation in the thematic decomposition of the text^{15,30,62,69,70}.

The final thematic classification combines the outliers detected by HDBSCAN with the sub-themes produced by K-Means³⁰. Cluster interpretation is

based on TF-IDF scores^{10,62}. For each group, the TF-IDF scores are calculated for all terms, and the top-scoring terms are then used to assign descriptive objective keywords to that group^{62,68}. TF-IDF remains the reference method for identifying and summarizing key terms, thereby improving the interpretability of clusters in contemporary surveys and social data analysis^{10,56,62,64,68}. Overall, this pipeline yielded robust and readily interpretable thematic segmentation by incorporating established best practices from the manifold learning and unsupervised clustering literature^{30,66}.

4.5 Experiment III: Employing CVPA to assess the quality of the derived clusters

To assess the sociological significance of the computationally generated clusters, this work employs Cross-Validated Predictive Accuracy (CVPA). In contrast to conventional geometric measures like the Silhouette coefficient, which focus on the mathematical distinctness of clusters, CVPA implements predictive validity by measuring how consistently cluster assignments align with real-world demographic characteristics. The validation approach defines a human upper bound by computing the CVPA for qualitative codes that expert researchers have already assigned manually.

By placing these manual baselines directly alongside the computationally clustered results, the method provides a formal, quantifiable evaluation of whether the computationally clustered results have successfully captured the same demographic-correlated signals as the human coding scheme. If computationally created clusters achieve CVPA similar to or higher than the manual baseline, it indicates that the automated clusters identify consistent linguistic patterns that are at least as sociologically informative and actionable as those derived through manual qualitative analysis.

Competing interests

The authors declare no competing interests.

Author contributions

L.S.F. developed the methodology, wrote all analysis code, conducted the formal analysis, and wrote the manuscript. S.W. and W.T. provided supervision and contributed critical feedback on successive drafts. All authors reviewed and approved the final manuscript.

Data availability

No new primary data were collected for this study. The COVID-19 access-to-nature survey data analysed in this paper were originally collected as part of Dolan (2024) and are available from Bangor University Research at <https://research.bangor.ac.uk/en/studentTheses/access-to-nature-and-its-impact-on-peoples-wellbeing/>. The RNLI tidal literacy survey data were collected as part of Morris-Webb et al. (2025), published open access in Ocean and Society (DOI: 10.17645/oas.9793); data access enquiries should be directed to the corresponding author of that study. The Norwegian Citizen Panel survey data on climate change perceptions were originally published in Fløttum et al. (2015, Nature Climate Change, DOI: 10.1038/nclimate2663) and are publicly available via Harvard Dataverse at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28689>. All data were used in accordance with the terms under which they were originally made available.

Code availability

The code used for all clustering experiments, CVPA computation, and figure generation is available from the corresponding author (L.S.F.) upon reasonable request and will be made publicly available via GitHub upon acceptance of the manuscript.

Ethics

This study is based entirely on pre-existing survey datasets collected and governed under the ethical

frameworks of their respective originating institutions. The COVID-19 access-to-nature survey data were collected under Bangor University’s ethical governance. The RNLI tidal literacy survey data were collected under the ethical framework of the Royal National Lifeboat Institution. The Norwegian open-ended survey data on climate change perceptions were originally collected as part of the Norwegian Citizen Panel and published in Fløttum et al. (2015, Nature Climate Change, DOI: 10.1038/nclimate2663); these data were used in accordance with the terms of that study. No new human participant data were collected by the authors of this study, and no additional ethical approval was required.

Funding

L.S.F. is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence, Machine Learning and Advanced Computing (AIMLAC), funded by UK Research and Innovation (UKRI) grant EP/S023992/1.

References

- [1] Hizirolu, A. Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications* **40**, 6491–6507 (2013). URL <https://www.sciencedirect.com/science/article/pii/S0957417413003503>.
- [2] Barnett, C. & Mahony, N. Segmenting Publics (2011).
- [3] He, Q., Liao, D. & Jiao, H. Clustering Behavioral Patterns Using Process Data in PIAAC Problem-Solving Items. In Veldkamp, B. P. & Shuijter, C. (eds.) *Theoretical and Practical Advances in Computer-based Educational Measurement*, 189–212 (Springer International Publishing, Cham, 2019). URL https://doi.org/10.1007/978-3-030-18480-3_10.
- [4] Ezugwu, A. E. *et al.* A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence* **110**, 104743 (2022). URL <https://linkinghub.elsevier.com/retrieve/pii/S095219762200046X>.
- [5] Williams, W. T. Principles of Clustering. *Annual Review of Ecology and Systematics* **2**, 303–326 (1971). URL <https://www.annualreviews.org/doi/10.1146/annurev.es.02.110171.001511>.
- [6] Bertsimas, D., Orfanoudaki, A. & Wiberg, H. Interpretable clustering: an optimization approach. *Machine Learning* **110**, 89–138 (2021). URL <https://doi.org/10.1007/s10994-020-05896-2>.
- [7] Ezugwu, A. E. *et al.* Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications* **33**, 6247–6306 (2021). URL <https://doi.org/10.1007/s00521-020-05395-4>.
- [8] Dahal, S. Effect of different distance measures in result of cluster analysis (2015). URL <https://aaltodoc.aalto.fi/handle/123456789/17787>.
- [9] Liu, T., Yu, H. & Blair, R. H. Stability estimation for unsupervised clustering: A review. *WIREs Computational Statistics* **14**, e1575 (2022). URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1575>. eprint: <https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1575>
- [10] Ahmed, M. H., Tiun, S., Omar, N. & Sani, N. S. Short Text Clustering Algorithms, Application and Challenges: A Survey. *Applied Sciences* **13**, 342 (2023). URL <https://www.mdpi.com/2076-3417/13/1/342>. Number: 1.
- [11] Ditton, E., Swinbourne, A., Myers, T. & Scovell, M. Applying Semi-Automated Hyperparameter Tuning for Clustering Algorithms (2021). URL <http://arxiv.org/abs/2108.11053>. ArXiv:2108.11053 [cs].

- [12] Trudie Strauss. Generalising Ward’s Method for Use with Manhattan Distances | PLOS One (2017). URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168288>.
- [13] Jaeger, A. & Banks, D. Cluster analysis: A modern statistical review. *WIREs Computational Statistics* **15**, e1597 (2023). URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1597>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1597>.
- [14] Mickus, T., Paperno, D., Constant, M. & van Deemter, K. What do you mean, BERT? Assessing BERT as a Distributional Semantics Model (2020).
- [15] Naeem, S. & Wumaier, A. Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K. *International Journal of Computer Applications* **182**, 7–14 (2018). URL <http://www.ijcaonline.org/archives/volume182/number31/naeem-2018-ijca-918234.pdf>.
- [16] Battaglia, O. R., Di Paola, B. & Fazio, C. A quantitative analysis of Educational Data through the Comparison between Hierarchical and Not-Hierarchical Clustering. *EURASIA Journal of Mathematics, Science and Technology Education* **13** (2017). URL <https://www.ejmste.com/article/a-quantitative-analysis-of-educational-data-through-the-comparison-between-hierarchical-and-4890>.
- [17] Bafna, P., Pramod, D. & Vaidya, A. Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61–66 (2016). URL <https://ieeexplore.ieee.org/abstract/document/7754750>.
- [18] Wang, L., Chen, P., Chen, L. & Mou, J. Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach. *Journal of Marine Science and Engineering* **9**, 566 (2021). URL <https://www.mdpi.com/2077-1312/9/6/566>.
- [19] Almeida, H., Guedes, D., Meira, W. & Zaki, M. J. Is There a Best Quality Metric for Graph Clusters? In Gunopulos, D., Hofmann, T., Malerba, D. & Vazirgiannis, M. (eds.) *Machine Learning and Knowledge Discovery in Databases*, 44–59 (Springer, Berlin, Heidelberg, 2011).
- [20] Engl, E., Smittenaar, P. & Sgaier, S. K. Identifying population segments for effective intervention design and targeting using unsupervised machine learning: an end-to-end guide. *Gates Open Research* **3**, 1503 (2019). URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC6820452/>.
- [21] Dao Thi-Bich-Hanh, Vrain Christel, Duong Khanh-Chuong & Davidson Ian. A Framework for Actionable Clustering Using Constraint Programming. In *Frontiers in Artificial Intelligence and Applications* (IOS Press, 2016). URL <https://www.medra.org/servlet/aliasResolver?alias=iospress&isbn=978-1-61499-671-2&spage=453&doi=10.3233/978-1-61499-672-9-453>.
- [22] Trivedi, S., Pardos, Z. A. & Heffernan, N. T. The Utility of Clustering in Prediction Tasks (2015). URL <http://arxiv.org/abs/1509.06163>. ArXiv:1509.06163 [cs].
- [23] Tibshirani, R. & Walther, G. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics* **14**, 511–528 (2005). URL <https://doi.org/10.1198/106186005X59243>. eprint: <https://doi.org/10.1198/106186005X59243>.
- [24] Hu, L., Jiang, M., Dong, J., Liu, X. & He, Z. Interpretable Clustering: A Survey (2024). URL <http://arxiv.org/abs/2409.00743>. ArXiv:2409.00743 [cs].
- [25] Bushra, A. A. & Yi, G. Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE*

- Access* **9**, 87918–87935 (2021). URL <https://ieeexplore.ieee.org/document/9453785/>.
- [26] Malzer, C. & Baum, M. A Hybrid Approach To Hierarchical Density-based Cluster Selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 223–228 (2020). URL <https://ieeexplore.ieee.org/abstract/document/9235263>.
- [27] Hsu, D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Applied Energy* **160**, 153–163 (2015). URL <https://www.sciencedirect.com/science/article/pii/S0306261915010624>.
- [28] Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData Mining* **9**, 5 (2016). URL <https://doi.org/10.1186/s13040-016-0084-6>.
- [29] Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (2019). URL <http://arxiv.org/abs/1908.10084>. ArXiv:1908.10084 [cs].
- [30] Badillo, B. R. Mathematical Methods to Uncovering Latent Structures in High-Dimensional Community Survey Data: Clustering, Dimensionality Reduction, and Sentiment Analysis for Community Insights (2025).
- [31] Tvinnereim, E. & Fløttum, K. Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change* **5**, 744–747 (2015). URL <https://www.nature.com/articles/nclimate2663>.
- [32] Morris-Webb, E. S. *et al.* Cut Off by the Tide: How Ocean Literacy Can Help Save Lives. *Ocean and Society* **2** (2025). URL <https://www.cogitatiopress.com/oceanandsociety/article/view/9793>.
- [33] Dolan, R. Access to nature and its impact on peoples’ wellbeing - Bangor University (2024). URL [https://research.bangor.ac.uk/portal/en/theses/access-to-nature-and-its-impact-on-peoples-wellbeing\(fd56e979-ddb6-46b4-bfa9-f4256865a7b5\).html](https://research.bangor.ac.uk/portal/en/theses/access-to-nature-and-its-impact-on-peoples-wellbeing(fd56e979-ddb6-46b4-bfa9-f4256865a7b5).html).
- [34] Tenbrink, T. & Willcock, S. Place attachment and perception of climate change as a threat in rural and urban areas. *PLOS ONE* **18**, e0290354 (2023). URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0290354>.
- [35] Zhang, W., Yoshida, T. & Tang, X. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications* **38**, 2758–2765 (2011). URL <https://www.sciencedirect.com/science/article/pii/S0957417410008626>.
- [36] Wang, B. & Kuo, C.-C. J. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 2146–2157 (2020). URL <https://ieeexplore.ieee.org/abstract/document/9140343>.
- [37] Jawahar, G., Sagot, B. & Seddah, D. What does BERT learn about the structure of language? (2019). URL <https://inria.hal.science/hal-02131630>.
- [38] Choi, H., Kim, J., Joe, S. & Gwon, Y. Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 5482–5487 (2021). URL <https://ieeexplore.ieee.org/abstract/document/9412102>.
- [39] Sharon A. Brown, Sandra L. Upchurch & Gayle J. Acton. A Framework for Developing a Coding Scheme for Meta-Analysis (2003). URL <https://journals.sagepub.com/doi/epdf/10.1177/0193945902250038>.

- [40] Opitz, J. & Frank, A. SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features (2022). URL <http://arxiv.org/abs/2206.07023>. ArXiv:2206.07023 [cs].
- [41] Farrelly, C. M. *et al.* The analysis of bridging constructs with hierarchical clustering methods: An application to identity. *Journal of Research in Personality* **70**, 93–106 (2017). URL <https://www.sciencedirect.com/science/article/pii/S0092656617300685>.
- [42] Amato, F. *Mixed data temporal clustering for modelling longitudinal surveys*. phdthesis, Université Lumière Lyon 2 (2025). URL <https://hal.science/tel-05108072>.
- [43] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL <https://aclanthology.org/N19-1423/>.
- [44] Yamani, Shirin. *Exploration of Techniques for Working with Sparse Data when Applying Natural Language Processing to Assist a Qualitative Data Analysis of a COVID-19 Open Innovation Community*. Ph.D. thesis, University of Calgary, Calgary, Canada (2024).
- [45] Zhong, T. *et al.* Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research (2024). URL <http://arxiv.org/abs/2412.04497>. ArXiv:2412.04497 [cs].
- [46] Andersson, J. A comparison of different methods in their ability to compare semantic similarity between articles and press releases (2022).
- [47] Wang, H., Li, J., Wu, H., Hovy, E. & Sun, Y. Pre-Trained Language Models and Their Applications. *Engineering* **25**, 51–65 (2023). URL <https://www.sciencedirect.com/science/article/pii/S2095809922006324>.
- [48] Alyafeai, Z., AlShaibani, M. S. & Ahmad, I. A Survey on Transfer Learning in Natural Language Processing (2020). URL <http://arxiv.org/abs/2007.04239>. ArXiv:2007.04239 [cs].
- [49] Gardazi, N. M. *et al.* BERT applications in natural language processing: a review. *Artificial Intelligence Review* **58**, 166 (2025). URL <https://doi.org/10.1007/s10462-025-11162-5>.
- [50] Wang, J. *et al.* Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Computing Surveys* **56**, 1–33 (2024). URL <https://dl.acm.org/doi/10.1145/3648471>.
- [51] Malmberg, J. *Evaluating semantic similarity using sentence embeddings* (2021). URL <https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-291425>.
- [52] Jamshidian, M. Evaluation of Text Transformers for Classifying Sentiment of Reviews by Using TF-IDF, BERT (word embedding), SBERT (sentence embedding) with Support Vector Machine Evaluation (2022).
- [53] Wang, S., Thompson, L. & Iyyer, M. PhraseBERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration (2021). URL <http://arxiv.org/abs/2109.06304>. ArXiv:2109.06304 [cs].
- [54] Stankevičius, L. & Lukoševičius, M. Extracting Sentence Embeddings from Pretrained Transformer Models. *Applied Sciences* **14**, 8887 (2024). URL <https://www.mdpi.com/2076-3417/14/19/8887>. Number: 19.
- [55] Venkata Subramanian, D., Thangakumar, J., Preetha, M., Ganga, S. & Sajeev, S. Similarities and Ranking of Documents Using TF-IDF,

- LDA and WAM. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 01–07 (2024). URL <https://ieeexplore.ieee.org/abstract/document/10533526>.
- [56] Al-Obaydy, W. N. I., Hashim, H. A., Najm, Y. A. & Jalal, A. A. Document classification using term frequency-inverse document frequency and K-means clustering. *Indonesian Journal of Electrical Engineering and Computer Science* **27**, 1517 (2022). URL <https://ijeecs.iaescore.com/index.php/IJECS/article/view/27717>.
- [57] Taha, K., Yoo, P. D., Yeun, C., Homouz, D. & Taha, A. A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review* **54**, 100664 (2024). URL <https://linkinghub.elsevier.com/retrieve/pii/S1574013724000480>.
- [58] Griffiths, W. E. & Hajargasht, G. GMM Estimation of Mixtures from Grouped Data: An Application to Income Distributions (2012).
- [59] Calìò, F. Advancing Keyword Clustering Techniques: A Comparative Exploration of Supervised and Unsupervised Methods: Investigating the Effectiveness and Performance of Supervised and Unsupervised Methods with Sentence Embeddings (2023).
- [60] Zaki Abdulhameed, T., Yousif, S. A., Samawi, V. W. & Imad Al-Shaikhli, H. SS-DBSCAN: Semi-Supervised Density-Based Spatial Clustering of Applications With Noise for Meaningful Clustering in Diverse Density Data. *IEEE Access* **12**, 131507–131520 (2024). URL <https://ieeexplore.ieee.org/abstract/document/10670579>.
- [61] Singh, G. DYNAMIC CLUSTER TRACKING USING DB SCAN ALGORITHM. *SSRN Electronic Journal* (2013). URL <http://www.ssrn.com/abstract=2266258>.
- [62] Orabi Alkhen, W. *Exploring Advanced Clustering Techniques for Business Descriptions : A Comparative Study and Analysis of DB-SCAN, K-Means, and Hierarchical Clustering* (2023). URL <https://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-48770>.
- [63] Firdaus, S. & Uddin, A. A Survey on Clustering Algorithms and Complexity Analysis. *International Journal of Computer Science Issues* **12** (2015).
- [64] Ibrahimoh, R. & Aderoba, A. Comparison of K-Means and HDBSCAN Clustering Approaches to Enhance Marketing Strategies **8** (2024).
- [65] Mohotti, W. A. & Nayak, R. Corpus-Based Augmented Media Posts with Density-Based Clustering for Community Detection. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, 379–386 (2018). URL <https://ieeexplore.ieee.org/abstract/document/8576064>.
- [66] McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* **2**, 205 (2017). URL <http://joss.theoj.org/papers/10.21105/joss.00205>.
- [67] Bhattacharjee, P. & Mitra, P. A survey of density based clustering algorithms. *Frontiers of Computer Science* **15**, 151308 (2020). URL <https://doi.org/10.1007/s11704-019-9059-3>.
- [68] Saeed, M. Y., Awais, M., Talib, R. & Younas, M. Unstructured Text Documents Summarization With Multi-Stage Clustering. *IEEE Access* **8**, 212838–212854 (2020). URL <https://ieeexplore.ieee.org/abstract/document/9269966>.
- [69] Zahra, S. *et al.* Novel centroid selection approaches for KMeans-clustering based recommender systems. *Information Sciences* **320**, 156–189 (2015). URL <https://www.sciencedirect.com/science/article/pii/S0020025515002352>.

- [70] Abdalgader, K. Clustering Short Text using a Centroid-Based Lexical Clustering Algorithm (2017).