9 Unsupervised Learning for Embodied Emotion Discovery

Nataša Kovač, Faculty of Applied Sciences, University of Donja Gorica, Oktoih 1, Podgorica, 81000, Montenegro, https://orcid.org/ [0000-0002-6671-2938](https://orcid.org/0000-0002-6671-2938)

Hojjatollah Farahani, Department of Psychology, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran, <https://orcid.org/0000-0002-9799-7008>

Peter Watson, Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom, https://orcid.org/0000-0002-9436-0693

**Abstract:** In the chapter, the focus is on paving the way for a novel approach to exploring embodied emotions without explicit labels. It begins by establishing the groundwork necessary for unsupervised exploration of emotions, emphasising the need for innovative techniques. The chapter then delves into the intricacies of learning algorithms tailored for detecting emotions in embodied contexts, emphasising their adaptability to unsupervised scenarios. Furthermore, it explores the application of unsupervised learning methods to various data modalities, including EEG signals, images, text data, and audio signals, showcasing the versatility of this approach in deciphering emotions without relying on predefined emotional labels.

**Keywords:** Unsupervised Embodied Emotion; Learning Algorithm; Unsupervised EEG Signal Analysis; Unsupervised Image Analysis; Unsupervised Text Data Analysis; Unsupervised Audio Signal Analysis

9.1 Introduction

Understanding human emotion is inherently complex. Emotional expressions emerge through intricate physiological patterns, subtle behavioural changes, nuanced textual expressions, and dynamic voice variations. As explored in previous chapters, supervised machine learning methods (algorithms trained explicitly on labelled emotional data) are powerful yet limited by the inherent need for extensive and accurately labelled datasets (Calvo & D’Mello, 2010). However, the process of labelling emotional states is subjective, labour-intensive, and sometimes inaccurate due to cultural variations, subjective judgments, and ambiguous emotional boundaries.

Unsupervised learning, in contrast, offers a novel avenue for embodied emotion discovery, bypassing the constraints of labelled datasets by enabling algorithms to explore and uncover hidden emotional structures directly from the raw data itself. Unlike supervised learning, unsupervised methods allow researchers to reveal natural groupings, subtle emotional distinctions, and unique physiological signatures without predefined emotion categories or labels (Hastie, Tibshirani, & Friedman, 2009).

Unsupervised learning has emerged as an essential methodological paradigm in the exploration of embodied emotional phenomena, allowing researchers to uncover latent patterns within multimodal behavioural and physiological data without relying on predefined emotional labels. In the context of affective science and computational psychology, *embodied emotion* refers to the dynamic coupling between bodily states, perceptual feedback, and internal affective experiences. Traditional supervised machine learning paradigms, which depend heavily on labelled datasets reflecting subjective human annotation schemes, often impose rigid categorical frameworks, such as discrete emotions or continuous affective dimensions, that may obscure the complex, context-dependent nature of emotional embodiment. Conversely, unsupervised learning algorithms enable the discovery of emergent emotional structures by directly engaging with high-dimensional sensor data, including posture, movement dynamics, muscle activation patterns, facial micro-expressions, heart rate variability, and electrodermal activity. Through clustering, dimensionality reduction, and manifold learning approaches, unsupervised systems can infer latent affective topologies that correspond to embodied states of arousal, valence, and motivational orientation, for example, techniques such as Gaussian mixture modelling, t-distributed stochastic neighbour embedding, and autoencoder networks, which allow the extraction of compact, interpretable representations of multimodal emotion-related data. The resulting low-dimensional embeddings capture relational patterns among bodily signals and afford the identification of implicit emotional prototypes that are not bound to preexisting psychological taxonomies. Additionally, more recent advances in deep unsupervised learning, i.e., variational autoencoders, contrastive learning frameworks, and diffusion-based generative models, further enhance the capacity to model nonlinear temporal dependencies and subtle intersubject variations in embodied expression. These models offer the possibility of formulating a data-driven cartography of emotional embodiment, where emotions are conceptualised as dynamic attractor states within sensorimotor and interoceptive spaces rather than as static categorical entities (Kingma & Welling, 2013; van der Maaten & Hinton, 2008). Integrating unsupervised learning with embodied emotion research also necessitates a shift in epistemological stance. Rather than treating emotion as an internal variable to be inferred from external behaviour, unsupervised algorithms approach emotion as an emergent property of coordinated bodily and contextual processes. This view aligns with inactivist and constructivist perspectives in affective neuroscience, which argue that emotions are enacted through bodily regulation and environmental engagement (Seth & Tsakiris, 2018; Barrett, 2017). When embodied data, such as motion capture trajectories or physiological oscillations, are subjected to unsupervised feature extraction, the emergent dimensions often map onto psychophysiological processes that underlie affective experience, such as autonomic synchrony, reactive motor readiness, or interoceptive awareness. Revealing these latent organisational principles, unsupervised learning provides a computational bridge between neural, bodily, and experiential components of emotion (Levenson, 2014). Moreover, the integration of unsupervised methods into embodied emotion holds significant implications for cross-cultural and developmental research. These models can identify population-specific or context-dependent structures of affective embodiment. because they don't presuppose culturally biased emotion categories. In developmental contexts, unsupervised analysis of infants’ or children’s movement and physiological data can uncover pre-linguistic affective patterns that precede the acquisition of explicit emotional concepts (Zhang et al., 2023). Similarly, in clinical applications, unsupervised learning can detect atypical emotional coordination patterns in disorders of affect regulation, such as depression, anxiety, or autism spectrum conditions, thus offering diagnostic and therapeutic insights grounded in embodied signal dynamics rather than subjective report (Kragel et al., 2016). Ultimately, unsupervised learning for embodied emotion represents an epistemic and methodological convergence between data-driven modelling and embodied cognition theory. By relinquishing the dependence on human-labelled emotion corpora, this approach opens a pathway towards understanding emotion as a distributed, emergent, and adaptive phenomenon, one that can be quantitatively described yet remains deeply grounded in bodily action and perception. The challenge moving forward lies in developing interpretability frameworks that translate the abstract representations generated by unsupervised models into psychologically meaningful constructs, thereby maintaining empirical rigour while respecting the complexity of embodied emotional life (Barsade & Gibson, 2012; Seth & Tsakiris, 2018).

This chapter focuses deeply on exploring unsupervised methods, demonstrating how these techniques powerfully enhance emotional AI research, offering fresh insights and uncovering new emotional phenomena previously undetectable through traditional methods. By learning directly from the inherent structure and patterns in physiological, behavioural, and textual signals, unsupervised algorithms can identify natural clusters of emotional experiences, delineate subtle emotional distinctions, and help generate entirely new emotional taxonomies aligned closely with actual human emotional experiences (Mehta, Siddiqui, & Javaid, 2019).

In the forthcoming sections, readers will navigate step-by-step through essential preparations for unsupervised emotional discovery, delve into powerful unsupervised learning algorithms, and apply these methods practically across key emotional modalities such as EEG signals, image analysis, text data, and audio signals. By embracing unsupervised learning methodologies, researchers and practitioners can more fully uncover and understand the embodied nature of emotional experience, shedding new light on the intricate emotional landscape inherent to human cognition, physiology, and social interaction.

The journey begins by carefully preparing the ground for unsupervised analysis, identifying optimal preprocessing techniques, and laying a robust methodological foundation, before advancing into diverse applications that illustrate the power and versatility of unsupervised emotion exploration. Throughout this exploration, practical Python examples, clear visualisations, methodological insights, and careful ethical considerations will guide readers to confidently apply unsupervised techniques to their research and applications.

9.2 Preparing the Ground for Unsupervised Embodied Emotion Exploration

Unsupervised embodied–emotion analysis starts with stable, label-free representations that already encode affective regularities in speech, vision, text, and physiology. Self-supervised encoders are the practical backbone here: emotion2vec for speech, DINOv2 for images, SBERT for text, and MNE-centric pipelines for EEG (Gramfort et al., 2013) that preserve physiologically meaningful variance before any clustering. These choices reduce reliance on annotation and improve generalisation across speakers, cameras, and sessions (Ma et al., 2023).

Benchmark datasets help align discovery with human judgments without driving the learning step itself. IEMOCAP (Busso et al., 2008) is widely used for speech and dyadic interaction. RAVDESS (Livingstone & Russo, 2018) provides carefully controlled audio-visual expressions, AffectNet (Mollahosseini et al., 2017) brings large-scale in-the-wild faces with both categorical and dimensional labels, and DEAP (Koelstra et al., 2011) integrates EEG with peripheral physiology under controlled stimuli. Using these sets for evaluation only keeps the exploration label-agnostic while enabling later triangulation.

EEG preparation benefits from standardised routines. The PREP pipeline (Bigdely-Shamlo et al., 2015) automates referencing and line-noise handling at scale, and independent component analysis remains a reliable method for ocular (Jung et al., 2000) and muscle artefact separation. Open tools in MNE-Python make these steps reproducible and transparent in research reports. Multimodal studies need accurate timing. When audio, video, and EEG are recorded on different devices, clock drift can mask true relations between channels. The Lab Streaming Layer provides per-sample timestamps, offset correction, and low-jitter synchronisation, which preserves the temporal structure that unsupervised discovery uses (Kothe et al., 2025).

Feature backbones should be compact and consistent across modalities. Wav2vec 2.0 and HuBERT (Hsu et al., 2021) remain strong general speech encoders, while emotion2vec focuses directly on affect transfer. DINOv2 (Oquab et al, 2023) yields robust visual features for frames or clips without labels, and SBERT (Reimers & Gurevych, 2019) provides sentence-level vectors that support topic and affect discovery in text. Standardising these representations and concatenating them into a single table prepares the ground for geometry and density analysis (Baevski et al., 2020).

Geometry and density must match the data. If the embedding space already separates affective neighbourhoods, partitions such as K-means are adequate and easy to report. When states are interleaved and uneven, HDBSCAN (Campello et al., 2020) identifies dense cores and flags noise instead of forcing every sample into a group. For visualisation without labels, UMAP preserves local neighbourhoods and often reveals smooth affective trajectories (McInnes et al., 2018). Multi-site or mixed-device studies can introduce batch effects that dominate unsupervised patterns. Recent EEG work demonstrates that ComBat-style harmonisation can reduce between-site artefacts in spectral features while preserving biological associations, thereby allowing clusters to reflect behaviour rather than hardware (Jaramillo-Jimenez et al., 2024).

Ethical preparation is part of the method. Contemporary analyses document demographic and domain biases in facial and speech emotion systems and call for fairness-aware evaluation, especially when models are trained or assessed with unbalanced datasets. In parallel, privacy-preserving strategies for affect recognition are moving forward with federated and differentially private methods for physiological and audio-visual signals (Sajjad et al., 2023).

Methodologically, the documentation standards you use in supervised psychology carry over directly to the unsupervised stage. Clear data preparation, diagnostics, and artefact sharing keep the subsequent interpretation of latent states defensible when linked to constructs such as internal shame or chronic pain (Kovač et al., 2024; Kovač et al., 2025a; Kovač et al., 2025b). This continuity supports transparent pipelines from discovery to inference.

Figure 1 illustrates the first visual check researchers typically perform. A two-component PCA projection of standardised multimodal features is coloured by K-means assignments, which reveals whether compact neighbourhoods exist before any sophisticated manifold analysis. When clusters appear in this view, analysts gain confidence that later density or graph-based methods will refine structure rather than create it.

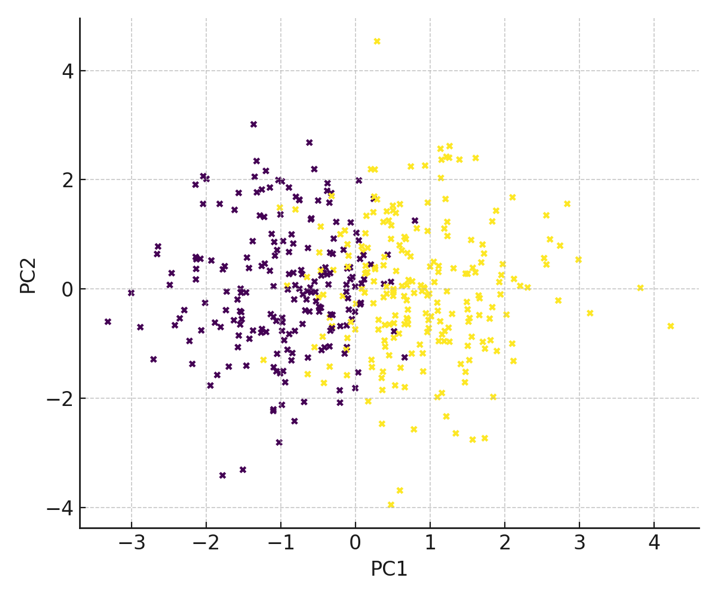
****

Figure 1. PCA projection of multimodal features with K-means clusters

Model parsimony follows next. Figure 2 shows the elbow curve of within-cluster sum of squares as K increases. The bend indicates diminishing returns, which is useful when the goal is to keep the affect map interpretable for clinical or behavioural reporting. Selecting the smallest K near the elbow keeps the representation compact without sacrificing major structure.

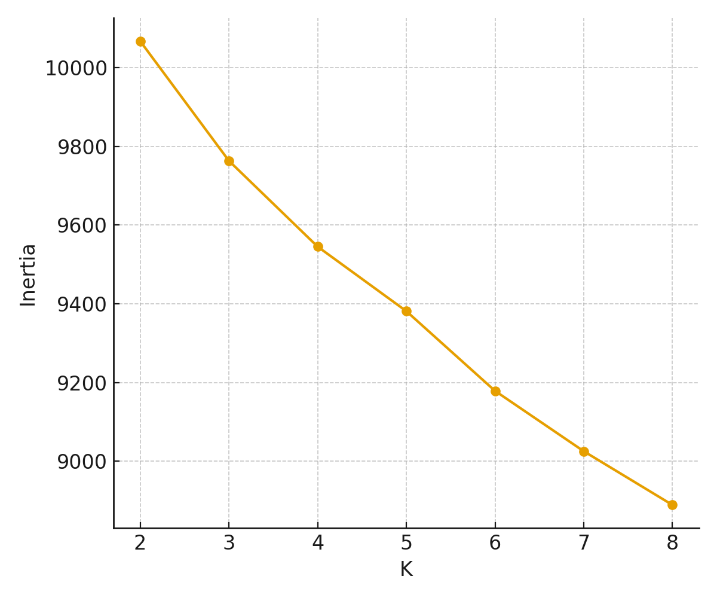
****

Figure 2. Elbow curve of the within-cluster sum of squares across K

Separation and cohesion should also be quantified. Figure 3 plots the silhouette across K, summarising how well samples match their own cluster compared to the nearest alternative. This diagnostic is quick to compute and helps guard against overly fine partitions that appear neat visually but have weak internal consistency (Rousseeuw, 1987).

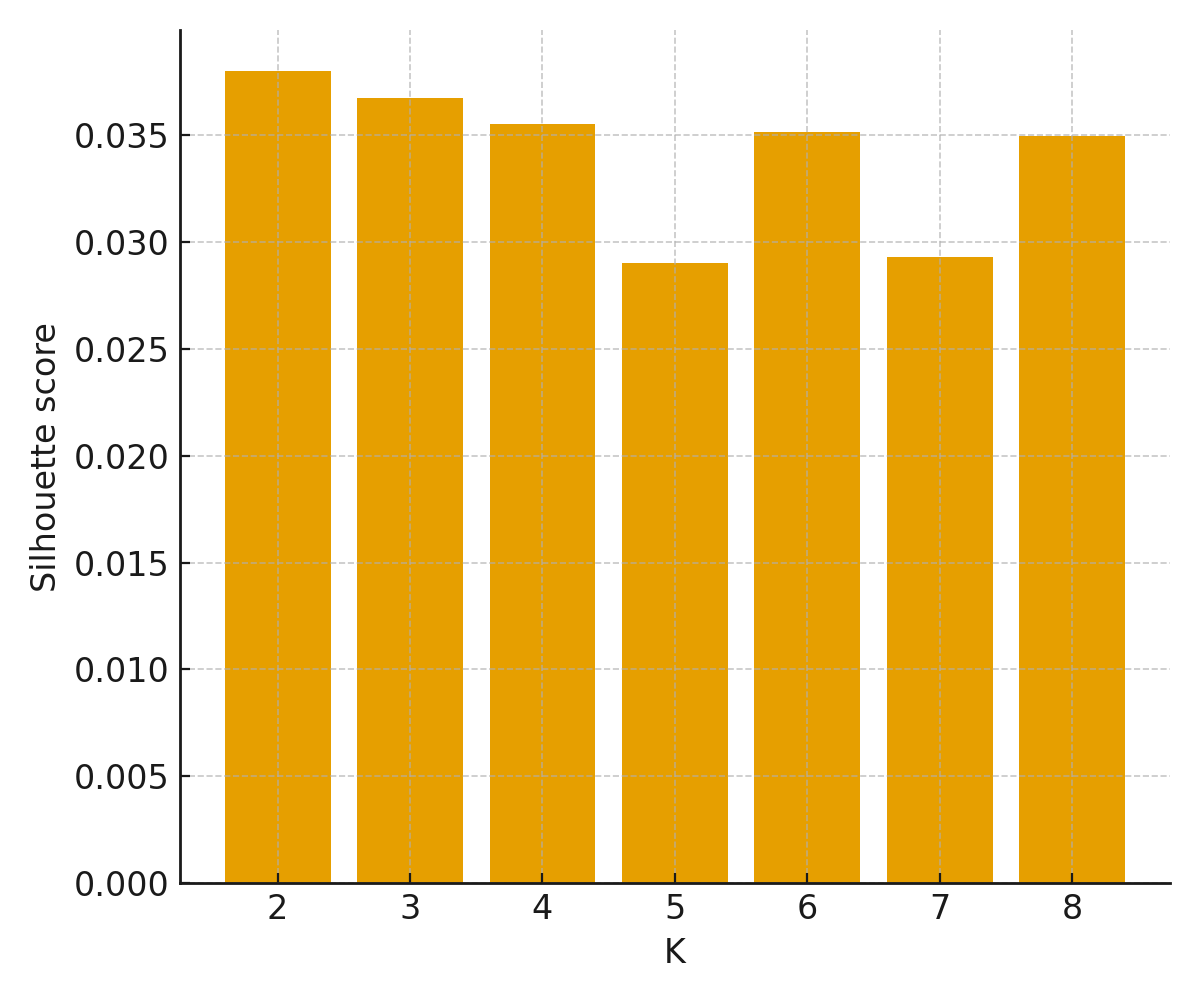


Figure 3. Silhouette scores across K for quick partition diagnostics

Preparation choices by modality are summarised in Table 1. The emphasis is on compact features and encoders that are stable across datasets, which reduces the risk that unsupervised structure reflects device quirks rather than behaviour. For EEG, the table highlights a practical sequence that many teams adopt before clustering. For speech and vision, the listed encoders have shown strong transfer even without labels. For text, sentence-level vectors remain the most dependable starting point.

Table 1. Modalities, compact feature choices, and encoder suggestions

|  |  |  |
| --- | --- | --- |
| Modality | Compact features in practice | Encoder or recipe |
| Speech | Utterance embeddings pooled from frames | emotion2vec for affect transfer (Ma et al., 2023); wav2vec-2.0 or HuBERT for general features (Hsu et al., 2021). |
| EEG | Band powers, connectivity, and short spatiotemporal blocks | PREP → ICA → features in MNE-Python; consider self‑supervised learning (SSL) pretraining before clustering (Bigdely-Shamlo et al., 2015). |
| Faces | Frame or clip features with identity suppression | DINOv2 for universal visual embeddings that cluster well without labels (Oquab et al, 2023). |
| Text | Sentence vectors and topic footprints | SBERT embeddings with topic discovery when needed (Reimers & Gurevych, 2019). |

Reproducibility is helped by a concise checklist, presented in Table 2, that you can attach to preregistrations or method appendices. The steps run from standardisation and visualisation to model selection and device harmonisation. UMAP is suggested for neighbourhood visualisation because it preserves local structure in two dimensions without imposing linear constraints, while HDBSCAN is a practical density option when data contain sparse episodes or artefacts (McInnes, Healy, & Melville, 2018; Campello, Moulavi, Zimek, & Sander, 2020). Synchronisation and batch control complete the list so that temporal and site effects do not masquerade as emotion structure (Kothe, 2025; Jaramillo-Jimenez et al., 2024).

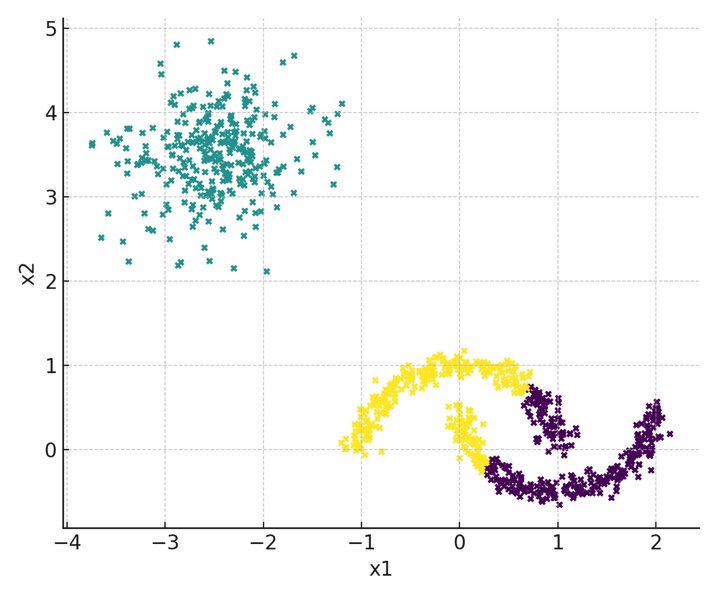
Table 2. Quick checklist for reproducible preparation

|  |  |  |
| --- | --- | --- |
| Step | Rationale | Typical tool |
| Standardize features | Comparable scales across modalities | StandardScaler |
| Visualize geometry | Early neighborhood sanity check | PCA or UMAP (McInnes et al., 2018) |
| Pick K or choose density | Match model to shape of data | K-means vs. HDBSCAN (Campello et al., 2020) |
| Synchronize streams | Preserve cross-modal timing | Lab Streaming Layer (Kothe et al., 2025) |
| Control batch effects | Prevent site or device drift | ComBat for EEG features (Jaramillo-Jimenez et al., 2024) |

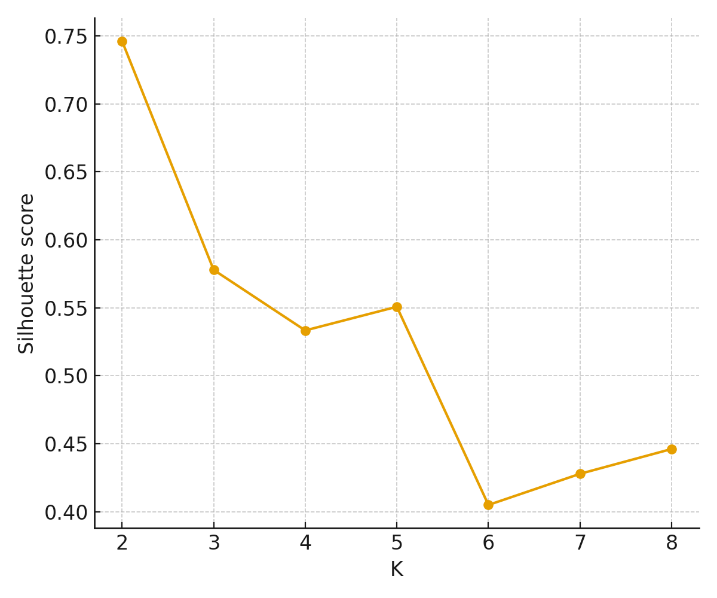
9.3 Learning Algorithms in Embodied Emotion Detection

Partitioning methods are often the first stop once multimodal embeddings are reasonably well separated, because compact regions in a standardised space map cleanly to affective *neighbourhoods* that practitioners can inspect and name (McQueen, 1967). Gaussian mixtures extend this idea with probabilistic membership, which handles gradual transitions between states that are common in spontaneous affective behaviour and psychotherapy sessions (Dempster, Laird, & Rubin, 1977; Bishop & Nasrabadi, 2006). Model selection can remain parsimonious by using penalised criteria when deciding the number of components so that clusters convey structure without overfitting (Schwarz, 1978).

This geometric picture is visible in Figure 4, where K-means partitions a dataset into three compact groups while ignoring subtle curvature. The same experiment, summarised by the silhouette curve in Figure 5, shows how a quick validity check can guide the number of partitions before any interpretation with domain experts (Rousseeuw, 1987; Caliński & Harabasz, 1974; Davies & Bouldin, 2009).



**Figure 4. K-means on a mixed-geometry synthetic dataset (K = 3): compact partitions despite non-convex structure**



**Figure 5. Silhouette score across K (2–8): a quick diagnostic for selecting the number of partitions**

When embodied states are uneven or interleaved, density-based discovery is usually more faithful to the data. DBSCAN identifies dense cores and leaves ambiguous points unlabelled, which is practical in in-the-wild logs where artefacts and rare episodes coexist (Ester, Kriegel, Sander, & Xu, 1996). HDBSCAN generalises to variable densities and selects clusters that are stable across scales, which tends to produce more consistent segments for speech prosody, micro-expressions, or EEG bursts (Campello, Moulavi, & Sander, 2013; McInnes, Healy, & Astels, 2017). The contrast is visible in **Figure 6**, where the non-convex moon-shaped structure emerges naturally under DBSCAN while noise points remain isolated.

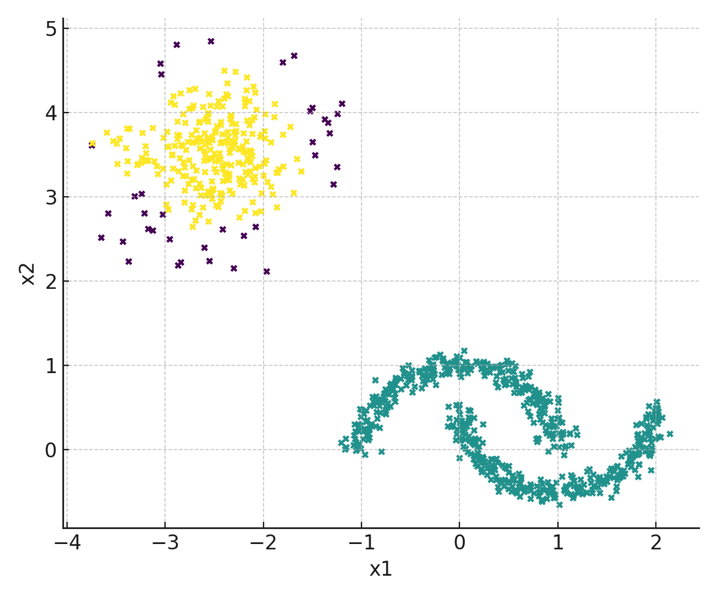
****

Figure 6. DBSCAN on the same dataset: recovery of non-convex clusters with explicit noise isolation

Graph-based perspectives are useful when latent affect lies on a manifold shaped by articulation, facial dynamics, or neural synchrony. Spectral clustering discovers communities from the Laplacian of a similarity graph and is known to recover non-linear boundaries that partitioning cannot capture (Ng, Jordan, & Weiss, 2001; von Luxburg, 2007). Closely related constructions turn neighbourhoods into coordinates. Laplacian Eigenmaps and Diffusion Maps embed samples along smooth trajectories, which often align with arousal ramps, regulation cycles, or appraisal shifts (Belkin & Niyogi, 2003; Coifman & Lafon, 2006). Kernel PCA remains a classical alternative when feature maps are easier to encode than explicit graphs (Schölkopf, Smola, & Müller, 1998). Modern practice frequently uses UMAP to visualise these neighbourhoods while delegating clustering to a separate step to avoid confounding geometry with grouping (McInnes, Healy, & Melville, 2018).

Deep clustering couples representation learning with grouping so that the feature space itself becomes more cluster-friendly. Deep Embedded Clustering refines an encoder using a soft assignment objective that pulls similar samples together and spreads clusters apart (Xie, Girshick, & Farhadi, 2016). Variational Deep Embedding combines a variational autoencoder with a mixture model to produce latent variables that are both generative and cluster-aware, an advantage when analysts later need counterfactual samples or uncertainty assessments (Jiang et al., 2016). In computer vision, self-labelling approaches use clusters as supervision to iteratively improve the backbone, which has proven effective for expression-like features in large, unlabelled corpora (Caron, Bojanowski, Joulin, & Douze, 2018; Caron, Misra, Mairal, Goyal, Bojanowski, & Joulin, 2020).

Temporal structure should be acknowledged because embodied emotion unfolds as episodes. A practical approach separates dynamics from geometry by detecting changepoints first and then clustering within segments, which reduces the burden on any single algorithm to capture both time and shape (Truong, Oudre, & Vayatis, 2020). Bayesian online schemes or optimal partitioning methods provide additional options when near real-time decisions are required (Adams & MacKay, 2007; Killick, Fearnhead, & Eckley, 2012). It has become central when data are sensitive or interventions must not degrade user well-being, since segmentation allows the system to abstain or defer decisions in ambiguous intervals (Farahani et al., 2024). Taken together, the families below serve complementary roles. Table 3 positions each algorithm in terms of assumptions, strengths, and caveats, which facilitates transparent preregistration. Table 4 lists internal indices that help select models without labels.

**Table 3. Algorithm families for embodied emotion discovery**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Family | Core assumption | Typical use in affect data | Strengths | Caveats |
| K-means | Globular clusters in Euclidean space | Clean, well-separated SSL embeddings | Fast and readable | Needs K and standardization (McQueen, 1967) |
| Gaussian mixtures (EM) | Mixture densities with soft membership | Overlapping states with gradual transitions | Uncertainty and flexible shapes | Risk of spurious components, select K with BIC (Dempster et al., 1977; Schwarz, 1978; Bishop & Nasrabadi) |
| DBSCAN | Dense regions separated by sparse areas | In-the-wild episodes with artifacts | No K, handles noise | Parameter sensitivity to scale (Ester et al., 1996) |
| HDBSCAN | Variable densities across clusters | Multimodal logs with rare states | Stable clusters across scales | Requires similarity choices (Campello et al., 2013; McInnes et al., 2017) |
| Spectral clustering | Manifold communities via Laplacian | Smooth affect trajectories | Non-linear boundaries | Graph construction is critical (Ng et al., 2001; von Luxburg, 2007) |
| Diffusion geometry | Heat-kernel neighbourhoods | Gradual affective progressions | Interpretable coordinates | Kernel and scale choices matter (Coifman & Lafon, 2006) |
| Deep clustering (DEC, VaDE) | Features and clusters co-learned | Large multimodal sets with weak labels | State-of-the-art transfer | Training stability and compute (Xie et al., 2016; Jiang et al., 2016) |
| Self-labeling vision (DeepCluster, SwAV) | Clusters drive backbone training | Expression features without labels | Scales to millions of samples | Careful augmentation design (Caron et al., 2018; Caron et al., 2020) |

**Table 4. Internal validity indices commonly used during method selection**

|  |  |  |
| --- | --- | --- |
| Index | What it measures | When it shines |
| Silhouette | Cohesion versus separation per point | Quick triage of K in partitions (Rousseeuw, 1987) |
| Calinski–Harabasz | Between to within dispersion ratio | Comparing many K on standardized features (Caliński & Harabasz, 1974) |
| Davies–Bouldin | Average similarity between clusters | Penalizing overlapping, diffuse clusters (Davies & Bouldin, 2009) |

**Table 5** distils pragmatic heuristics that align with real constraints in psychology labs and clinics. In translational settings, these choices can be integrated with supervised outcomes and theory, as demonstrated by recent cross-domain applications linking latent structure to complex constructs in mental health and interpersonal functioning (Farahani et al., 2025) and by methodologically careful pipelines used in predictive modelling (Kovač et al., 2025a).

**Table 5. Pragmatic selection heuristics**

|  |  |  |
| --- | --- | --- |
| Situation | Prefer | Rationale |
| SSL embeddings already separate well | K-means or GMM | Simpler models suffice; soft labels enable graded states |
| Many outliers and rare episodes | HDBSCAN | Labels noise explicitly and adapts to variable density |
| Non-linear boundaries and smooth progressions | Spectral or diffusion geometry | Captures manifold structure and temporal continuity |
| Very large unlabelled corpora | DeepCluster or SwAV | Use clusters as supervision to shape the encoder |
| Long sequences with regime shifts | Changepoint plus within-segment clustering | Factors dynamics from geometry for clearer interpretation (Truong et al., 2020) |

9.4 Embodied Emotion Detection by Unsupervised EEG Signal Analysis

Unsupervised discovery with EEG starts long before clustering: the physics of the sensor, the low SNR, and cross-subject variability make early, standardised preprocessing a scientific necessity. A practical baseline is to combine automated referencing and line-noise handling with artefact separation, keeping every step reproducible and reportable (Bigdely-Shamlo, Mullen, Kothe, Su, & Robbins, 2015; Gramfort et al., 2013). In practice, this means PREP-style early processing, then common average reference (or a robust variant), notch and band-pass filtering, and ICA for ocular and myogenic components, all logged in a pipeline that can be re-run on future cohorts.

The following Python code illustrates the minimal MNE-centric preprocessing for unsupervised learning. This code captures a reproducible baseline, filtering, average referencing, ICA cleanup, and short overlapping epochs, without entangling discovery with labels (Bigdely-Shamlo et al., 2015; Gramfort et al., 2013).

# Goal: produce clean epochs ready for unsupervised features.

import mne

raw = mne.io.read\_raw\_fif("subject\_raw.fif", preload=True)

raw.notch\_filter([50, 100]) # or 60 Hz depending on locale

raw.filter(1., 45., phase="zero-double") # affect-friendly passband

raw.set\_eeg\_reference("average") # robust CAR

ica = mne.preprocessing.ICA(n\_components=20, random\_state=97)

ica.fit(raw.copy().filter(1., None))

ica.detect\_artifacts(raw, eog\_ch=["HEO","VEO"]) # or use correlation templates

raw\_clean = ica.apply(raw)

events = mne.make\_fixed\_length\_events(raw\_clean, duration=2.0, overlap=1.0)

epochs = mne.Epochs(raw\_clean, events, tmin=0.0, tmax=2.0, baseline=None, preload=True)

# 'epochs' is now ready for spectral, covariance, or SSL embeddings.

After cleaning, the aim is a representation that respects EEG’s spatial structure while remaining friendly to unsupervised geometry and density. Two families dominate. The first summarises spectral content to capture arousal-like and engagement-like variation without labels. The second summarises spatial structure via channel covariance: symmetric positive definite (SPD) matrices that encode how sensors co-fluctuate; mapping these matrices to a Riemannian tangent space yields Euclidean features amenable to clustering and manifold learning (Barachant, Bonnet, Congedo, & Jutten, 2012; Yger, Lotte, & Sugiyama, 2016). Tooling such as *pyRiemann* standardises this SPD-to-tangent workflow so you can reuse it across studies with minimal code (Barachant et al., 2022).

Connectivity features add a complementary lens when you expect coordinated network behaviour during emotion induction. Phase-based measures such as the phase-locking value (PLV) or phase slope index summarise frequency-specific synchrony; coherence variants capture magnitude-phase relations. These tools are powerful but require careful interpretation because volume conduction, filtering, and reference choices can bias apparent coupling (Bastos & Schoffelen, 2016; Chiarion, Artoni, & Makeig, 2023). For unsupervised discovery, a pragmatic approach is to compute a compact panel, e.g., PLV in theta/alpha between a handful of regions of interest, and concatenate with bandpower or tangent-space features before clustering. A small panel (log-bandpower per channel and a targeted PLV) is shown in the Python code below. It offers an interpretable structure that plays well with unsupervised algorithms, while acknowledging the interpretational caveats of connectivity in EEG (Bastos & Schoffelen, 2016; Chiarion et al., 2023).

# Goal: compute per-epoch bandpower and a small PLV panel between ROIs.

import numpy as np

from scipy.signal import welch, hilbert

sfreq = epochs.info["sfreq"]

data = epochs.get\_data() # (n\_epochs, n\_channels, n\_times)

# 1) Log-bandpower (theta 4-8 Hz, alpha 8-12 Hz, beta 13-30 Hz)

def bandpower(x, fs, fmin, fmax):

f, Pxx = welch(x, fs=fs, nperseg=int(fs\*1.0))

idx = (f >= fmin) & (f <= fmax)

return np.log(Pxx[..., idx].mean())

bp = []

for ep in data:

bp.append([

bandpower(ep[ch], sfreq, 4, 8),

bandpower(ep[ch], sfreq, 8, 12),

bandpower(ep[ch], sfreq, 13, 30)

] for ch in range(ep.shape[0]))

bp = np.array([np.ravel(x) for x in bp])

# 2) Simple PLV between two ROI averages in alpha band

roi1, roi2 = [0,1,2,3], [20,21,22,23]

def band\_hilbert(ep, fs, fmin, fmax):

# narrowband via FFT or IIR (omitted for brevity), then analytic signal

x = ep # assume prefiltered

return hilbert(x, axis=-1)

plvs = []

for ep in data:

s1 = ep[roi1].mean(axis=0)

s2 = ep[roi2].mean(axis=0)

a1 = band\_hilbert(s1, sfreq, 8, 12)

a2 = band\_hilbert(s2, sfreq, 8, 12)

phase\_diff = np.angle(a1) - np.angle(a2)

plv = np.abs(np.exp(1j\*phase\_diff)).mean()

plvs.append(plv)

# Concatenate bandpower and PLV into a compact feature vector per epoch.

features = np.column\_stack([bp, np.array(plvs)[:, None]])

A second, intrinsically unsupervised tradition in EEG, microstate analysis, segments continuous data into brief, quasi-stable topographic patterns and studies their sequence statistics. Microstates act like a symbolic alphabet for the brain’s ongoing dynamics and can be related, post hoc, to affective conditions or regulation episodes without ever using labels during discovery (Michel & Koenig, 2018; Haydock, Custo, & Brunet, 2025). While microstate modelling is not a drop-in replacement for feature-space clustering, it offers a complementary viewpoint when the temporal grammar of states is central.

Recent advances show that self-supervised learning (SSL) can pretrain EEG encoders on large unlabelled recordings using masked prediction or contrastive tasks, then expose low-dimensional embeddings for clustering. The effect is practical: cleaner separations and better cross-subject transfer before any labels are introduced (Zhang, Ma, Xiao, & Wang, 2024; see also newer spatial–temporal SSL variants). In effect, these studies allow you to discover latent EEG states, then relate them to behavioural markers, questionnaires, or physiological covariates in a second stage.

Algorithm choice should match the EEG geometry. When tangent-space or SSL embeddings already look well separated, K-means or Gaussian mixtures are adequate and easy to report. When the state space contains rare bursts, artefacts, or uneven dwell times, HDBSCAN excels by finding dense cores while leaving ambiguous samples unlabelled. This is crucial when you prefer abstention over forced assignment in sensitive applications (McInnes, Healy, & Astels, 2017). In all cases, cluster in the full feature space; use 2-D projections only for inspection.

Covariance lives on an SPD manifold, while tangent-space mapping preserves geometry while enabling standard clustering. HDBSCAN labels dense EEG states and leaves borderline epochs as noise, which is desirable when you’d rather defer decisions than over-commit (Barachant et al., 2012; McInnes et al., 2017). Illustrative Python code for mapping epoch covariance to tangent space, embedding, then clustering without labels by HDBSCAN is as follows:

import numpy as np

from pyriemann.estimation import Covariances

from pyriemann.tangentspace import TangentSpace

from sklearn.preprocessing import StandardScaler

from sklearn.manifold import TSNE

import hdbscan

# epochs.get\_data() -> shape (n\_epochs, n\_channels, n\_times)

X = epochs.get\_data()

covs = Covariances(estimator="oas").transform(X)

TS = TangentSpace().fit(covs).transform(covs) # Euclidean features

Z = StandardScaler().fit\_transform(TS)

# 2-D inspection only (cluster in full Z ideally)

Z2 = TSNE(n\_components=2, perplexity=30, random\_state=7).fit\_transform(Z)

labels = hdbscan.HDBSCAN(min\_cluster\_size=25).fit\_predict(Z) # abstains on ambiguous points (-1)

Finally, keep the bridge to downstream psychology explicit. Latent EEG states, cluster posteriors, distances to centroids, or dwell-time summaries, can enter regression or classification models as structured covariates. This mirrors transparent modelling practices already used in supervised work, while keeping discovery label-agnostic at the outset (Kovač et al., 2025a).

9.5 Embodied Emotion Detection by Unsupervised Image Analysis

Unsupervised image analysis seeks affective structure directly from frames, without relying on expression labels, and then links the discovered states to behaviour or context. The practical starting point is a label-free visual backbone whose embeddings already emphasise content over nuisance factors like pose or lighting. In recent practice, DINOv2 offers robust, out-of-the-box image features that transfer well across datasets and can be clustered with minimal tuning (Oquab et al., 2024).

A persistent challenge is identity leakage: when the embedding space inadvertently clusters by who the person is rather than how they feel. Expression-specific self-supervision helps reduce this problem by shaping features so that local neighbourhoods reflect facial actions and not identity style (Shu et al., 2022). In that spirit, a pragmatic recipe is to extract general features with a strong universal backbone and then adapt with a light, expression-focused self-supervised objective if identity cues dominate (Shu et al., 2022).

Discovery should be divorced from evaluation. For images *in the wild*, AffectNet is a natural target for validation because it spans both categorical labels and continuous valence–arousal ratings at scale (Mollahosseini, Hasani, & Mahoor, 2017). For broader diversity and spontaneous footage, Aff-Wild2 adds long, unconstrained videos annotated in continuous space, and it is routinely used for cross-database tests after discovery (Kollias & Zafeiriou, 2018, 2019). RAF-DB complements these with curated, real-world faces and well-documented splits (Li, Deng & Du, 2017). Using these corpora for evaluation only keeps the unsupervised stage clean while offering rigorous checks once clusters are found.

When analysts need interpretable anchors, action units (AUs) remain valuable because they provide a mechanistic lens on facial muscle activations. OpenFace 2.0 exposes AUs, landmarks, and head pose with a transparent pipeline; these signals can be summarised per cluster to name states post hoc without turning the discovery step into supervised learning (Baltrušaitis et al., 2016).

Clustering choices follow geometry. If embeddings already separate well, compact partitions are defensible and easy to report; when data contain rare episodes or ambiguous frames, HDBSCAN tends to recover stable cores and abstain elsewhere, which is often preferable in sensitive applications (McInnes, Healy, & Astels, 2017). For visualisation only, UMAP reveals neighbourhoods without imposing linearity; the clustering itself should occur in the full embedding space to avoid projection artefacts (McInnes, Healy, & Melville, 2018).

Fairness and privacy belong in the design loop. Facial datasets are not demographically neutral, and model performance can vary by age, gender, or ethnoracial group. Balanced attribute datasets such as FairFace can be used to assess representational balance of discovered states or to audit downstream classifiers trained on pseudo-labels (Karkkainen & Joo, 2021). Recent studies document bias patterns specifically in facial expression recognition and provide templates for measurement and mitigation. These steps should be acknowledged even in unsupervised workflows (Sajjad et al., 2023; Domínguez-Catena et al., 2024).

Methodologically, the reporting discipline you apply in supervised psychology (clear preprocessing, explicit hyperparameters, and transparent diagnostics) transfers directly to this unsupervised image setting. That consistency keeps the path from latent image states to psychological interpretation defensible (Kovač et al., 2024). Next code walks through a compact, label-free pipeline: it subsamples video frames, extracts embeddings with a universal backbone (DINOv2), clusters the full-dimensional features with a density method that can abstain on ambiguous frames (HDBSCAN), and then renders a 2-D neighbourhood map purely for inspection with UMAP, while keeping visualization separate from clustering to avoid projection artifacts (Oquab et al., 2024; McInnes, Healy, & Astels, 2017; McInnes, Healy, & Melville, 2018).

# Goal: extract frame embeddings with a universal backbone, cluster without labels,

# and inspect neighborhoods in 2-D (clustering should happen in full space).

# Requires: pip install timm umap-learn hdbscan opencv-python

import cv2, glob, numpy as np, torch, timm

from sklearn.preprocessing import StandardScaler

import hdbscan

import umap

# 1) Load frames (e.g., one every 8–10 frames to reduce redundancy)

def sample\_frames(video\_path, stride=10, max\_frames=512):

cap, frames = cv2.VideoCapture(video\_path), []

i = 0

while cap.isOpened() and len(frames) < max\_frames:

ret, frame = cap.read()

if not ret: break

if i % stride == 0:

frames.append(cv2.cvtColor(frame, cv2.COLOR\_BGR2RGB))

i += 1

cap.release()

return frames

frames = sample\_frames("session.mp4", stride=10)

# 2) Create DINOv2 model from timm (choose a small/medium ViT for speed)

device = "cuda" if torch.cuda.is\_available() else "cpu"

model = timm.create\_model("vit\_small\_patch14\_dinov2.lvd142m", pretrained=True).to(device).eval()

transform = timm.data.transforms\_factory.create\_transform(

input\_size=model.default\_cfg["input\_size"][1], is\_training=False

)

# 3) Compute L2-normalized embeddings

with torch.inference\_mode():

feats = []

for img in frames:

x = transform(img).unsqueeze(0).to(device)

z = model.forward\_features(x)

z = torch.nn.functional.normalize(z.mean(dim=1), dim=1) # global average

feats.append(z.squeeze(0).cpu().numpy())

X = np.vstack(feats)

# 4) Unsupervised clustering (density-friendly; abstains on noise)

Z = StandardScaler().fit\_transform(X)

labels = hdbscan.HDBSCAN(min\_cluster\_size=30).fit\_predict(Z)

# 5) 2-D view for inspection only (do not cluster in 2-D)

U = umap.UMAP(n\_neighbors=20, min\_dist=0.05, random\_state=7).fit\_transform(Z)

# Use U[:,0], U[:,1] to make a scatter plot colored by 'labels'.

To use this code, replace session.mp4 with your clip or folder of images. The clustering result labels gives an integer per frame (with -1 for ambiguous frames), which you can map back to timestamps to visualise episodes. The UMAP coordinates are for sanity-check plots only.

Identity cues can dominate unsupervised image embeddings. Thus, it helps to neutralise static, person-specific style before discovery. Next listing implements a simple within-subject centring step: subtract each individual’s mean embedding from their frames, preserving relative expression dynamics while dampening identity and background signatures. In practice, this reduces *who you are* variance and sharpens *how you feel* structure, improving density-based discovery in multi-person corpora; afterwards, you can recover interpretability by summarising cluster-wise action units (AUs) rather than reintroducing labels (Shu et al., 2022; Baltrušaitis et al., 2016). Because demographic imbalance can still imprint on representation quality, auditing clusters across balanced attribute benchmarks remains advisable even with centring (Karkkainen & Joo, 2021; Sajjad et al., 2023).

# Goal: reduce identity leakage by removing each person's embedding mean before clustering.

# Assumes you have person IDs per frame (e.g., from a tracker or face re-ID).

import numpy as np

from sklearn.preprocessing import StandardScaler

import hdbscan

# X: (n\_frames, d) embeddings; person\_id: (n\_frames,) integers

def within\_subject\_centering(X, person\_id):

Xc = X.copy()

for pid in np.unique(person\_id):

idx = np.where(person\_id == pid)[0]

Xc[idx] -= X[idx].mean(axis=0, keepdims=True)

return Xc

Xc = within\_subject\_centering(X, person\_id) # identity-mean removed

Z = StandardScaler().fit\_transform(Xc)

labels = hdbscan.HDBSCAN(min\_cluster\_size=30).fit\_predict(Z)

When you have multiple people in the same recording, subtracting the per-person mean neutralises static identity style while preserving relative expression changes. If AUs are available (e.g., via OpenFace 2.0), compute AU rates per cluster to name the states afterwards without injecting labels into discovery (Baltrušaitis et al., 2016).

9.6 Embodied Emotion Detection by Unsupervised Text Data Analysis

Text offers a rich substrate for embodied emotion because language co-occurs with behaviour, physiology, and context; yet labels are costly and inconsistent. An unsupervised route is therefore attractive: build semantically faithful sentence embeddings, let geometry or density reveal latent themes, and only then map those themes into affective space (e.g., valence–arousal–dominance) for interpretation. In practice, sentence-level encoders such as Sentence-BERT, SimCSE, INSTRUCTOR, and E5 produce robust vectors that transfer across domains without task-specific labels. This is an essential property for diaries, counselling transcripts, or social streams (Reimers & Gurevych, 2019; Gao, Yao, & Chen, 2021; Su et al., 2022; Wang et al., 2022).

Once embedded, discovery typically proceeds with modern topic models that **do not** require a preset number of topics and work directly in the continuous space. BERTopic clusters embeddings (often with HDBSCAN/UMAP under the hood) and then uses class-based TF–IDF to label topics; Top2Vec jointly organises words, documents, and topic vectors, letting the corpus determine topic counts automatically (Grootendorst, 2022; Angelov, 2020). Parameter choices (e.g., HDBSCAN min\_cluster\_size, UMAP n\_neighbors) control granularity and interpretability; small adjustments can turn *too many micro-topics* into *stable, nameable themes* (Grootendorst, 2022).

To relate unlabelled themes to affect without turning the pipeline into supervision, a practical step is to project topics into the VAD space. Word-level norms such as Warriner–Kuperman–Brysbaert (2013) and the NRC VAD lexicon provide continuous scores for large vocabularies. Aggregating those over each topic’s salient terms yields a rough affective fingerprint useful for naming or ordering topics, e.g., *high arousal / low valence* (Warriner et al., 2013; Mohammad, 2025). This post hoc mapping does not supply labels during discovery; it only guides interpretation and reporting.

When sequences matter, dynamic topic modelling supplies temporal structure, tracking how themes drift, split, or merge over time. Classical Dynamic Topic Models (DTM) handle discrete time slices; continuous-time variants and neural extensions incorporate word embeddings or networks to model smooth semantic evolution (Blei & Lafferty, 2006; Wang et al., 2012; Dieng et al., 2020; Zhang et al., 2022). In practice, many studies obtain a strong baseline by estimating static topics on a rolling window, then plotting topic prevalence as a function of time for exploratory inspection.

Fairness and privacy obligations remain, even when labels are absent. Surveys in NLP caution that *bias* is multidimensional and demands explicit normative framing; unsupervised clusters can still encode or amplify representational harms if sources are skewed (Blodgett, Barocas, Daumé III, & Wallach, 2020). Broader critiques of large-scale text modelling emphasise documentation, dataset governance, and harm-aware evaluation, and other considerations that apply equally when topics become pseudo-labels for downstream prediction (Bender et al., 2021; Gallegos et al., 2024; Hovy & Prabhumoye, 2021). It has become central when data are sensitive or interventions must not degrade user well-being, so abstention, aggregation, and conservative naming are good defaults (Blodgett et al., 2020). Finally, keep the bridge to outcomes explicit: topic memberships, distances to topic centroids, and VAD fingerprints become structured predictors in transparent psychological analyses, aligning with reporting practices your group already uses for supervised models (Kovač et al., 2024). This continuity lets latent text structure inform theory without retrofitting labels post hoc.

To make the text pipeline concrete, the following code assembles a fully label-free workflow: it first encodes each utterance with a domain-robust sentence transformer (e.g., SBERT) to preserve semantics without supervision (Reimers & Gurevych, 2019), then induces topics using BERTopic, which couples density clustering and neighbourhood preservation via HDBSCAN and UMAP for stable, nameable themes (Grootendorst, 2022; McInnes, Healy, & Astels, 2017; McInnes, Healy, & Melville, 2018). Finally, it assigns each discovered topic an affective fingerprint by projecting its salient terms into valence–arousal–dominance space using lexicons such as Warriner–Kuperman–Brysbaert or NRC-VAD; crucially, this VAD step is post hoc and guides interpretation without supplying labels during discovery (Warriner, Kuperman, & Brysbaert, 2013; Mohammad, 2025).

# pip install sentence-transformers bertopic umap-learn hdbscan pandas

from sentence\_transformers import SentenceTransformer

from bertopic import BERTopic

from sklearn.metrics.pairwise import cosine\_distances

import pandas as pd

# 1) Load texts: one row per message/utterance/entry (with optional timestamp)

df = pd.read\_csv("corpus.csv") # columns: ["text", "timestamp"]

texts = df["text"].fillna("").tolist()

# 2) Embed with a strong general-purpose model (swap for SimCSE/INSTRUCTOR/E5 as needed)

embedder = SentenceTransformer("all-MiniLM-L6-v2")

X = embedder.encode(texts, batch\_size=64, show\_progress\_bar=True, normalize\_embeddings=True)

# 3) Unsupervised topics (automatic count). For coarser themes: increase min\_cluster\_size

topic\_model = BERTopic(min\_topic\_size=40, calculate\_probabilities=True, verbose=True)

topics, probs = topic\_model.fit\_transform(texts, X)

# 4) Get representative words per topic

topic\_info = topic\_model.get\_topic\_info() # topic sizes and names

topic\_words = {t: [w for w, \_ in topic\_model.get\_topic(t)[:20]] # top-20 words

for t in topic\_info["Topic"] if t != -1}

# 5) Post hoc VAD mapping: merge topic words with a VAD lexicon (Warriner or NRC VAD)

vad = pd.read\_csv("vad\_lexicon.csv") # columns: word,valence,arousal,dominance

vad["word"] = vad["word"].str.lower()

def topic\_vad(words):

m = vad[vad["word"].isin([w.lower() for w in words])]

return m[["valence","arousal","dominance"]].mean()

topic\_vads = {t: topic\_vad(ws) for t, ws in topic\_words.items()}

topic\_vads = pd.DataFrame(topic\_vads).T # index=topic, columns=V/A/D

# 6) (Optional) distance-to-topic-centroids as features for downstream models

centroids = topic\_model.c\_tf\_idf\_ # or compute mean embedding per topic

Sentence embeddings preserve semantics without labels; BERTopic groups them and generates interpretable descriptors; VAD projection provides an affective *fingerprint* for each latent theme without supervising discovery (Reimers & Gurevych, 2019; Grootendorst, 2022; Warriner et al., 2013).

To trace how unlabelled themes ebb and flow in real data, the following Python code turns static topics into a time series: it bins documents by period (e.g., weeks), normalizes within each bin to obtain topic prevalence, and optionally joins these trajectories with each topic’s valence–arousal–dominance fingerprint from the code for affect-aware plots. This rolling, label-free view is a transparent baseline that often suffices before fitting full dynamic topic models, which introduce temporal priors to capture smooth evolution or regime shifts (Blei & Lafferty, 2006; Wang et al., 2012). When abrupt changes are suspected (e.g., around interventions), the same prevalence series can be paired with standard changepoint analyses for exploratory checks before confirmatory modelling (Truong, Oudre, & Vayatis, 2020).

# Assume you have 'topics' from Listing 9.5.1 and a datetime column 'timestamp'

df["topic"] = topics

df["date"] = pd.to\_datetime(df["timestamp"]).dt.to\_period("W") # weekly bins

# 1) Topic prevalence timeline

timeline = (df[df["topic"]!=-1]

.groupby(["date","topic"])

.size()

.groupby(level=0)

.apply(lambda x: x / x.sum()) # normalize per period

.reset\_index(name="prevalence"))

# 2) Merge with topic VAD for annotated plots or regime-change analysis

timeline = timeline.merge(topic\_vads.reset\_index().rename(columns={"index":"topic"}),

on="topic", how="left")

# Now plot prevalence vs. time per topic, or summarize high-arousal themes across weeks.

While full DTM provides principled temporal priors, a rolling-window prevalence view offers a strong, transparent baseline; it often suffices for exploratory analyses before fitting dynamic models (Blei & Lafferty, 2006; Zhang et al., 2022).

9.7 Embodied Emotion Detection by Unsupervised Audio Signal Analysis

Unsupervised audio analysis targets structure in vocal behaviour without relying on rater-provided emotion tags, which is essential in naturalistic settings where labels are sparse, subjective, or ethically sensitive. The modern recipe is straightforward: ensure signal hygiene, extract embeddings that capture paralinguistic cues, let geometry or density expose latent states, and only then examine how those states relate to context or outcomes. Self-supervised encoders trained on large unlabelled corpora, such as *wav2vec 2.0*, *HuBERT*, and *WavLM*, supply rich utterance-level representations that transfer to affect with minimal adaptation, while emotion-specific pretraining such as *emotion2vec* can sharpen separation among arousal- and valence-related acoustics (Baevski, Zhou, Mohamed, & Auli, 2020; Hsu et al., 2021; Chen et al., 2022; Zhang, Ma, Xiao, & Wang, 2024). Classic acoustic summaries (e.g., MFCCs, energy, and pitch trajectories) remain valuable for interpretability and complement neural embeddings during discovery (Eyben et al., 2015).

Before discovery, the research should resample to a consistent rate (e.g., 16 kHz), normalise loudness, remove long silences with voice activity detection, and diarise when multiple speakers are present. Unsupervised segmentation helps disentangle state changes from speaker turns: spectral-clustering diarization and offline changepoint detection frequently suffice to carve the stream into interpretable units without labels (Park et al., 2019; Truong, Oudre, & Vayatis, 2020). In settings involving sensitive populations, abstaining from decisions in ambiguous segments (e.g., by labelling them as *noise* in density clustering) reduces the risk of spurious inferences from poor signal quality.

Embeddings from *WavLM*, *HuBERT*, or *wav2vec 2.0* summarise timbre, articulation, and prosodic contours; they are robust to lexical content and often cluster by speaking style or state even before fine-tuning (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). For broader non-speech affect (e.g., laughter, sighs, crying), audio-tagging backbones such as PANNs, BYOL-A, and OpenL3 provide complementary cues that capture nonverbal events absent from ASR-focused encoders (Kong, Chen, & Wang, 2020; Niizumi et al., 2021; Cramer, Wu, Salamon, & Bello, 2019). Pooling choices matter: mean–std pooling is a strong default; attention pooling can emphasise high-energy or high-F0 regions typical of heightened arousal.

If SSL embeddings already separate cleanly, K-means or Gaussian mixtures provide simple, reportable partitions. Where recordings contain rare vocalisations, overlaps, or artefacts, HDBSCAN isolates dense cores and marks ambiguous segments as noise. This is a desirable behaviour when abstention is safer than forced assignment (Rousseeuw, 1987; McInnes, Healy, & Astels, 2017). Two-dimensional projections (e.g., UMAP) are useful for sanity checks, but clusters should be computed in the full space to avoid projection artefacts. After unsupervised states are found, posterior memberships, dwell times, and transition statistics can enter transparent outcome models (Kovač et al., 2025a; Kovač et al., 2024). This separation (discovery first, linkage later) keeps the pipeline label-agnostic at the representation stage while preserving interpretability at the analysis stage.

To operationalise the audio workflow, the following code implements a compact, label-free pipeline: it first segments the waveform into short, overlapping windows; then extracts self-supervised speech embeddings (illustrated with WavLM, but wav2vec 2.0 or HuBERT are drop-in alternatives) so paralinguistic cues are captured without transcripts (Chen et al., 2022; Baevski et al., 2020; Hsu et al., 2021). The embeddings are standardised and passed to HDBSCAN, which uncovers stable affective episodes while abstaining on ambiguous segments rather than forcing assignments (McInnes, Healy, & Astels, 2017). A UMAP map is generated only for neighbourhood inspection and produces a visualisation that remains strictly separate from clustering to avoid projection artefacts (McInnes, Healy, & Melville, 2018).

# pip install librosa torch torchaudio transformers hdbscan umap-learn

import librosa, numpy as np

from transformers import WavLMModel, WavLMConfig, WavLMProcessor

from sklearn.preprocessing import StandardScaler

import hdbscan, umap

# 1) Load audio and segment (short fixed windows with overlap work well)

y, sr = librosa.load("session.wav", sr=16000, mono=True)

frames = librosa.util.frame(y, frame\_length=int(1.6\*sr), hop\_length=int(0.8\*sr)).T # 1.6s windows, 50% overlap

# 2) SSL embeddings (WavLM Base as an example)

processor = WavLMProcessor.from\_pretrained("microsoft/wavlm-base")

model = WavLMModel.from\_pretrained("microsoft/wavlm-base").eval()

def embed(wave):

inputs = processor(wave, sampling\_rate=16000, return\_tensors="pt", padding=True)

with torch.inference\_mode():

h = model(\*\*inputs).last\_hidden\_state # (1, T, D)

# Mean pooling over time for an utterance-level embedding

return h.mean(dim=1).squeeze(0).numpy()

X = np.vstack([embed(f) for f in frames])

# 3) Standardize and cluster (density; abstains on ambiguous segments with label -1)

Z = StandardScaler().fit\_transform(X)

labels = hdbscan.HDBSCAN(min\_cluster\_size=25).fit\_predict(Z)

# 4) Optional 2-D map for inspection (do not cluster in 2-D)

U = umap.UMAP(n\_neighbors=20, min\_dist=0.05, random\_state=7).fit\_transform(Z)

# Plot U[:,0], U[:,1] colored by 'labels' to inspect neighborhoods.

SSL embeddings summarise paralinguistic cues without transcript labels; HDBSCAN surfaces stable affective *episodes* and leaves the rest unlabelled, which is safer in noisy or clinical audio (Chen et al., 2022; McInnes et al., 2017).

To emphasise prosodic structure before any representation learning, the next Python code builds episodes directly from low-level acoustics: it extracts frame-level fundamental frequency (F0) and loudness (RMS), detects changepoints in the 2-D prosody stream to carve the audio into coherent segments, then pools simple statistics (mean, variance) per segment and clusters them (e.g., with K-means). This prosody-first route is intentionally transparent: F0 tracks tension and vocal effort while RMS indexes energy, giving segments that map naturally onto arousal-related interpretations without relying on transcripts or labels (Eyben et al., 2015). Using a principled F0 estimator such as YIN stabilises pitch in noisy, spontaneous speech (De Cheveigné & Kawahara, 2002), and offline changepoint detection separates regime shifts (e.g., calm → excited) from mere local fluctuations (Truong, Oudre, & Vayatis, 2020). By decoupling segmentation from clustering, the method avoids asking one model to learn both dynamics and geometry at once, and it supports cautious inference, so segments with ambiguous prosody can be left unclustered or flagged for review.

# pip install librosa ruptures scikit-learn

import librosa, numpy as np, ruptures as rpt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

y, sr = librosa.load("session.wav", sr=16000, mono=True)

# 1) Frame-level prosody: pitch (F0) and loudness proxy

f0 = librosa.yin(y, fmin=50, fmax=400, sr=sr, frame\_length=2048, hop\_length=160) # ~10 ms hop

rms = librosa.feature.rms(y=y, frame\_length=2048, hop\_length=160).squeeze()

# 2) Changepoints on a compact 2-D prosody series

X = np.vstack([librosa.util.normalize(np.nan\_to\_num(f0)), librosa.util.normalize(rms)]).T

algo = rpt.KernelCPD(kernel="rbf").fit(X)

bps = algo.predict(pen=10) # breakpoints (indices in frames)

# 3) Pool features within segments, then cluster

segs = [ (0, bps[0]) ] + [ (bps[i], bps[i+1]) for i in range(len(bps)-1) ]

feat = []

for s, e in segs:

fseg = X[s:e]

feat.append(np.hstack([fseg.mean(axis=0), fseg.std(axis=0)])) # mean/std of F0, RMS

feat = np.vstack(feat)

Z = StandardScaler().fit\_transform(feat)

labels = KMeans(n\_clusters=3, n\_init=20, random\_state=0).fit\_predict(Z)

# 'labels' marks prosody-defined episodes that you can align with time, context, or physiology.

A prosody-first approach makes the segmentation step explicit, reducing pressure on clustering to capture both dynamics and geometry (Truong et al., 2020). It also yields interpretable features (pitch/energy) that map naturally to arousal-related constructs.

To capture non-speech affect (laughter, sobs, gasps, sighs, breathiness) next listing swaps ASR-oriented SSL encoders for audio-tagging backbones that are pre-trained to recognise diverse acoustic events. The pipeline extracts event-aware embeddings from models such as PANNs, BYOL-A, or OpenL3, pools them at the segment level, standardises, and then applies HDBSCAN so dense, recurrent vocal events emerge as unsupervised states while ambiguous segments are left unlabelled. This complements speech encoders by highlighting affective paralinguistic cues that carry a strong emotional signal even in the absence of words (Kong, Chen, & Wang, 2020; Niizumi et al., 2021; Cramer, Wu, Salamon, & Bello, 2019).

# pip install torchlibrosa pytorch-lightning

# Example with PANNs-like features (pseudocode; replace with an actual pretrained tagger)

import torch, numpy as np

from sklearn.preprocessing import StandardScaler

import hdbscan

class TaggerBackbone(torch.nn.Module):

def \_\_init\_\_(self): ...

def forward(self, x): ... # returns (batch, time, d)

model = TaggerBackbone().eval()

# Assume 'spec' is a log-mel spectrogram of shape (time, mel)

with torch.inference\_mode():

h = model(torch.from\_numpy(spec)[None, None, ...]) # (1, T, D)

X = h.mean(dim=1).squeeze(0).numpy() # event-aware embedding

Z = StandardScaler().fit\_transform(X)

labels = hdbscan.HDBSCAN(min\_cluster\_size=20).fit\_predict(Z)

Audio-tagging backbones capture non-speech events (e.g., laughter, sobs, gasps) that are central to embodied emotion yet underrepresented in ASR-focused SSL features (Kong et al., 2020; Niizumi et al., 2021; Cramer et al., 2019). These findings indicate that the selection of model backbone functions is not a neutral engineering decision but is rather an implicit theoretical stance regarding which acoustic phenomena are treated as meaningful signals in human audio. Audio‑tagging encoders, typically trained on broad sound ontologies such as AudioSet (Gemmeke et al., 2017), allocate substantial representational capacity to short, high‑arousal non‑verbal human vocalisations. As a result, embodied affective cues, including laughter, sobbing, gasping, and sighing, tend to be preserved within their latent spaces (Cramer et al., 2019; Kong et al., 2020). In contrast, ASR‑oriented SSL encoders are optimised primarily for lexical robustness. Non‑speech intervals are frequently collapsed into generic silence or background noise categories and thus contribute minimally to the self‑supervised objective (Baevski et al., 2020; Hsu et al., 2021). Consequently, emotionally salient non‑lexical vocal events are systematically underrepresented. For downstream emotion‑recognition tasks, this pattern implies that systems relying exclusively on ASR‑derived SSL features may inherit a structural linguistic bias and overlook embodied, non‑verbal affective information, whereas audio‑tagging backbones provide a more appropriate computational substrate for capturing the broader spectrum of human emotional expression.

References

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

Barachant, A., Barthélemy, Q., King, J. R., Gramfort, A., Chevallier, S., Rodrigues, P. L. C., ... & Corsi, M. C. (2022). pyRiemann/pyRiemann: v0. 5. *Zenodo, Jul*.

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering, 59*(4), 920–928.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Barsade, S. G., & Gibson, D. E. (2012). Group affect: Its influence on individual and group outcomes. *Current Directions in Psychological Science*, *21*(2), 119-123.

Bastos, A. M., & Schoffelen, J. M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, *9*, 175.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, *15*(6), 1373-1396.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in neuroinformatics*, *9*, 16.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: Springer.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120).

Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, *1*(1), 18-37.

Campello, R. J., Kröger, P., Sander, J., & Zimek, A. (2020). Density‐based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(2), e1343.

Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 160-172). Berlin, Heidelberg: Springer Berlin Heidelberg.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 132-149).

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, *33*, 9912-9924.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505-1518.

Chiarion, G., Sparacino, L., Antonacci, Y., Faes, L., & Mesin, L. (2023). Connectivity analysis in EEG data: a tutorial review of the state of the art and emerging trends. *Bioengineering*, *10*(3), 372.

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, *21*(1), 5-30.

Cramer, A. L., Wu, H. H., Salamon, J., & Bello, J. P. (2019, May). Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3852-3856). IEEE.

Davies, D. L., & Bouldin, D. W. (2009). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917-1930.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1-22.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439-453.

Domínguez-Catena, I., Paternain, D., Jurio, A., & Galar, M. (2024). Less can be more: representational vs. stereotypical gender bias in facial expression recognition. *Progress in Artificial Intelligence*, *14*(1), 11-31.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190-202.

Farahani, H., Watson, P., Bezan, T., Kovač, N., Winter, L. C., Blagojević, M., … Jiménez, P. (2024). PsAIchology: An intelligent direction in psychological sciences. In 10th International Scientific Conference Technics, Informatics and Education—TIE 2024. Faculty of Technical Sciences Čačak, University of Kragujevac.

Farahani, H., Watson, P., Kovač, N., Sheykhangafshe, F. B., Azadfallah, P., Allahyari, A., … Chesli, R. R. (2025). Unraveling the complexity of love addiction using machine learning algorithms: The influence of positive and negative affect, interpersonal needs, and self-hate. In International Handbook of Love: Transcultural and Transdisciplinary Perspectives (pp. 1–22). Springer Nature Switzerland.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097-1179.

Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776-780). IEEE.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroinformatics*, *7*, 267.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Haydock, D., Kadir, S., Leech, R., Nehaniv, C. L., & Antonova, E. (2025). EEG microstate syntax analysis: A review of methodological challenges and advances. *NeuroImage*, 121090.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and linguistics compass*, *15*(8), e12432.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451-3460.

Jaramillo-Jimenez, A., Tovar-Rios, D. A., Mantilla-Ramos, Y. J., Ochoa-Gomez, J. F., Bonanni, L., & Brønnick, K. (2024). ComBat models for harmonization of resting-state EEG features in multisite studies. *Clinical Neurophysiology*, *167*, 241-253.

Jiang, Z., Zheng, Y., Tan, H., Tang, B., & Zhou, H. (2016). Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163-178.

Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1548-1558).

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590-1598.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, *3*(1), 18-31.

Kollias, D., & Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*.

Kollias, D., & Zafeiriou, S. (2019). Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNSs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880-2894.

Kothe, C., Shirazi, S. Y., Stenner, T., Medine, D., Boulay, C., Grivich, M. I., ... & Makeig, S. (2025). The lab streaming layer for synchronized multimodal recording. *Imaging Neuroscience*, *3*, IMAG-a.

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2024). A practical applications guide to machine learning regression models in psychology with Python. *Methods in Psychology*, 11, 100156.

Kovač, N., Ratković, K., Farahani, H., & Watson, P. (2025a). Machine learning regression models for internal shame. Acta Psychologica, 260, 105721.

Kovač, N., Ratković, K., Watson, P., Farahani, H., & Bagheri Sheykhangafshe, F. (2025b). Machine learning classification models for predicting chronic pain. *Current Psychology*, 1–14.

Kragel, P. A., & LaBar, K. S. (2016). Decoding the nature of emotion in the brain. *Trends in cognitive sciences*, *20*(6), 444-455.

Levenson, R. W. (2014). The autonomic nervous system and emotion. *Emotion review*, *6*(2), 100-112.

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2852-2861).

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2023). emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. *J. Open Source Softw.*, *2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.* (pp. 281-297).

Mehta, D., Siddiqui, M. F. H., & Javaid, A. Y. (2019). Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors*, *19*(8), 1897.

Michel, C. M., & Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage*, *180*, 577-593.

Mohammad, S. M. (2025). NRC VAD Lexicon v2: Norms for valence, arousal, and dominance for over 55k English terms. *arXiv preprint arXiv:2503.23547*.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, *14*.

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2021, July). Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2024). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, T. J., Han, K. J., Kumar, M., & Narayanan, S. (2019). Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, *27*, 381-385.

Peter, J., & Rousseeuw, S. (1987). A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*, *20*.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Cheikh, F. A., Hijji, M., ... & Rodrigues, J. J. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alexandria Engineering Journal*, *68*, 817-840.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, *10*(5), 1299-1319.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in cognitive sciences*, *22*(11), 969-981.

Shu, Y., Gu, X., Yang, G. Z., & Lo, B. (2022). Revisiting self-supervised contrastive learning for facial expression recognition. *arXiv preprint arXiv:2210.03853*.

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., ... & Yu, T. (2022). One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, 107299.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395-416.

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ... & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191-1207.

Xie, J., Girshick, R., & Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In *International conference on machine learning* (pp. 478-487). PMLR.

Yger, F., Berar, M., & Lotte, F. (2016). Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(10), 1753-1762.

Zhang, D. C., & Lauw, H. (2022, June). Dynamic topic models for temporal document networks. In *International Conference on Machine Learning* (pp. 26281-26292). PMLR.

Zhang, M., Ma, B., Xiao, C., & Wang, C. (2024). Self-supervised learning-based emotion recognition using EEG. *Frontiers in Human Neuroscience, 18*, 1334721.

Zhang, Z., Wang, L., & Yang, J. (2023). Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18888-18897).