Unsupervised Learning for Detecting Cognitive Distortions

A novel framework for analyzing patient narratives in psychotherapy.

Bobo, S., & Kolonin, A. (2025). Unsupervised learning for detection of cognitive distortions in patient narratives . In Twenty-Seventh International Conference on Neural Networks “Neuroinformatics-2025” . https://openreview.net/forum?id=RiDhbCyws2

Introduction to Cognitive Distortions

Cognitive distortions (CDs) are systematic thought errors common in mental health conditions, fueling anxiety and negatively impacting life. Bobo, S., & Kolonin, A. (2025).

1

Definition

Internal subconscious and conscious mental filters or biases that increase self-misery.

2

Prevalence

37% increase in CD prevalence in public discourse since 1980.

3

Impact

Manifest as maladaptive automatic thoughts perpetuating psychological distress.

Problem and Motivation

Current methods face significant limitations, hindering accessible mental healthcare.

1

Time–Intensive

Diagnostics take an average of 5 minutes per case. (Asghar et al. (2020) Automatic detection of Cognitive Distortions form text using Machine learning)

2

Subjectivity

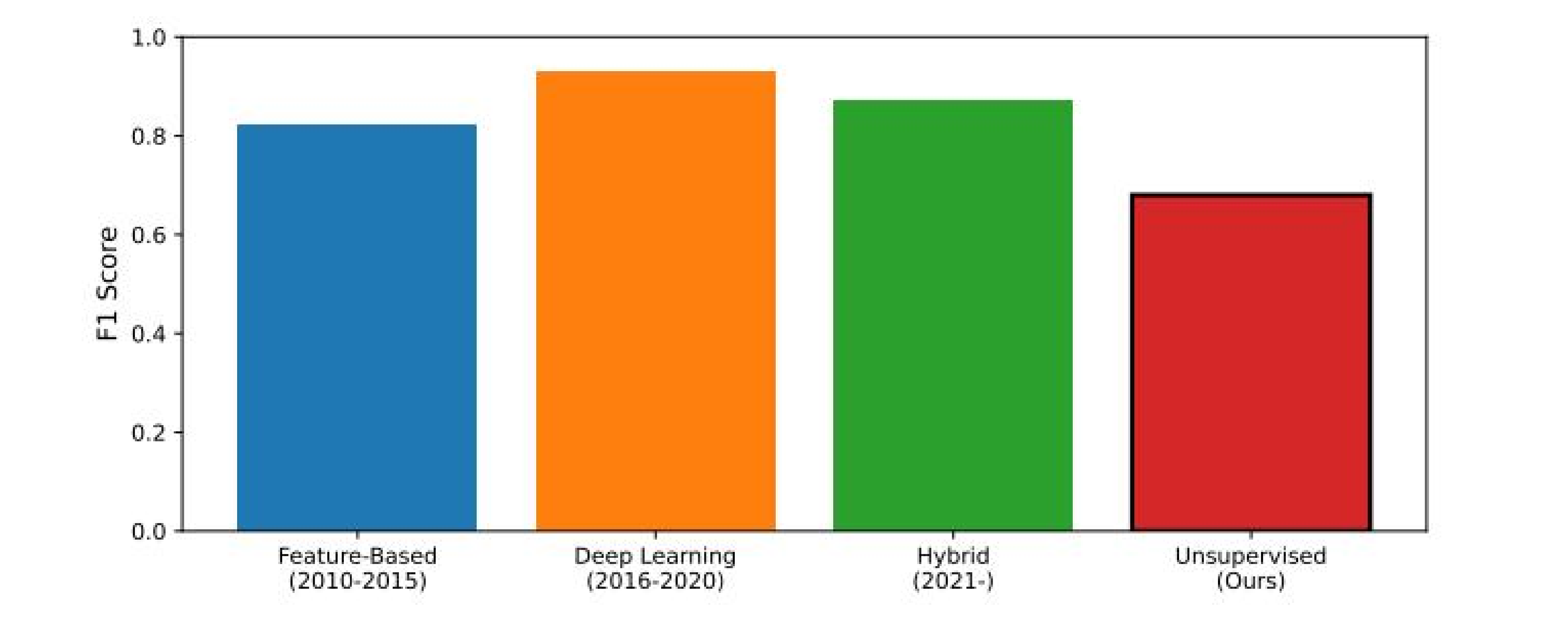
Clinician judgment varies (Fleiss Kappa = 0.45-0.60]).

3

Scalability

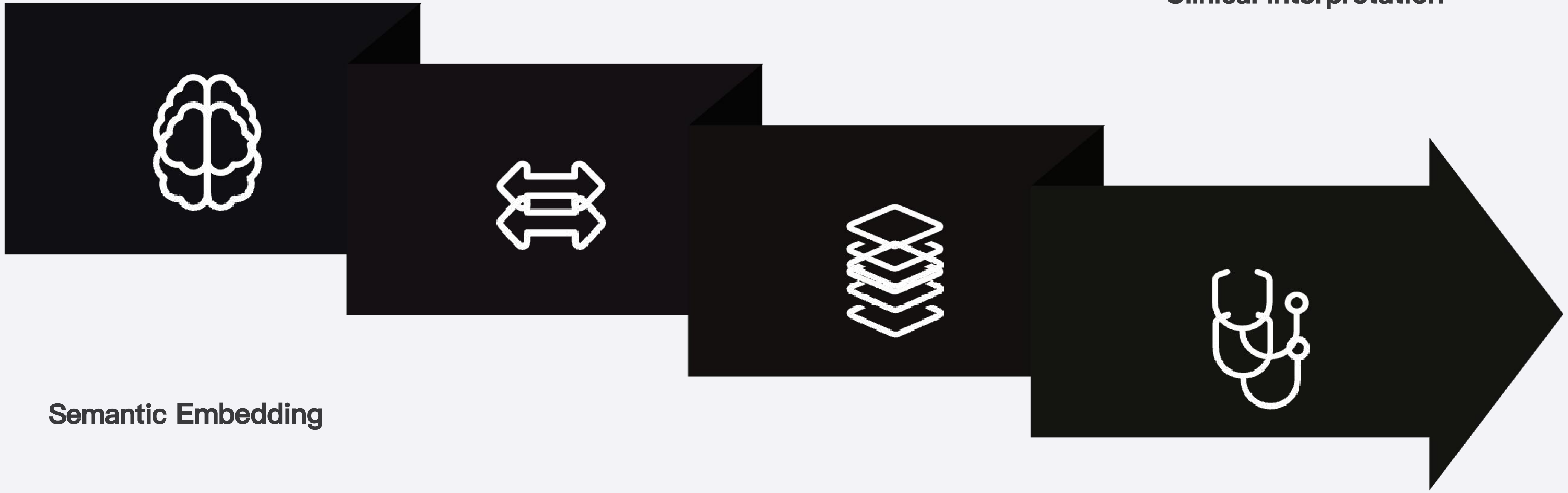
Limited for large datasets, restricting accessibility.

EVOLUTION OF COGNITIVE DISTORTION DETECTION METHODS



Our Unsupervised Approach

We propose a framework using unsupervised machine learning to address current challenges.



Methodological Innovations

Our framework integrates several techniques for robust CD detection.

Text Processing	Cleanup from formatting and irrelevant/junk characters
Clustering	Dynamic HDBSCAN thresholds adapting to distortion density.
Validation	Mixed quantitative-clinical evaluation protocol against Public anonymized labeled clinical and synthetic data, AI-assisted
Interpretability	KeyBERT-assisted clinical labeling of clusters.

$$C = \mathrm{HDBSCAN}\left(\mathrm{PCA}_{75}\left(\mathrm{MiniLM}(T)\right)\right)$$

Where
C = Cluster
T = Text
MiniLM = Semantic embedding

1 $$\mathrm{PCA}_{75} : \mathbb{R}^{n\times 384} \to \mathbb{R}^{n\times 75}, \quad \mathrm{Var}_{\mathrm{retained}} = 92\%$$

Where R = Semantic embeddings

Clustering Optimisation (HDBSCAN)

2 $$\mathrm{min\_cluster\_size} = \max(5,\ 0.01n), \quad \varepsilon = 0.5, \quad \mathrm{Silhouette\ Score} = 0.098$$

With K = 4 (The number of clusters obtained) at 1st level clustering with HDBSCAN

Key Results - Cluster Profiles

- - Dominant CD Profiles:
- • Performance Anxiety (n=93): 100% CDs
- • Social Anxiety (n=3,680): 64.9% CDs
- • Social Exclusion (n=122): 98.4% CDs
- • Mixed Symptoms (n=2,162): 74.1% CDs
- - Hierarchical Insights: Subclusters ↑ topic specificity improved

SUBCLUSTER DISTRIBUTION AND THEMATIC SCOPE

CLUSTER	THEME	# SUB CLUSTERS	SCOPE OF SUBCLUSTER THEMES
Perfomance Anxiety	Professional Failures	19	Project rejections, academic failures, promotion denials
Social Anxiety	Social rejection	18	Wedding anxieties, party exclusions, friend disputes
Social Exclusion	Daily Challenges	20	Mental health subtypes, work stress, social insecurities
Mixed Symptoms	Relationship concerns	20	Romantic conflicts, family tensions, peer exclusion

Hierarchical Subclustering with KeyBERT

KeyBERT keyword extraction refines broad clusters into clinically meaningful subgroups.

Work/School	Mind Reading	"must think I'm incompetent"
Work/School	Personalization	"my fault project failed"
Relationships	Labeling	"I'm a terrible friend"
Relationships	Catastrophizing	"Nobody will ever love me"

This approach enhances therapeutic relevance and uncovers latent themes, enabling precise linkage between life contexts and cognitive distortions.

Key Subclustering Observations

Our hierarchical approach significantly improves thematic specificity and clinical relevance.

1

Thematic Expansion

Subclusters increased topic specificity.

2

Noise Utilization

100% of "Mixed symptoms" (noise) cluster entries reorganized into clinically relevant subgroups.

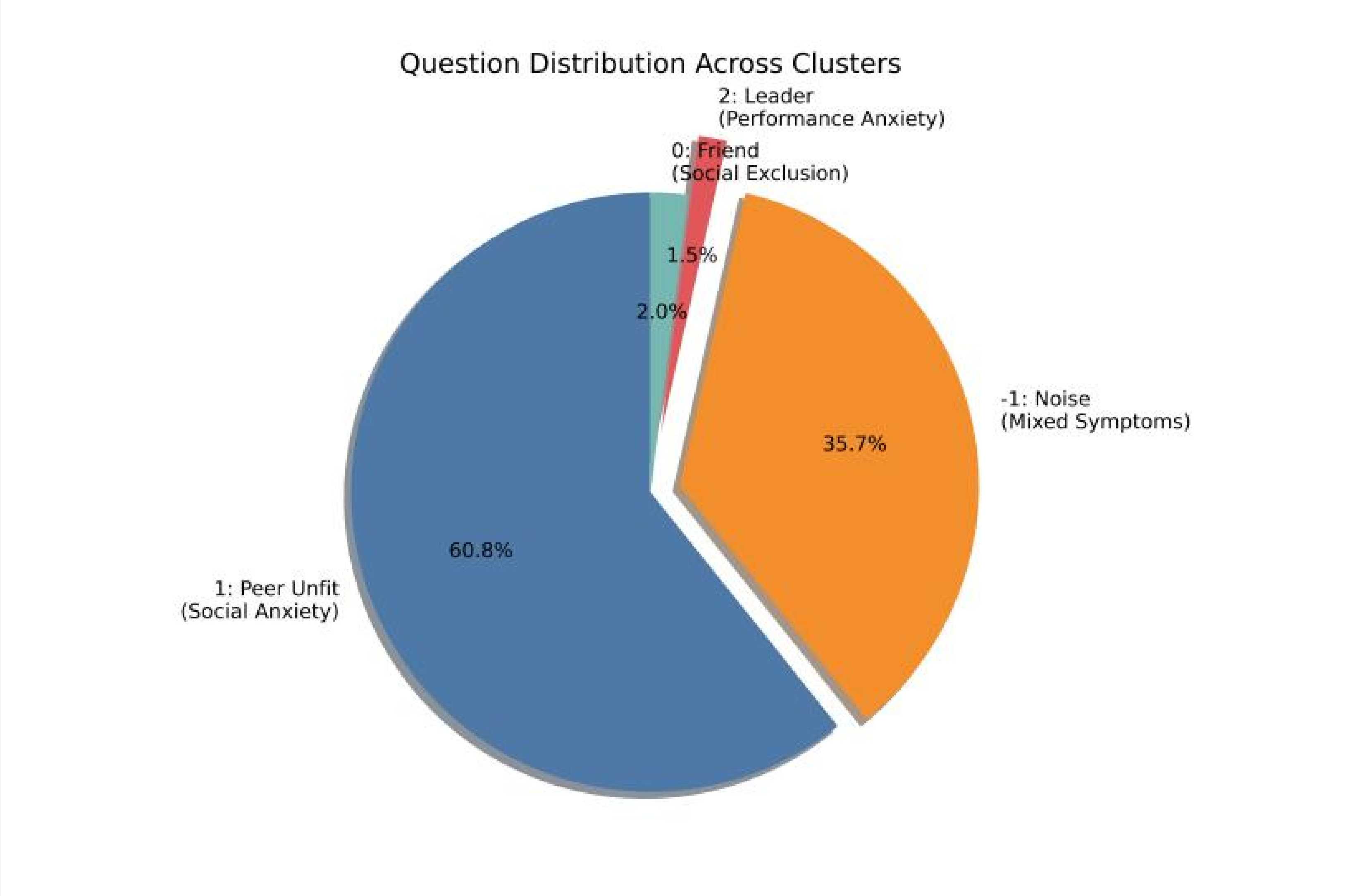
3

Distortion Resolution

Achieved 97% keyword non-overlap (Jaccard similarity = 2.2×10^{-5}).

Clinical Patterns in Cognitive Distortions

Our clustering pipeline revealed distinct patient profiles with high distortion prevalence.



Cluster Profiles with clinical interpretation

ID	Clinical Label	N	Distortion %	Avg Size	Prototypical Phrase
1	Social Anxiety	3,680	64.9	184	”I’m scared nobody will ever love me”
-1	Mixed Symptoms	2,162	74.1	108	”I’m afraid I’ll die alone”
2	Performance Anxiety	93	100.0	5	”My work is totally worthless”
0	Social Exclusion	122	98.4	7	”They excluded me on purpose”

Key Findings & Clinical Utility

Our research provides actionable insights for mental healthcare.

100%

High-Risk Group

Performance Anxiety cluster shows 100% distortion.

0.82

Linguistic Markers

Absolute terms ("Total"/"Never") predict distortions (AUC).

92%

Clinical Alignment

Our model aligns with CBT intervention targets.

2.7

Efficient Pipeline

Processes 6,057 texts in under 3 hours.

Clinical Validation

- Expert Evaluation: AI-Assisted validation with (MetaA!, ChatGPT,

DeepSeek

- Agreement: Fleiss’ κ = 0.68 ("substantial agreement")

- Linguistic-Distortion Correlations:

  • Absolute terms → 100% CDs

Advantages & Applications

- - Efficiency: 6,057 texts in <3 hrs (CPU, 47× faster than manual)
- - Clinical Use Cases:
-   1. Cognitive distortion diagnostics time reduced
-   2. Population Health (real-time CD tracking)
- - Scalability: No predefined labels; adapts to new CD patterns

Limitations

- English-only corpus
- Small clusters (e.g., n=93)
- Concept drift

Conclusion & Future Directions

This work establishes foundational advances in unsupervised clinical validity and operational scalability.

Risk Prioritization

Urgency scoring for immediate intervention.

Linguistic Markers

Absolute terms ("never," "totally") as diagnostic indicators.

Treatment Personalization

Cluster-specific CBT protocol recommendations.

$$P(CD \mid AbsoluteTerm) \propto \text{Frequency of Absolute Terms}$$

Future work includes integrating patient clinical profiles and EHR for enhanced utility.

Future Directions and Scalability

- Expand to multilingual datasets using mBERT.
- Integrate Electronic Health Records (EHR) for real-time monitoring.
- Link patient phenotype (age, comorbidities) with CD patterns.
- Enhance model adaptability to concept drift (19% new terms/year).
- Deploy on edge devices (CPU-only inference in <3 hours).

Q&A

- - Acknowledgements: Dataset providers, clinical validators
- - Contact: bobosamson8@gmail.com