A Survey of Cognitive Distortion Detection and Classification in NLP

Archie Sage, Jeroen Keppens, Helen Yannakoudakis
Department of Informatics, King's College London
{archie.sage, jeroen.keppens, helen.yannakoudakis}@kcl.ac.uk

Abstract

As interest grows in applying natural language processing (NLP) techniques to mental health, an expanding body of work explores the automatic detection and classification of cognitive distortions (CDs). CDs are habitual patterns of negatively biased or flawed thinking that distort how people perceive events, judge themselves, and react to the world. Identifying and addressing them is a central goal of therapy. Despite this momentum, the field remains fragmented, with inconsistencies in CD taxonomies, task formulations, and evaluation practices limiting comparability across studies. This survey presents the first comprehensive review of 38 studies spanning two decades, mapping how CDs have been implemented in computational research and evaluating the methods applied. We provide a consolidated CD taxonomy reference, summarise common task setups, and highlight persistent challenges to support more coherent and reproducible research. Alongside our review, we introduce practical resources, including curated evaluation metrics from surveyed papers, a standardised datasheet template, and an ethics flowchart, available online.[1]

1 Introduction

Cognitive distortions (CDs) are habitual patterns of negatively biased or logically flawed thinking that distort how people perceive events, judge themselves, and react to the world around them. These distortions play a central role in emotional distress and are a core target of evidence-based psychological interventions such as cognitive behavioural therapy (CBT) (Beck, 1963; Burns, 1999).

Common examples[2] include *Catastrophising* ('I let them down, so they'll never trust me again'), *Mind Reading* ('They haven't replied, so they must be angry at me'), and *All or Nothing Thinking* ('If I don't get this right the first time, I'm a complete failure'). These patterns often appear intuitive or harmless at first, but they have been shown to maintain and exacerbate conditions like depression, anxiety, and post-traumatic stress disorder. In therapeutic settings, recognising and *reframing* such distortions is a core goal of CBT. Crucially, these distortions are primarily expressed through language, making them well-suited to computational modelling. Recent work in natural language processing (NLP) has begun to explore the automatic detection and classification of CDs, with applications ranging from clinical decision support tools to mental health chatbots, journaling tools, and triage systems. Studies have shown that incorporating CD-level features can improve outcomes in related tasks such as depression detection (Wang et al., 2023b), complementing more traditional sentiment or topic based approaches. By identifying distorted cognitive patterns in everyday text, NLP systems may support more timely, personalised, and psychologically-informed interventions.

Despite rapid growth in the field, the literature remains fragmented. Computational approaches use inconsistent taxonomies for defining CDs, making it difficult to compare findings across studies. Task formulations (e.g., detection vs classification, single-label vs multi-label) vary widely, often reflecting implicit assumptions that shape evaluation and outcomes. Benchmarks are scarce, metrics inconsistently applied, and variations in domain and dataset usage further complicate comparisons, making it hard to pinpoint gaps or establish best practices.

To our knowledge, the most comparable prior survey is by Suputra et al. (2023), which examined 12 studies and provided an initial synthesis of modelling approaches used at the time. Our work expands on this by covering 38 publications, in-

[1] https://github.com/archiesage/cognitive-distortion-nlp-survey

[2] A complete list of CD definitions, examples, and synonyms is provided in Table 3, with additional psychological background in Appendix B.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14884–14899
November 4-9, 2025 ©2025 Association for Computational Linguistics

Code	Cognitive Distortion	THERAPISTQA: Shreevastava and Foltz (2021)	Chen et al. (2023)	Lim et al. (2024)	Pico et al. (2025)	Zhang et al. (2025)	Babacan et al. (2025)	Varadarajan et al. (2025)	TEXT INTERVENTION: Lybarger et al. (2022)	Ding et al. (2022)	Tauscher et al. (2023)	THINKING TRAP: Sharma et al. (2023)	Agarwal and Sirts (2025)	C-JOURNAL: Elsharawi and El Bolock (2024)	Rasmy et al. (2025)	CDS: Bathina et al. (2021)	Lalk et al. (2024)	OTHERS: Wiemer-Hastings et al. (2004)	Xing et al. (2017)	Rojas-Barahona et al. (2018)	Shickel et al. (2020)	Lee et al. (2021)	Mostafa et al. (2021)	Alhaj et al. (2022)	Wang et al. (2023b)	Wang et al. (2023a)	Maddela et al. (2023)	Lin et al. (2024)	Qi et al. (2024)	Kim and Kim (2025)	%†
Widely Adopted: Frequently seen in NLP, typically with clearer semantic distinctions, and recommended as a focus for future research.																															
OVG	Overgeneralisation	●	◐	◐	◐	◐	◐	◐	●	◐	◐	●	◐	●	◐	●	◐	●	●	●	●	○	●	●	●	●	●	○	●	●	100%
SHD	Should Statements	●	◐	◐	◐	◐	◐	◐	●	◐	◐	●	◐	●	◐	●	◐	●	●	●	●		●	●		●	●	○	●	●	93%
LBL	Labelling	●	◐	◐	◐	◐	◐	◐				●	◐	●	◐	●	◐	●	●	●	●	●		●	●	●	●	○	●	●	86%
AON	All or Nothing Thinking	●	◐	◐	◐	◐	◐	◐				●	◐	●	◐	●	◐	●	●	●	●	●			●	●	●	○	●	●	83%
EMR	Emotional Reasoning	●	◐	◐	◐	◐	◐	◐				●	◐	●	◐	●	◐	●	●	●	●			●	●	●		○	●	●	79%
PRS	Personalisation	●	◐	◐	◐	◐	◐	◐				●	◐	●	◐	●	◐	●	●	●	●			●	●	●	●	○	●	●	79%
MFL	Mental Filter	●	◐	◐	◐	◐	◐	◐	●	◐	◐			●	◐	●	◐		●	●	●			●		●	●	○	●	●	76%
MDR	Mind Reading	●	◐	◐	◐	◐	◐	◐				●	◐			●	◐	●	○	●	●				●	●	●	○	●		69%
FTL	Fortune Telling	●	◐	◐	◐	◐	◐	◐				●	◐			●	◐	●	○	●		●			●	●	●	○	●		69%
CAT	Catastrophising							●	●	◐	◐	●	◐	●	◐	●	◐			●	●			●			●				48%
DQP	Disqualifying the Positive											●	◐			●	◐	●	●	●		○				●	●		●	●	41%
Occasionally Adopted: Includes semantically overlapping or synonymous variants, which are often merged in practice.																															
MAG	Magnification	●	◐	◐	◐	◐	◐					●					●	●										○	●		38%
JTC	Jumping to Conclusions								●	◐	◐			●	◐				●	●									●	28%
BLM	Blaming											●	◐	●	◐					●	●								●		24%
CMP	Comparing											●	◐				●			●											14%
MXN	Magnification or Minimisation															●			●							●				●	14%
Rarely Adopted: Poorly represented in NLP studies, often appearing only in isolated datasets.																															
BRT	Being Right													●	◐						●										10%
CTL	Control Fallacy													●	◐						●										10%
FOC	Fallacy of Change													●	◐						●										10%
FOF	Fallacy of Fairness													●	◐						●										10%
NFE	Negative Feeling or Emotion											●	◐																		7%
HRF	Heaven's Reward Fallacy																				●										3%
LFT	Low Frustration Tolerance																			●											3%
MIN	Minimisation																●														3%

Table 1: Consolidated CD categories and their inclusion across NLP papers in our survey that address the CD classification task, either directly or conceptually. Where applicable, papers are grouped by the CD dataset they rely on, reflecting the underlying taxonomy of those datasets. Definitions of CD categories can be found in Table 3.
† Percentage of all papers in this table that reference the given CD in any form (●, ◐, or ○).

● Used in experiments ◐ Inherited taxonomy from dataset usage ○ Mentioned conceptually only

cluding recent preprints, and offering a structured overview of the computational landscape for CD detection and classification. This survey aims to provide a clear and practical entry point for researchers engaged in the growing field of CD detection and classification from a computational perspective. Our focus is firmly on how computational methods approach these tasks, without seeking to redefine or adjudicate psychological constructs themselves. The contributions in this paper are threefold. (i) It provides a consolidated reference of the CD taxonomies used across computational studies, highlighting inconsistencies and listing common synonyms (Tables 1 and 3). (ii) It defines and analyses task setups, datasets, computational methods, and approaches to performance evaluation, highlighting key patterns and gaps (§2, §4, §5, §6). (iii) It identifies open challenges (§7) and proposes best practices to guide future research (§8). In doing so, this paper aims to enable more consistent, comparable, and reproducible work in this emerging field.

2 Task Definitions

At its core, computational work on CDs involves a basic classification task: determining whether a given text reflects distorted thinking, and if so, identifying the specific type(s) of distortion. However, a number of different versions of this definition have been adopted in distinct groups of studies, inhibiting direct comparison between them. Some publications combine the classification task with additional tasks that may or may not inform classification. This section reviews how existing research defines CD tasks, highlights key differences, and links these to their clinical foundations.

2.1 Binary Classification (Detection)

The simplest way to frame the computational task is to ask whether a given text span contains any

instance of distorted thinking. This is often called *detection* and is usually treated as a binary classification problem (Distorted vs Undistorted). Methodologically, detection is a basic case of classification with just two labels - the approach remains the same, only the label set is coarser. This binary framing reflects early clinical aims, where simply noticing the presence of distorted thinking is an important first step before exploring the person's thoughts in more detail during therapy. Even so, most computational studies go further, aiming to identify specific CD types.

2.2 Single-label vs Multi-label Classification

While detection addresses the presence of distorted thinking, classification aims to specify which type(s) of CD are present in a given text. Here, task definitions diverge based on assumptions around label cardinality:

Single-label classification. This assumes that each text span reflects exactly one type of distortion. This simplifying assumption is often motivated by practical constraints, such as dataset design or the brevity of inputs (e.g., tweets, short user queries). However, it neglects the fact that distorted thoughts frequently exhibit multiple overlapping CD categories, particularly in longer or more detailed texts.

Multi-label classification. This allows a text to be assigned multiple CD categories simultaneously. This formulation more accurately reflects clinical reality, where distortions co-occur and interact. Studies adopting multi-label setups typically model each CD type as an independent binary label, which simplifies analysis and inter-annotator agreement calculations (Lybarger et al., 2022; Tauscher et al., 2023).

2.3 Auxiliary Tasks

Beyond isolated classification, some studies incorporate auxiliary tasks that extend the utility or interpretability of CD models. These are not necessarily distinct problem categories, but rather downstream or complementary tasks that build upon classification outputs:

Reasoning generation. This involves producing explanatory rationales for why a particular text span was classified as distorted. Methods such as Diagnosis of Thought (DoT) prompting (Chen et al., 2023) and the later ERD (Extraction, Reasoning, Debate) framework (Lim et al., 2024) aim to mimic clinician-like reasoning, improving model transparency and trustworthiness.

Reframing generation. This focuses on producing healthier rephrasings of distorted thoughts, consistent with CBT interventions. Studies in this area (Sharma et al., 2023; Maddela et al., 2023) treat reframing as a natural extension of detection and classification.

Multi-task learning. These setups combine CD classification with related objectives, such as depression severity prediction (Wang et al., 2023a) or emotion cause extraction (Singh et al., 2023). These formulations tend to leverage the diagnostic value of CD features to improve performance on auxiliary tasks.

Multi-modal approaches. While the vast majority of computational CD research focuses on textual data, recent work has begun to explore multi-modal approaches. For example, Singh et al. (2023) integrated text, audio, and video inputs from therapist-patient interactions to enhance CD detection. Though still nascent, these efforts highlight the potential of multi-modal signals in capturing the subtle nuances of distorted thinking in real-world settings.

3 Taxonomies of CDs in NLP

Despite the shared objective of classifying CDs, studies seeking to label CDs have adopted diverse taxonomies. Without standardisation, it is hard to keep annotations consistent, compare models properly, or clearly interpret the results. Table 1 illustrates this fragmentation by mapping which CD categories are recognised across the papers surveyed. While some categories, such as *Labelling* and *Should Statements*, are commonly used, others are inconsistently applied or redefined in various ways. Early conceptions of CDs were proposed by Beck (1963), who described patterns of dysfunctional thinking with examples such as arbitrary inference and overgeneralisation. This work was later popularised by Burns (1999), whose ten-category taxonomy is frequently cited in psychology literature.[3] However, computational studies do not uniformly follow this framework.

[3]While Burns (1999) is not the only taxonomy of CDs used in psychological contexts, and clinical practice often employs broader or alternative frameworks, our focus is on the inconsistency in how these definitions are interpreted and applied within computational research.

Some papers, such as Shickel et al. (2020), draw on definitions from popular psychology sources (e.g., PsychCentral, Psychology Today), resulting in the inclusion of broader or differently framed categories. Other works make subtle assumptions in how the Burns taxonomy is applied, for example, by splitting *Jumping to Conclusions* into its subcategories *Mind Reading* and *Fortune Telling*, sometimes without explicit rationale.

Terminology also varies across studies. The CD category referred to as *All or Nothing Thinking* also frequently appears under alternative labels such as *Black and White Thinking*, *Polarised Thinking*, or *Dichotomous Reasoning*. This terminological variety complicates efforts to harmonise and compare datasets. Since Rojas-Barahona et al. (2018) noted that CDs were 'fairly well standardised' in computational research, the field has grown considerably - from a handful of studies to 38 now in this survey - leading to a surge in differing taxonomies. The aim of Table 1 is to offer a consolidated view of this evolving landscape, providing a resource for future NLP research. Additionally, we include an appendix table (Table 3) listing synonyms, definitions, and hierarchical relationships between CD types, to support more consistent and transparent classification efforts in computational contexts.

4 Datasets

Datasets are the foundation for research on CD classification, providing the labelled examples needed to develop and evaluate detection methods. However, existing datasets vary widely in scope, annotation practices, and accessibility. To organise this diversity, we group datasets by their underlying data sources and contexts of use.

4.1 Domains

We use the term *domain* to describe the broader context from which a dataset's text data originates. Domains shape the linguistic style of examples, affect annotation reliability, and carry practical considerations such as data privacy and availability. The following six domains reflect the main sources of data in current CD classification research.

Literature Examples. Early work, such as Wiemer-Hastings et al. (2004), used CD examples from existing psychological literature (Beck, 1979; Burns, 1999). While these examples are clear and well-labelled, they are typically idealised and explicit, limiting their applicability to real-world, patient-generated language.

Social Media Platforms. Public posts from platforms such as Reddit (Aureus et al., 2021), Twitter (Alhaj et al., 2022), and Weibo (Qi et al., 2024) provide naturally occurring CD instances in user-generated content. This domain offers large volumes of data but poses challenges related to linguistic noise and context ambiguity.

Digital Mental Health Platforms. Peer-support services, such as Koko (Rojas-Barahona et al., 2018) and TaoConnect (Shickel et al., 2020), have been valuable sources of data rich in CDs. The widely used THERAPISTQA dataset (Shreevastava and Foltz, 2021) originates from a Kaggle Q&A repository and has since been extended in multiple studies (Chen et al., 2023; Babacan et al., 2025; Lalk et al., 2024).

Crowd-Sourced Approaches. To tackle issues of data scarcity and privacy, several studies have turned to crowdworkers to generate or annotate CD examples. Well-known corpora created this way include CROWDDIST (Shickel et al., 2020), PATTERNREFRAME (Maddela et al., 2023), and THINKING TRAP (Sharma et al., 2023). These datasets are scalable and flexible but often lack the subtlety of real-world data, as crowdworkers may produce overly explicit examples.

Clinical Interventions. These datasets, derived from real therapeutic conversations, reflect how people communicate in real-world settings. Notable examples include annotated patient-therapist text message exchanges (Lybarger et al., 2022; Tauscher et al., 2023) and psychotherapy transcripts (Lalk et al., 2024). Multimodal corpora such as CODEC and CODER (Singh et al., 2023, 2024) also fall into this category, although they involve a mix of authentic and staged interactions. Despite their value, these datasets are often subject to access restrictions due to privacy and ethical considerations.

Synthetic Datasets Recent work has explored using large language models (LLMs) to generate synthetic CD data. Babacan et al. (2025) created GPT-4-generated corpora, while Kim and Kim (2025) recently released KOACD, a Korean dataset augmenting social media data with synthetic samples. Synthetic datasets support balanced and scalable resource creation, but may fail to capture the nu-

Dataset†	Language	Size (# Samples)*	Labelling‡	Annotators	Access
Literature Examples					
Wiemer-Hastings et al. (2004)	English	261	Single-label (10)	Expert	Private
Social Media					
Alhaj et al. (2022)	Arabic	9,250	Single-label (5)	Non-Expert (Unspecified)	Private
SocialCD-3K, Qi et al. (2024)	Mandarin	3,407	Multi-label (12)	Domain-Informed	Public
Aureus et al. (2021)	English	586	Binary (2)	Mixed	Private
Simms et al. (2017)	English	459	Binary (2)	Mixed	Private
Digital Mental Health Platform					
Rojas-Barahona et al. (2018)	English	4,035	Multi-label (15)	Expert	Private
Lin et al. (2024)	Mandarin	4,001	Binary (2)	Domain-Informed	Public
TherapistQA, Shreevastava and Foltz (2021)	English	2,529	Multi-label (10)	Non-Expert (Unspecified)	Public
MH-D, Shickel et al. (2020)	English	1,799	Binary (2)	Domain-Informed	Private
MH-C, Shickel et al. (2020)	English	1,164	Single-label (15)	Domain-Informed	Private
CBT-CD, Zhang et al. (2025)	English	146	Multi-label (10)	Expert	Public
Crowd-sourced					
Elsharawi and El Bolock (2024)	English	34,370	Single-label (14)	Expert	Private
PatternReframe, Maddela et al. (2023)	English	9,688	Multi-label (10)	Crowd-Generated	Public
CrowdDist, Shickel et al. (2020)	English	7,666	Single-label (15)	Crowd-Generated	Private
C2D2, Wang et al. (2023b)	Mandarin	7,500	Single-label (7)	Crowd-Generated	Request
Thinking Trap, Sharma et al. (2023)	English	600	Multi-label (13)	Expert	Public
Synthetic					
GPT-4 Synthetic, Babacan et al. (2025)	English	2,000	Single-label (10)	Automated (LLM)	Public
Clinical Intervention					
Lalk et al. (2024)	German	104,557	Multi-label (14)	Automated (Lexicon)	Request
Lybarger et al. (2022)	English	7,436	Multi-label (5)	Expert	Private
Hybrid (Mixed Domains)					
KoACD, Kim and Kim (2025)	Korean	108,717	Single-label (10)	Automated (LLM)	Request
GPT-4 Combined, Babacan et al. (2025)	English	4,530	Single-label (10)	Automated (LLM)	Request
CoDEC, Singh et al. (2023)	English	3,773	Binary (2)	Non-Expert (Unspecified)	Request
CoDeR, Singh et al. (2024)	English	3,773	Binary (2)	Trained	Public
Wang et al. (2023a)	English	3,644	Single-label (11)	Automated (BERT)	Private
Mostafa et al. (2021)	English	2,409	Single-label (2)	Domain-Informed	Private

Table 2: Overview of datasets for CD detection and classification, grouped by domain. See Appendix Table 4 for an expanded version with agreement metrics, access details, and subdomains. † Corpus name, or earliest study to use it for CD tasks. * Number of annotated units (e.g., posts, speech turns); for automated annotations, items processed. ‡ Number of CD categories used, excluding 'Undistorted' for classification.

anced linguistic and contextual patterns present in genuine human language.

4.2 Comparative Discussion

Each domain offers distinct strengths and drawbacks. Clinical intervention datasets are highly representative of real-world therapeutic contexts but are usually small and difficult to obtain. In contrast, social media and digital mental health platforms provide scalable, naturally occurring data, though they often exhibit linguistic noise and structural inconsistency. Crowd-sourced datasets allow for controlled creation of CD examples but can introduce stylistic artefacts that may not mirror authentic language use. Synthetic datasets, including those generated by LLMs, support large-scale experimentation and balanced class distributions, yet require thorough validation to ensure their realism. Most existing corpora are monolingual, predominantly in English, as summarised in Table 2. However, recent efforts have started expanding into other languages. For instance, Wang et al. (2023b) introduced C2D2, a Mandarin corpus, while Kim and Kim (2025) developed KoACD, a Korean dataset. Additional dataset details, including subdomains, access links, and inter-annotator agreement (IAA) figures, are provided in Appendix Table 4.

4.3 Annotation Strategies

Annotation strategies vary considerably across datasets. Clinical and literature-derived corpora typically rely on expert annotators, prioritising la-

bel quality at the expense of scalability. In contrast, social media and crowd-sourced datasets often involve non-expert annotators, sometimes supported by brief, domain-specific training from qualified psychologists to improve consistency. While these strategies enable large-scale annotation, reliably classifying CDs remains challenging. The task demands subtle, often subjective judgements, and studies consistently report low IAA, particularly when annotators lack deeper domain expertise. To mitigate this, some works adopt strict inclusion criteria, only retaining examples where annotators fully agree on the label or a subset of distortion types (Aureus et al., 2021; Shickel et al., 2020). Though this approach improves label precision, it risks introducing bias by systematically excluding ambiguous or borderline cases - which are arguably the most reflective of real-world CD occurrences.

5 Modelling Approaches

This section outlines computational approaches to CD detection and classification, grouped into six methodological categories that reflect major developments in the field.

5.1 Rule-Based

The first systems for CD classification were rule-based, using hand-crafted keyword patterns and syntactic features. Wiemer-Hastings et al. (2004) developed COGNO, a system that mapped surface linguistic cues (e.g. verb tense, negation, person markers) to predefined CD categories. It performed well on a 10-class single-label task (Macro-F1 = 0.61) but was only tested on 'polished' textbook-style CD examples. One of the main strengths of rule-based systems is their interpretability, a feature still highly valued in clinical settings, where transparency is critical. For instance, Lalk et al. (2024) employed a manually curated list of n-grams, based on previous work (Bathina et al., 2021), to monitor distortion frequency in psychotherapy transcripts and predict patient depression severity.

5.2 Traditional Machine Learning (Feature-based)

As more annotated corpora became available, early rule-based systems gave way to feature-based statistical models. These approaches combined classic classifiers, such as logistic regression (LR) and support vector machines (SVMs), with engineered features such as Linguistic Inquiry and Word Count (LIWC) scores and Term Frequency-Inverse Document Frequency (TF-IDF) vectors. Simms et al. (2017) demonstrated that LR trained on LIWC features performed well on the detection task using Tumblr data, while Shickel et al. (2020) found that TF-IDF with LR outperformed CNNs on a 15-class single-label synthetic dataset (F1 = 0.68), indicating that shallow linguistic cues can remain competitive even in classification tasks. Similarly, Shreevastava and Foltz (2021) used SVM with smooth inverse frequency (SIF) embeddings, achieving strong performance (F1 = 0.77) on the detection task despite the insensitivity of SIF to word order.

5.3 Deep Learning with Static Embeddings

Static word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) introduced vector-based representations of words derived from co-occurrence patterns. These effectively captured word similarity but failed to account for contextual nuance (e.g., 'riverbank' vs 'bank'). Despite this limitation, such embeddings formed the backbone of early deep-learning based CD classifiers. For example, Rojas-Barahona et al. (2018) combined GloVe with CNNs for multi-label CD classification, outperforming traditional models. Similarly, Mostafa et al. (2021) trained LSTM models using GloVe vectors to classify different CD types, although their findings highlighted concerns around overfitting due to the use of limited and synthetic data. To address data sparsity in Arabic texts, Alhaj et al. (2022) applied contextual topic modelling (CTM), integrating static embeddings with domain-specific topic information via BERTopic - an approach that proved helpful in low-resource settings. While static embeddings offered richer representations than earlier methods, their inability to capture context, especially for polysemous words and subtle pragmatic distinctions, ultimately led to the rise of contextual models.

5.4 Transformer-based Architectures

Transformers, particularly BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), introduced contextual embeddings that capture word meaning based on the surrounding sentence, marking a significant leap forward in many NLP tasks, including CD detection and classification. Shreevastava and Foltz (2021) used fine-tuned Sentence-BERT (SBERT) for binary CD detection, showing notable improvements over earlier models. Similarly, Lybarger et al. (2022)

demonstrated that incorporating conversational history further improves classification performance on therapy dialogues. Domain-adapted transformers like MentalBERT (Ji et al., 2022), which are pre-trained on mental-health-related data, showed benefits over general-purpose models, while augmentation techniques such as mixup (Zhang et al., 2018) were explored to improve performance on rare CD classes (minor improvements). Maddela et al. (2023) found that models like RoBERTa performed better than larger language models such as GPT-3.5 for this task, though distinguishing between closely related distortions remained a challenge. Overall, transformers brought significant gains in robustness and accuracy, but these came at the cost of reduced interpretability and a higher risk of overfitting, especially in datasets with extreme class imbalance. These limitations set the stage for the emergence of prompt-based models.

5.5 LLMs and Prompting Frameworks

LLMs like GPT-3 have enabled CD detection through natural language prompting, allowing models to perform the task without the need for tailored training. Recent work in this area can broadly be divided into two approaches: zero-shot prompting and chain-of-thought frameworks.

Zero-shot prompting. Chen et al. (2023) introduced Diagnosis-of-Thought (DoT) prompting, guiding LLMs to reason through CD detection and classification with structured outputs. While this method sometimes outperformed fine-tuned transformers on the THERAPISTQA corpus, it also is prone to hallucinations and inconsistent rationales, especially on ambiguous cases. Similarly, Pico et al. (2025) compared multiple LLMs, finding that well-prompted open-source models could approach the performance of larger proprietary models, though results were not always consistent across runs.

Chain-of-thought (CoT). Lim et al. (2024) presented the ERD framework, where multiple LLM agents simulate therapist-like reasoning, extracting emotional cues and providing structured explanations. Though this method produced richer rationales, it was highly sensitive to how prompts were designed and faced challenges in scalability and validation. Another key application of LLMs has been synthetic data generation. For example, Babacan et al. (2025) used GPT-4 to create a balanced CD dataset. While initial results were promising, the synthetic data struggled to generalise to the noisier, more complex language found in real-world user inputs.

Overall, prompted LLMs offer advantages in reducing training costs and improving interpretability, but they remain limited by issues such as prompt fragility, hallucinations, and inconsistent evaluation results.

5.6 Multimodal and Multi-task Architectures

To overcome the limitations of text-only models, recent research has explored incorporating additional modalities and joint tasks. Singh et al. (2023) introduced CODEC, a multimodal dataset combining video, audio, and text from therapy simulations. By leveraging intonation and facial expressions, their model achieved improvements in detecting emotion-related CDs such as *Emotional Reasoning*. Building on CODEC, Singh et al. (2024) developed CODER, which added annotated reasoning spans to support explanation-aware CD classification. Multi-task learning has also been employed to utilise the diagnostic value of CDs. For example, Lee et al. (2021) repurposed micromodel outputs to improve depression and PTSD prediction, while Wang et al. (2023b) demonstrated that incorporating CD frequency improved mental illness detection pipelines. These architectures have shown promise in boosting robustness, particularly in low-resource or noisy data settings. However, the scarcity of multimodal datasets and challenges around annotation and privacy continue to hinder wider adoption.

5.7 Feasibility of Meta-Analysis

While we initially intended to include comparative performance tables, meaningful aggregation proved infeasible. The studies diverge across critical axes - (i) task formulation, (ii) CD taxonomy choice, (iii) dataset domain, (iv) evaluation metrics, (v) granularity of the unit of analysis, (vi) context inclusion and window size, and (vii) modality - often incompatibly. Sample sizes within aligned subgroups are too small for robust comparison, and IAA is reported inconsistently, further limiting comparability. Pooled tables would therefore risk suggesting misleading trends. We release the full set of extracted results in our GitHub repository, enabling researchers to build their own comparisons. In this paper, we restrict ourselves to qualitative synthesis, deferring formal meta-analysis until the evidence base is larger and more standardised.

6 Evaluation

Despite progress in CD classification, evaluation practices remain inconsistent, with studies differing markedly in their choice and reporting of metrics, which hinders comparability. Although F1 score is the most commonly used metric, distinctions between macro, micro, and weighted variants are frequently overlooked - a significant issue for class-imbalanced datasets where per-class performance is critical. Some studies now report AUPRC (Area Under the Precision-Recall Curve) to account for skewed label distributions, providing a more informative measure of performance on rare CD types (Ding et al., 2022). Nonetheless, per-class metrics are still underreported, hiding weaknesses in addressing infrequent distortions. Similarly, dataset quality is often inconsistently assessed. IAA is either inconsistently measured or reported using incomparable metrics. While Cohen's Kappa (κ), which adjusts for chance, is typically more appropriate, many studies instead rely on raw agreement or non-standard metrics, blurring the line between annotation reliability and model performance. Baseline comparisons are further complicated by inconsistent CD taxonomies and datasets.

7 Challenges & Future Directions

Despite recent progress, the automatic detection and classification of CDs remains a challenging task, both conceptually and computationally. In this section, we outline three key challenges currently limiting the field: (1) inconsistency in CD taxonomies, (2) data scarcity and imbalance, and (3) the overreliance on short-form text. Addressing these issues is essential for improving model performance, evaluation fairness, and eventual clinical applicability.

7.1 Inconsistent CD Taxonomies

A longstanding challenge is the lack of a standardised taxonomy for CDs. While foundational frameworks such as the Burns ten-category list (Burns, 1999) are commonly cited, computational studies diverge significantly in how they define, split, or rename distortion types. For instance, *Jumping to Conclusions* is frequently subdivided into *Mind Reading* and *Fortune Telling*, and terms such as *All or Nothing Thinking* appear under multiple aliases (e.g., *Black and White Thinking*, *Polarised Thinking*), making it difficult to compare models across studies, reproduce results, or interpret outputs reliably. These inconsistencies also affect annotation quality, as ambiguous or overly granular label sets introduce subjectivity and reduce IAA - an issue compounded by the lack of formal guidance on taxonomy use.

7.2 Data Scarcity, Imbalance, and Annotation Limits

The field remains constrained by a lack of large, high-quality datasets that capture authentic, context-rich examples of distorted thinking. Clinical corpora are scarce and often inaccessible due to privacy constraints, while many widely used datasets are synthetic, crowd-sourced, or compiled from multiple domains, which may lack the nuance and ambiguity of real-world language. This limits the complexity of distortions that models can learn and typically results in heavy class imbalance, with rare distortion types being underrepresented or excluded altogether. Although augmentation strategies such as mixup or back-translation offer minor gains for rare categories (Ding et al., 2022), a deeper issue lies in the ceiling imposed by annotation reliability itself. In some settings, even expert annotators show limited agreement, particularly for subtle or overlapping categories. For instance, Tauscher et al. (2023) report that for the presence of 'Any Distortion' in a text, human F1 agreement was 0.63, while a fine-tuned BERT model reached 0.62 - a very small difference. This suggests that for certain formulations, such as binary detection or high-frequency classes, current models may already be approaching the upper bound set by annotation quality. It also reinforces the need for clearer task definitions and more consistent annotation protocols before investing in model complexity.

7.3 Overreliance on Short Text

The vast majority of existing datasets frame CD detection at the sentence or single-utterance level. This simplifies annotation and model design but introduces strong limitations, as many distortions are context-dependent or only weakly signalled lexically. By stripping away discourse-level cues, models are forced to rely on surface-level patterns and may perform poorly on more ambiguous cases. Predictably, this has contributed to the dominance of distortion types with overt markers (e.g., *Should Statements*) in both datasets and model outputs. Empirical studies confirm the value of richer context, showing that including prior conversational

turns improved detection F1 from 0.68 to 0.73 (Lybarger et al., 2022), while frameworks such as ERD achieve greater interpretability by explicitly reasoning over multi-sentence inputs (Lim et al., 2024). Still, most benchmarks continue to prioritise short-form inputs. Moving forward, we argue that context-aware models should become the norm rather than the exception. In parallel, new datasets should prioritise multi-turn conversations, real patient narratives, and longer-form content that more closely mirrors therapeutic language.

8 Best Practices and Recommendations

Alongside consolidating existing research, it is important to address the main sources of fragmentation in the field. We therefore propose a set of best practices for future work, which should be followed where possible or clearly justified if deviated from.

8.1 Taxonomy Adoption

As shown in Table 1, inconsistent use of CD taxonomies has made cross-study comparison difficult. In the absence of a universally accepted taxonomy, we recommend Burns' taxonomy[4] (Burns, 1999) as a sensible default, since it is the most widely cited and most alternatives used in NLP are partial reinterpretations of it. Researchers should (i) report the source and rationale for their chosen taxonomy, (ii) avoid introducing new or expanded taxonomies without justification, and (iii) prioritise taxonomies grounded in clinical consensus. Reliance on loosely defined online taxonomies is discouraged. Where deviations from Burns' taxonomy are necessary, the rationale should be documented in the study, or, in the case of new datasets, in the corresponding datasheet, for which we provide a template online.

8.2 Unambiguous Evaluation Reporting

Inconsistent reporting is a major barrier to comparing results across studies. To improve comparability, we recommend that future work (i) clearly state the task formulation (detection, single-label, or multi-label classification; §2), (ii) specify the analysis unit (sentence, turn, session) and, if relevant, document the exact context window, (iii) indicate the CD taxonomy used, (iv) report per-class scores alongside macro and weighted F1[5] with unambiguous labels, and (v) explain the choice of evaluation metrics. While a single metric cannot suit all applications, departures from macro or weighted F1 should always be accompanied by a clear rationale.

8.3 Dataset Development and Use

For researchers creating new CD datasets, we recommend providing a datasheet that documents the dataset's origin, annotation protocol, size, taxonomy, analysis unit, and licensing. To support this, we provide a standardised datasheet template in our GitHub repository. For reuse of existing datasets, we encourage researchers to apply our *Ethics Flowchart*, which provides practical guidance on assessing provenance, consent, and documentation before experimentation.

8.4 Annotation Reliability and Inter-annotator Agreement

Annotation processes should be reported transparently. Researchers should provide standard IAA metrics (e.g., Cohen's κ, Fleiss' κ, Krippendorff's α) rather than vague statements of 'agreement'. The rationale for the chosen metric should be stated, and partial dataset sampling for IAA is acceptable provided that procedures are clearly documented.

We further recommend that future work (i) reports human-model performance comparison metrics where possible, (ii) prioritises the release of multilingual and multi-domain corpora, and (iii) explicitly documents how annotation disagreements are resolved. Without such practices, gains in model performance may reflect noise-fitting rather than genuine progress.

8.5 Code and Dataset Release

To support replication and benchmarking, researchers should release code and, where licensing and privacy considerations permit, datasets. As shown in Table 4, many of the surveyed studies do not provide public implementations. We recommend that future work adopt code and data release as standard practice, in line with recent broader calls for stronger reproducibility standards in AI research and governance (Semmelrock et al., 2025; Mason-Williams and Mason-Williams, 2025).

9 Limitations

While this survey offers a structured overview of methods, datasets, and evaluation practices for CD detection and classification in NLP, it has several limitations. The focus is primarily computational,

[4] We provide recommended label sets online.

[5] For imbalanced or multi-label settings, AUPRC (micro or macro) and per-class PR curves may be more informative. Ultimately, metrics should fit the use case.

with limited integration of insights from clinical psychology or cognitive science, and deeper conceptual analyses of CDs are beyond its scope. The survey also centres on English-language datasets and approaches, which may limit generalisability to other languages and cultural contexts. Although emerging work on multimodal and conversational systems is noted, the emphasis remains on text-based methods and classification tasks, rather than auxiliary tasks such as cognitive reframing. Finally, due to space constraints, some datasets and methods could not be covered in detail, and despite systematic efforts, some relevant studies may have been missed.

As with any literature survey, our analysis is constrained by the scope, reporting quality, and coverage of the included studies, and should be viewed as a snapshot of a rapidly evolving field. As such, some very recent preprints may not be included. While we have aimed for balanced representation, our synthesis reflects our methodological choices and interpretive framing, which may influence the emphasis placed on particular themes.

10 Ethical Considerations

Given the nature of CDs within psychotherapy contexts, this survey acknowledges several important ethical considerations. As our work is a synthesis of existing studies, we did not collect new data or propose new models. Nevertheless, the scope of our review touches on certain areas of concern that warrant attention.

Dataset Origins. Many of the datasets discussed in this survey are derived from sources where individuals may have disclosed personal, and often highly sensitive, information. This is particularly true in domains such as digital mental health platforms, social media, and therapy transcripts. In reviewing these studies, we noted that some datasets have limited publicly available information on certain aspects of their origins or collection processes, often due to constraints inherited from upstream sources.

For instance, with THERAPISTQA, the authors provide clear documentation - including detailed labelling guidelines and procedures for resolving disagreements between annotators (Shreevastava and Foltz, 2021). However, because the dataset draws on an upstream public source, some provenance details reflect the level of information made available by that original source rather than any omission by the curators themselves. In this case, the publicly available version is based on a *Kaggle Q&A* dataset,[6] for which we have not found publicly accessible details specifying the original platform or data collection process. This situation is not unique to THERAPISTQA; several widely used mental health corpora draw on similar repositories, highlighting a broader and ongoing challenge in achieving complete transparency of data origins within the field.

Linguistic & Cultural Biases. The literature we surveyed remains heavily focused on English-language data, with only limited, though encouragingly increasing, attention paid to other languages or cultural contexts. This linguistic bias introduces significant limitations, especially given that CDs are almost certainly shaped by cultural norms, stigma, and may manifest quite differently across populations. We repeat calls from prior work for the development and evaluation of more CD classification methods that are sensitive to cross-linguistic and cross-cultural variation.

Risks of Misuse & Overreliance. We also acknowledge that the automatic detection of CDs carries serious risks if applied irresponsibly, particularly outside of therapeutic settings. Misclassification or over-reliance on automated outputs could result in harm - reinforcing stigma, invalidating personal experiences, or leading to inappropriate interventions. We therefore stress that CD classification systems should not be deployed without careful validation, the involvement of mental health professionals, and appropriate safeguards to protect user autonomy and well-being. As highlighted in Section 5, CD classification performance is variable and often limited on rarer classes. As such, applications in clinical settings should be approached cautiously.

In presenting this survey, our aim is to support the NLP community in enabling more ethically considerate, transparent, and responsible research practices - particularly when working in sensitive domains such as mental health.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number

[6] https://www.kaggle.com/datasets/arnmaud/therapist-qa/data

EP/W524475/1]. We thank the anonymous reviewers for their constructive feedback.

References

Navneet Agarwal and Kairit Sirts. 2025. Exploratory study into relations between cognitive distortions and emotional appraisals. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 127–139, Albuquerque, New Mexico. Association for Computational Linguistics.

Fatima Alhaj, Ali Al-Haj, Ahmad Sharieh, and Riad Jabri. 2022. Improving arabic cognitive distortion classification in twitter using bertopic. *International Journal of Advanced Computer Science and Applications*, 13(1).

Jelly P. Aureus, Ma. Regina Justina E. Estuar, Dorothy C. Mapua, Roland P. Abao, and Anna Angeline M. Cataluña. 2021. Determining linguistic markers in cognitive distortions from covid-19 pandemic-related reddit texts. In *2021 1st International Conference in Information and Computing Research (iCORE)*, pages 56–61.

Hakkı Halil Babacan, Ramazan Oğuz, and Yahya Kemal Beyitoğlu. 2025. Creating a clinical psychology dataset with synthetic data: Automatic detection of cognitive distortions classified with nlp. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 37(1):83–92.

Krishna C. Bathina, Marijn ten Thij, Lorenzo Lorenzo-Luaces, Lauren A. Rutter, and Johan Bollen. 2021. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 5(4):458–466. Epub 2021 Feb 11.

Aaron T. Beck. 1963. Thinking and depression. i. idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, 9(4):324–333.

Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

David D Burns. 1999. *The feeling good handbook: The groundbreaking program with powerful new techniques and step-by-step exercises to overcome depression, conquer anxiety, and enjoy greater intimacy*. Penguin.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 68–75, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Windy Dryden. 2021. *Rational emotive behaviour therapy: Distinctive features*. Routledge.

Albert Ellis. 1957. Rational psychotherapy and individual psychology. *Journal of individual psychology*, 13(1):38.

Albert Ellis. 1994. *Reason and emotion in psychotherapy, revised and updated*. Carol Publishing Group.

Nada Elsharawi and Alia El Bolock. 2024. C-journal: A journaling application for detecting and classifying cognitive distortions using deep-learning based on a crowd-sourced dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3224–3234, Torino, Italia. ELRA and ICCL.

Steven C Hayes, Kirk D Strosahl, and Kelly G Wilson. 2011. *Acceptance and commitment therapy: The process and practice of mindful change*. Guilford press.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

JunSeo Kim and HyeHyeon Kim. 2025. Koacd: The first korean adolescent dataset for cognitive distortion analysis. *Preprint*, arXiv:2505.00367.

C. Lalk, T. Steinbrenner, J.S. Pena, and 1 others. 2024. Depression symptoms are associated with frequency of cognitive distortions in psychotherapy transcripts. *Cognitive Therapy and Research*.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for efficient, explainable, and reusable systems: A case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Shuya Lin, Yuxiong Wang, Jonathan Dong, and Shiguang Ni. 2024. Detection and positive reconstruction of cognitive distortion sentences: Mandarin dataset and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6686–6701, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Lybarger, Justin Tauscher, Xiruo Ding, Dror Benzeev, and Trevor Cohen. 2022. Identifying distorted thinking in patient-therapist text message exchanges by leveraging dynamic multi-turn context. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136, Seattle, USA. Association for Computational Linguistics.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

Israel Mason-Williams and Gabryel Mason-Williams. 2025. Reproducibility: The new frontier in AI governance. In *ICML Workshop on Technical AI Governance (TAIG)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Mai Mostafa, Alia El Bolock, and Slim Abdennadher. 2021. Automatic detection and classification of cognitive distortions in journaling text. In *WEBIST*, pages 444–452.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Aaron Pico, Joaquin Taverner, Emilio Vivancos, and Ana Garcia-Fornes. 2025. Comparative analysis of the efficacy in the classification of cognitive distortions using llms. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 1: EAA*, pages 957–965. INSTICC, SciTePress.

Hongzhi Qi, Qing Zhao, Jianqiang Li, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yi Jing Yu, Fan Wang, Huijing Zou, Bing Xiang Yang, and Guanghui Fu. 2024. Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media. *Preprint*, arXiv:2309.03564.

M. Rasmy, C. Sabty, N. Sakr, and A. El Bolock. 2025. Enhanced cognitive distortions detection and classification through data augmentation techniques. In *PRICAI 2024: Trends in Artificial Intelligence*, volume 15281 of *Lecture Notes in Computer Science*, Singapore. Springer.

Lina Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health concepts derived from Cognitive Behavioural Therapy. *arXiv preprint*. ArXiv:1809.00640 [cs].

Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine learning-based research: Overview, barriers and drivers. *Preprint*, arXiv:2406.14325.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez, and C. Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 508–512.

Gopendra Vikram Singh, Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Decode: Detection of cognitive distortion and emotion cause extraction in clinical conversations. In *Advances in Information Retrieval*, pages 156–171, Cham. Springer Nature Switzerland.

Gopendra Vikram Singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal LLM-based detection and reasoning framework. In *Proceedings*

of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22546–22570, Miami, Florida, USA. Association for Computational Linguistics.

I Putu Gede Hendra Suputra, Linawati Linawati, Nyoman Putra Sastra, Gede Sukadarmika, Ngurah Agus Sanjaya ER, Diana Purwitasari, and I Made Agus Setiawan. 2023. Detection and classification of cognitive distortions: A literature review. In *2023 International Conference on Smart-Green Technology in Electrical and Information Systems (ICSGTEIS)*, pages 166–171.

JS Tauscher, K Lybarger, X Ding, A Chander, WJ Hudenko, T Cohen, and D Ben-Zeev. 2023. Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric Services*, 74(4):407–410. Epub 2022 Sep 27.

Vasudha Varadarajan, Allison Lahnala, Sujeeth Vankudari, Akshay Raghavan, Scott Feltman, Syeda Mahwish, Camilo Ruggero, Roman Kotov, and H. Andrew Schwartz. 2025. Linking language-based distortion detection to mental health outcomes. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 62–68, Albuquerque, New Mexico. Association for Computational Linguistics.

B. Wang, Y. Zhao, X. Lu, and B. Qin. 2023a. Cognitive distortion based explainable depression detection and analysis technologies for the adolescent internet users on social media. *Frontiers in Public Health*, 10:1045777.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023b. C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160, Singapore. Association for Computational Linguistics.

Katja Wiemer-Hastings, Adrian S Janit, Peter M Wiemer-Hastings, Steve Cromer, and Jennifer Kinser. 2004. Automatic classification of dysfunctional thoughts: a feasibility test. *Behavior Research Methods, Instruments, & Computers*, 36:203–212.

Zhenchang Xing, Xuejiao Zhao, and Chunyan Miao. 2017. Identifying cognitive distortion by convolutional neural network based text classification.

Jeffrey E Young, Janet S Klosko, and Marjorie E Weishaar. 2006. *Schema therapy: A practitioner's guide*. guilford press.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *Preprint*, arXiv:1710.09412.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. CBT-bench: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

A Survey Methodology

To compile a comprehensive list of relevant research publications, we drew from the following sources:

1. Searches conducted across the ACL Anthology[7], arXiv[8], PubMed[9], and IEEE Xplore[10], with no date restrictions. Search queries included terms such as *'cognitive distortion'* and *'dysfunctional thought'*.

2. Additional papers identified organically via Google Scholar, Semantic Scholar, and reference lists from relevant work.

After manual filtering, we retained 38 primary publications and preprints, spanning from 2004 to May 2025. Studies were included if they (i) implemented a CD detection or classification model, (ii) introduced a CD-related dataset, or (iii) computationally explored the taxonomy of CDs. To support reproducibility, we release our supplementary resources and paper list, along with corrections and updates, on GitHub: `https://github.com/archiesage/cognitive-distortion-nlp-survey`.

B Psychological Foundations of CDs

This appendix offers an overview of CDs, providing additional context for NLP researchers who may be unfamiliar with them.

B.1 What are CDs?

As introduced in Section 1, CDs, sometimes called *thinking errors*, are habitual patterns of negatively biased or flawed thinking that shape how people interpret events, evaluate themselves, and respond to the world (Beck, 1963). There are many different types of CDs, as outlined in Table 3. In this paper, we use the term *CD taxonomy* to refer to

[7] `https://aclanthology.org/`
[8] `https://arxiv.org/`
[9] `https://pubmed.ncbi.nlm.nih.gov/`
[10] `https://ieeexplore.ieee.org/`

the particular set of CDs adopted in any given NLP study.

CDs are not restricted to clinical settings, as most people exhibit distorted thinking, often automatically, in response to certain situations (a key aspect studied in CBT). However, in conditions such as depression, anxiety, and post-traumatic stress disorder, these patterns tend to occur more frequently (Lalk et al., 2024), become harder to shift, and carry a heightened emotional impact.

B.2 Origins

The origins of CDs are often traced back to the early work of Beck in the 1960s. While he identified fewer types of distortions than are commonly recognised today, he did describe well-known types such as *Overgeneralisation*, as well as others that are less frequently cited. Quoted directly from his work (Beck, 1963):[11]

- *Arbitrary Interpretation* - 'the process of forming an interpretation of a situation, event, or experience when there is no factual evidence to support the conclusion, or when the conclusion is contrary to the evidence.'

- *Selective Abstraction* - 'focusing on a detail taken out of context, ignoring other more salient features of the situation, and conceptualising the whole experience on the basis of this element.'

The concept of CDs gained further traction in the 1980s through the work of David Burns, who outlined a widely used taxonomy of ten CDs in *Feeling Good: The New Mood Therapy*, later revising it in 1999 (Burns, 1999). Burns' list more closely resembles the taxonomies commonly used in NLP studies today.

B.3 Context & Clinical Relevance

Although this survey focuses on computational approaches to CD detection and classification, it is important to situate this area within its broader psychological context. Several therapeutic frameworks address distorted thinking, either directly or indirectly. For example:

Rational Emotive Behaviour Therapy (REBT) where the focus is on the identification and challenge of irrational beliefs, which are broader than the discrete thought patterns typically considered as CDs. REBT focuses on core value beliefs (e.g., 'I must be the best') that tend to underlie many distortions, aiming to replace them with rational alternatives (Ellis, 1957, 1994; Dryden, 2021).

Schema therapy moves past the present-moment focus of traditional CBT by addressing deep rooted, and often unhelpful, patterns, which are known as schemas. These can develop in childhood when core emotional needs are not met. Such schemas can lead to ongoing problems in how a person thinks, feels, behaves, and relates to others, often requiring longer term and more intensive treatment (Young et al., 2006).

Acceptance and Commitment Therapy (ACT) does not frame problematic thinking in terms of CDs, but instead chooses to recognise unhelpful thoughts as a normal part of human thinking. The focus is not on *challenging* the content of these thoughts, but on changing the individual's relationship to them through various strategies (Hayes et al., 2011).

While these approaches differ in focus, they generally agree on the importance of recognising distorted thinking patterns as a route to improved emotional or behavioural regulation. This clinical grounding remains a key motivation behind the computational modelling of CDs.

C Additional Tables

[11] Both *Arbitrary Interpretation* and *Selective Abstraction* align reasonably well with the consolidated taxonomy reference in Table 1, corresponding to *Jumping to Conclusions* and *Mental Filter*, respectively.

Code	Cognitive Distortion	Description	Example	Synonyms
Burns' Taxonomy Distortions (Burns, 1999)				
AON	All or Nothing Thinking	Viewing situations in black-and-white terms, without acknowledging nuance or grey areas.	*Since our method didn't outperform all baselines in every metric, the entire study feels like a failure.*	Black and White Thinking, Polarised Thinking, Dichotomous Reasoning
DQP	Disqualifying the Positive	Rejecting positive outcomes or feedback as unimportant, accidental, or unearned.	*Our paper was accepted, but probably only because the reviewers didn't scrutinise it deeply enough.*	Discounting the Positive
EMR	Emotional Reasoning	Believing that negative emotions reflect objective truths.	*I feel uneasy about presenting this model, so it must be inherently flawed in ways I'm not seeing.*	
FTL	Fortune Telling†	Predicting negative outcomes as inevitable, without sufficient evidence.	*Given how niche our contribution is, there's no chance it will get noticed by the review committee.*	Negative Predictions, The Fortune Teller Error
JTC	Jumping to Conclusions†	Making assumptions with insufficient evidence.	*The editor's brief reply likely means they've already decided to reject our manuscript.*	Jumping to Negative Conclusions
LBL	Labelling	Defining oneself or others by a single trait or outcome.	*I misinterpreted that reviewer comment, clearly I'm not cut out for academic writing.*	Global Labelling, Labelling and Mislabelling
MAG	Magnification*	Exaggerating the significance of errors or flaws.	*This small formatting mistake will probably make the reviewers think we lack attention to detail.*	Catastrophising*
MIN	Minimisation	Downplaying the significance of positive outcomes, achievements, or strengths, reducing their perceived value or relevance.	*Sure, the paper was accepted, but it didn't get the best reviews, so it doesn't really count as a proper success.*	
MTF	Mental Filter	Focusing exclusively on negative details.	*One weakness in our ablation study keeps bothering me, despite the overall positive experimental results.*	Filtering
MDR	Mind Reading†	Assuming you know what others are thinking, often negatively.	*The session chair looked disinterested, our work must have been irrelevant to the audience.*	
OVG	Overgeneralisation	Drawing broad conclusions from a single incident.	*Since our last submission was desk-rejected, it's obvious our current work will face the same fate.*	Overgeneralising
PRS	Personalisation	Attributing external events or failures entirely to oneself.	*The collaboration didn't materialise, probably because my proposal wasn't convincing enough.*	Personalisation and Blame, Personalising, Blaming Oneself
SHD	Should Statements	Holding rigid expectations about how oneself or others ought to behave.	*I should always produce novel ideas quickly, taking this long feels like professional incompetence.*	Shoulds, Inflexibility
Other Distortions				
BRT	Being Right	Placing too high value on proving yourself correct, often at your own or others' expense.	*I'm certain my annotation guidelines are the best. Any disagreement from the team simply indicates they don't understand the task properly.*	Always Being Right
BLM	Blaming	Attributing too high responsibility for negative outcomes to others, avoiding self-reflection or your own shared responsibility.	*The demo crashed because the organisers didn't provide adequate technical support, not because of any oversight on our side.*	Blaming Others
CAT	Catastrophising*	Imagining worst-case scenarios and exaggerating potential negative consequences far beyond their realistic likelihood.	*If this preprint has a minor oversight, it could irreparably damage our lab's reputation and future collaborations.*	
CMP	Comparing	Measuring self-worth against others in a way that undermines your own accomplishments.	*Another lab published a similar paper first - clearly they're much more capable researchers than we are.*	Comparing and Despairing, Comparison
CTL	Control Fallacy	Believing either complete control over everything or total helplessness in a situation, without middle ground.	*If I don't oversee every single preprocessing step myself, the entire pipeline will end up flawed.*	
FOC	Fallacy of Change	Assuming others should or will change to meet your own personal expectations.	*If only the dataset creators had annotated according to our taxonomy, our analysis would be so much clearer.*	Control of Fallacies
FOF	Fallacy of Fairness	Presuming life or systems must work in a way that aligns with personal standards of fairness.	*It's unfair that methodologically weaker papers receive more attention just because they're trendy.*	
HRF	Heaven's Reward Fallacy	Expecting a guaranteed reward for one's hard work.	*After months of hyperparameter tuning, this model surely deserves to be the new state-of-the-art.*	
LFT	Low Frustration Tolerance*	Overestimating the severity of minor inconveniences.	*Dealing with this reviewer rebuttal feels impossible. I can't imagine going through it again.*	
NFE	Negative Feeling or Emotion	Taking emotional discomfort as proof something is wrong.	*Feeling stuck while writing this paper draft surely means the research itself is inherently flawed.*	

Table 3: Categories of CDs observed in computational research. Descriptions and examples are reflective of common interpretations of these distortions in NLP contexts. All examples are fictional and not about any specific work or group. To ensure consistency across studies, we also include synonyms and related terms where applicable. † Jumping to Conclusions (JTC) is frequently considered a parent category that includes Fortune Telling (FTL) and Mind Reading (MDR). * Although Magnification (MAG) and Catastrophising (CAT) are often treated as equivalent, we list them separately to highlight subtle conceptual distinctions, following prior work (Lalk et al., 2024; Agarwal and Sirts, 2025). Similarly, Low Frustration Tolerance (LFT), while similar to CAT, is presented as a distinct category.

Dataset†	Language	Subdomain	Size (# Samples)*	Labelling‡	Annotators	Agreement	Access
Literature Examples							
Wiemer-Hastings et al. (2004)	English	Psychology literature	261	Single-label (10)	Expert	–	Private
Social Media							
Alhaj et al. (2022)	Arabic	Twitter	9,250	Single-label (5)	Non-Expert (Unspecified)	$\kappa = 0.817_c$	Private
SocialCD-3K, Qi et al. (2024)	Mandarin	Weibo 'Zoufan' blog	3,407	Multi-label (12)	Domain-Informed	–	Public[1]
Aureus et al. (2021)	English	Reddit: r/COVID19_support	586	Binary (2)	Mixed	–	Private
Simms et al. (2017)	English	Tumblr	459	Binary (2)	Mixed	–	Private
Digital Mental Health Platform							
Rojas-Barahona et al. (2018)	English	Koko	4,035	Multi-label (15)	Expert	$\kappa = 0.61_c$	Private[2]
Lin et al. (2024)	Mandarin	PsyQA counselling forums	4,001	Binary (2)	Domain-Informed	$\mathrm{JP} = 0.88_d$	Public[3]
TherapistQA, Shreevastava and Foltz (2021)	English	–	2,529	Multi-label (10)	Non-Expert (Unspecified)	$\mathrm{JP} = 0.34_c, 0.61_d$	Public[4]
MH-D, Shickel et al. (2020)	English	TaoConnect	1,799	Binary (2)	Domain-Informed	–	Private
MH-C, Shickel et al. (2020)	English	TaoConnect	1,164	Single-label (15)	Domain-Informed	–	Private
CBT-CD, Zhang et al. (2025)	English	Patient-therapist QA	146	Multi-label (10)	Expert	–	Public[5]
Crowd-sourced							
Elsharawi and El Bolock (2024)	English	–	34,370	Single-label (14)	Expert	–	Private
PatternReframe, Maddela et al. (2023)	English	MTurk, Mephisto	9,688	Multi-label (10)	Crowd-Generated	$\alpha = 0.355_c$	Public[6]
CrowdDist, Shickel et al. (2020)	English	MTurk	7,666	Single-label (15)	Crowd-Generated	–	Private
C2D2, Wang et al. (2023b)	Mandarin	–	7,500	Single-label (7)	Crowd-Generated	$\kappa = 0.67_c$	Request[7]
Thinking Trap, Sharma et al. (2023)	English	–	600	Multi-label (13)	Expert	–	Public[8]
Synthetic							
GPT-4 Synthetic, Babacan et al. (2025)	English	GPT-4	2,000	Single-label (10)	Automated (LLM)	–	Public[9]
Clinical Intervention							
Lalk et al. (2024)	German	CBT psychotherapy transcripts	104,557	Multi-label (14)	Automated (Lexicon)	–	Request[10]
Lybarger et al. (2022)	English	Patient-therapist text exchanges	7,436	Multi-label (5)	Expert	$\kappa = 0.53_d$	Private
Hybrid (Mixed Domains)							
KoACD, Kim and Kim (2025)	Korean	NAVER Knowledge iN + LLM	108,717	Single-label (10)	Automated (LLM)	$\kappa = 0.78$	Request[11]
GPT-4 Combined, Babacan et al. (2025)	English	GPT-4 synthetic + TherapistQA	4,530	Single-label (10)	Automated (LLM)	–	Request[12]
CoDEC, Singh et al. (2023)	English	Real + staged patient-therapist videos	3,773	Binary (2)	Non-Expert (Unspecified)	$F = 0.83_d$	Request[13]
CoDER, Singh et al. (2024)	English	Real + staged patient-therapist videos	3,773	Binary (2)	Trained	$F = 0.83_d$	Public[14]
Wang et al. (2023a)	English	Lit. examples + social media augment	3,644	Single-label (11)	Automated (BERT)	–	Private
Mostafa et al. (2021)	English	Twitter, Surveys, HappyDB	2,409	Single-label (2)	Domain-Informed	–	Private

Table 4: Extended overview of datasets for CD detection and classification, grouped by domain. Agreement metrics: κ = Cohen's kappa; α = Krippendorff's alpha; F = Fleiss's kappa; JP = joint probability; $_d$ = detection; $_c$ = classification. '–' indicates not applicable or not reported. † Corpus name, or earliest study to use it for CD tasks. * Number of annotated units (e.g., posts, speech turns); for automated methods, items processed. ‡ Number of CD categories used, excluding 'Undistorted' for classification.

[1] https://github.com/HongzhiQ/SupervisedVsLLM-EfficacyEval/tree/main/data/SocialCD-3k
[2] https://github.com/YinpeiDai/NAUM
[3] https://github.com/405200144/Dataset-of-Cognitive-Distortion-detection-and-Positive-Reconstruction/tree/main
[4] https://www.kaggle.com/datasets/sagarikashreevastava/cognitive-distortion-detetction-dataset
[5] https://huggingface.co/datasets/Psychotherapy-LLM/CBT-Bench
[6] https://github.com/facebookresearch/ParlAI/tree/main/projects/reframe_thoughts
[7] https://github.com/bcwangavailable/C2D2-Cognitive-Distortion
[8] https://github.com/behavioral-data/Cognitive-Reframing
[9] https://huggingface.co/datasets/halilbabacan/cognitive_distortions_gpt4
[10] https://osf.io/rsy4z/?view_only=41dc962f0c924c0e87e7bfc044535bd3
[11] https://github.com/cocoboldongle/KoACD
[12] https://huggingface.co/datasets/halilbabacan/combined_synthetic_cognitive_distortions
[13] https://www.iitp.ac.in/~ai-nlp-ml/resources.html#DeCoDE-CoDEC
[14] https://github.com/clang1234/ZS-CoDR.git