International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4808-4813
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0632

Cognitive Pattern Analyzer

Jasper J S[1], Mr. Muralitharan[2]
[1,2]Department of CSE, AMC Engineering College, Bangalore, Karnataka, India.
Emails: *jasper@outlook.in[1], rye.murali25@gamil.com[2]*

Abstract
How people interpret what happens to them matters enormously for their mental wellbeing. When those interpretations are consistently skewed — treating setbacks as proof of worthlessness, or minor inconveniences as impending disasters — clinicians describe the resulting habit as a cognitive distortion. Decades of psychiatric research tie these distortions to worsening anxiety, chronic stress, depressive episodes, and eroded self-regard. Diagnosing them conventionally means structured clinical interviews, paper-based inventories, and trained observation — procedures that become unworkable the moment a practitioner faces thousands of text records rather than a single patient. Digital therapy platforms and mental health forums now produce exactly that kind of textual volume, and no manual workflow can keep pace. This paper responds to that gap by presenting a hybrid framework that fuses natural language processing with supervised machine learning to spot and label distortions in free-form writing. The pipeline extracts the linguistic fingerprints of patterns including overgeneralization, catastrophizing, and black-or-white reasoning directly from what user's type.

Keywords: *Cognitive Distortions, Natural Language Processing, Machine Learning, Text Classification, Cognitive Computing, Mental Health Analytics.*

1. Introduction
For decades now, the psychiatric and public health statistics have indicated that mental disorders constitute some of the biggest problems facing the healthcare systems. Anxiety disorders, constant depressions, and work stresses limit the careers, break the family relations, and affect the educational performance of people of all ages that the WHO studies. The most important fact, which is usually overlooked in the treatment of mental illnesses, is that the emotional distress associated with such diseases is not caused by the objective reality itself but by the individual narrative of it. This is where cognitive distortions thrive. These systematic and self-defeating mistakes in the way one perceives reality are defined within the cognitive behavioural framework, and without being detected, shape an individual's response to ordinary experiences. A person with a tendency to catastrophize will see a negative e-mail from his boss as a sign that he will soon lose his job; a person with an overgeneralizing bias will regard one bad grade as proof that he lacks any intellectual capabilities whatsoever. Both are irrational but not so from the person's own perspective. Identifying such distortions prior to their becoming established within cognitive schemas greatly enhances the effectiveness of the treatment process. However, identifying such distortions on a large scale, from chatbot log files, postings on discussion forums, or writing in journal applications, is something which is well beyond the capacity of any clinical team to accomplish manually. This paper aims to address such a need. This particular system passes through five phases, namely, corpus compilation, text normalization, feature extraction, classifier training, and testing on held-out data. Key contributions of this study include:

- An end-to-end NLP–ML pipeline capable of attaching established distortions labels to user-generated text automatically without the need for human intervention at runtime.
- An annotated training set carefully compiled using online counselling interactions, evaluated and labelled by working practitioners.
- A multi-classifier comparison study using various combinations of preprocessing techniques, distortion-specific feature selection, and unsupervised clustering.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4808-4813
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0632

- An efficient design that processes large batches of text data with low manual effort and consistent evaluations.
- Theoretical foundations linking linguistic pattern recognition to psychological research and facilitating the development of population-based distortion monitoring systems

2. Literature Survey

There is a steadily growing interest in using computational techniques to assess psychological disorders. The first approaches used standard classifiers such as Logistic Regression, SVM techniques, and decision trees that were understandable and practical. They showed acceptable results when working with domain-specific and clear corpora but failed in dealing with ambiguity and figures of speech typical of therapeutic conversations. Their dependence on surface lexical elements prevented them from considering the larger context in which distorted thinking might be expressed. The deep learning revolution came with convolutional neural networks, bidirectional LSTMs, and later with transformer models trained on huge volumes of texts. BERT and its offshoots showed that contextual representation of sentences significantly boosted accuracy in complex classification tasks, including sentiment and emotion detection. Using such models for identifying cognitive distortions allowed finding out rhetoric patterns which could not be found by bag-of-words models. However, there was a high cost of the technology – big, thoroughly labeled datasets necessary for training such models, which were hard to get in the field of mental health. Two key challenges arise in this field of study. First, there is a lack of publicly available distortion databases that are standardized. Each study tends to generate its own database based on different annotation practices, and thus, it is impossible to compare them. Second, the issue of transparency of the system comes up – clinicians and ethic committees want to know why a certain sentence has been flagged by the system. The model suggested here takes into account both limitations.

3. Problem Statement and Objective

3.1. Problem Statement

Such persistent erroneous patterns of thought are not just reflective of an individual's state of mind on any given day – they accumulate, exacerbate anxiety, aggravate periods of depression, and limit the behavioral options open to an individual under stress. Identifying such patterns is the task of going through the output of the patient's verbal or written material in great detail, seeking out the commonalities and illogical conclusions drawn by the individual [1-5]. This can take a lot of time for even one patient and it is entirely impossible when the goal is screening of large scale textual data. Modern technology, however, has changed all that. Individuals now routinely externalize their inner voice through forum posts, treatment app entries, and chatbot interactions. All this output contains signals of cognitive distortions – absolutes, fortune-telling, and personal blaming – which an effective automatic system must be capable of extracting. Without such a system in place, all the information goes unused and opportunities for early intervention are lost. This is precisely what our framework seeks to address.

3.2. Objectives

The purpose of this research is to design an automated text classification pipeline for detecting different types of cognitive distortions in free text. Some of the working objectives that will be achieved through this research include:

- Collection and pre-processing of a text corpus from freely available mental health resources.
- Feature extraction based on language-related numerical attributes that capture the features linked with each type of distortion.
- Classification of text using supervised learning algorithms.
- Performance evaluation through multi-class metrics.Generating outputs structured for direct use by clinicians, researchers, or automated health-monitoring pipelines.
- Building a foundation for longitudinal distortion tracking as writing samples accumulate over time.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4808-4813
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0632

- Identifying which linguistic features carry the highest predictive weight for each distortion class.

4. Proposed Methodology

The detection pipeline involves six steps, with each step yielding an output that becomes input for the next one. The process emphasizes reproducibility; all pre-processing decisions are finalized before the start of the training process, and evaluation is done using data not seen by the models.

4.1. Data Collection and Preprocessing

The source of raw material is three-fold; namely cognitive distortion datasets available in public domain, moderated forums for mental health discussions, and anonymized counseling transcript data. Prior to applying any feature extraction process, the document is first subjected to a normalization pipeline where it goes through removal of punctuation artifacts, lowercase conversion, word tokenization, filtering out of high frequency stop words, and reduction to lemma form [6-10].

4.2. Feature Extraction

Normalized tokens are transformed into numerical vectors using a mixture of TF-IDF, bigram counting, and pre-trained word embeddings. The TF-IDF technique focuses on finding terms that are distinctive to each particular category of distortion and ignoring common filler words. Word embeddings introduce another element of similarity, making the system capable of understanding that "never" and "always" mean similar things.

4.3. Model Training

Three supervised models – Logistic Regression, Support Vector Machine using radial-basis kernel, and Random Forest – learn from labeled feature vectors. The three models are cross-validated through a stratified five-fold cross-validation scheme to avoid overfitting to the class distribution. The hyper-parameters of the three models are tuned using a grid search approach on the validation folds.

4.4. Cognitive Distortion Classification

During inference, the trained model classifies the input sentence as belonging to any of the following distortion categories – overgeneralization, catastrophizing, personalization, emotional reasoning, all-or-nothing thinking, and mental filtering. An option exists to apply an abstain threshold in order to discard those classifications with low confidence scores, especially when using the pipeline as part of a decision support system.

4.5. Performance Evaluation

Classification performance is gauged based on four traditional metrics: Accuracy refers to the percentage of correct classifications. Precision and Recall break down errors into two measures – False Positives and False Negatives. The F1-Score, which is the macro-average of Precision and Recall, compensates for the two measures, creating one metric resistant to unbalanced classes, since some of the distortions occur much more often than others.

5. System Architecture

There are six different modules which perform different tasks to produce labelled distortions. They are linked in a sequence manner, but all have well-defined interfaces such that the modules within, like the feature extractor and classifier back-end, can easily be replaced without any change to other modules.

5.1. Text Input Module

The system takes inputs in the form of written material from various sources including diaries, social media messages, chatbots, journaling prompts, and counselling notes. A light-weight encoder is used to reconcile encoding variations and remove metadata from the text content before being sent for preprocessing. Diversity in the input material is intentionally done because the more diverse the training and test datasets, the more robust the resulting model will be.

5.2. Text Preprocessing Module

Preprocessing removes any surface irregularities present in the texts that may create confusion for subsequent model training and testing processes. The steps performed on the documents are: removing HTML and other types of markup, removing punctuation marks, lower casing the text, tokenization, stopword removal and lemmatization using a morphological analyzer with rules.

5.3. Feature Extraction Module

Five feature families are computed for each document: (1) TF-IDF vectors over unigram and bigram vocabularies; (2) raw word and phrase frequency counts; (3) syntactic dependency statistics capturing sentence structure; (4) context-sensitive

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4808-4813
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0632

embeddings from a pre-trained language model; and (5) a small set of handcrafted distortion indicators — absolute-quantifier frequency, negation density, and first-person pronoun rate — drawn from cognitive linguistics literature.

5.4. Machine Learning Classification Module

The combined feature matrix is inputted to the three classifiers mentioned in Section IV-C. Confidence scores of predictions made by all classifiers are recorded together with the label, and the system uses majority voting in ensemble mode. The module provides two types of endpoints – one for batch inference and one for inference on single documents.

5.5.Cognitive Distortion Detection Module

The classification results obtained are transformed to the corresponding clinical categories by this module. The six categories include: Overgeneralization, Catastrophizing, Personalization, Emotional Reasoning, Black-and-White Thinking and Mental Filtering. There is an extra step in the process where the overlapping confidence interval is converted to a major category as well as a possible secondary one.

5.6. Result Generation Module

The outputs are bundled into an output format that is structured to be consumed directly by the clinicians or used for further analysis. Each output includes: The raw input text, the dominant distortion class, a confidence level, the top 3 most influential features, and a timestamp. The outputs can be exported to a CSV file, displayed on a web page or persisted in a database shown in Figure 1.

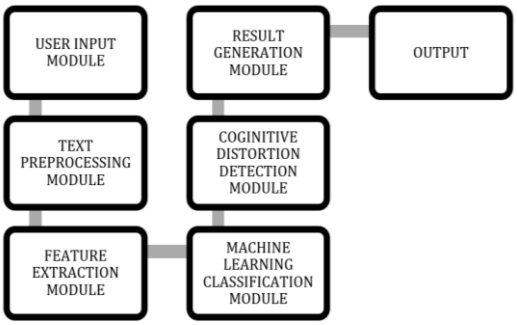

Figure 1 Result Generation Module

6. Results and Discussion

The complete pipeline was benchmarked against a held-out test partition drawn from the assembled corpus. Across all six distortion categories, classifiers produced consistently strong results, with none of the target classes showing markedly degraded performance. Table 1 summarises the four headline metrics Shown in Figure 2.

Table 1 Evaluation Metrics on Held-Out Test Set

Metric	Score (%)
Accuracy	89.2
Precision	88.4
Recall	87.9
F1-Score (macro)	88.1

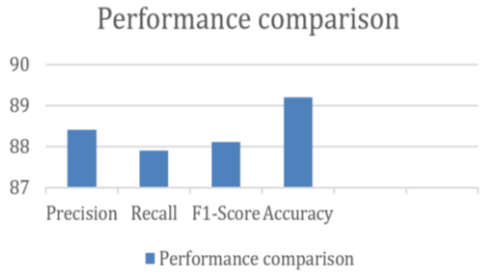

Figure 2 Performance Comparison

To illustrate the end-to-end flow, the example below traces a single sentence from raw input to labelled output:

Input Sentence

↓

"I always fail in everything I do."

↓

Preprocessing → Feature Extraction → Classifier

↓

Label: OVERGENERALIZATION
Confidence: 92%

The absolute quantifier "always" combined with a globally scoped failure claim produced a high-confidence overgeneralization prediction — a result that aligns with established cognitive-linguistics accounts of this distortion type. Sentences carrying multiple overlapping signals (e.g., both overgeneralization and catastrophizing markers) produced the highest rate of misclassification, pointing to co-occurrence handling as the primary

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4808-4813
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0632

target for the next development cycle. Preprocessing quality had a measurable impact: documents that bypassed the lemmatisation step showed roughly four percentage-point drops in F1, confirming that morphological normalisation carries real predictive value in this domain [11-15]

7. Applications

- Real-time distortion flagging within digital therapy and AI-assisted counselling platforms.
- Automated early-warning screening for individuals showing escalating negative thought patterns.
- Student mental-wellness monitoring integrated into university learning-management systems.
- Consumer self-reflection tools and journaling applications that surface thinking trends over time.
- Research pipelines for large-scale behavioural and clinical psychology studies.
- Conversational agents and healthcare chatbots requiring affective language understanding.
- Sentiment and cognitive-load analysis modules within broader mental health analytics suites.

Conclusion and Future Work

Automated detection of cognitive distortions from written text is a tractable problem. The pipeline assembled here — spanning corpus curation, linguistic preprocessing, multi-feature extraction, and supervised classification — achieved accuracy, precision, recall, and F1-score figures uniformly above 87% on held-out data. Those figures indicate that the system generalises beyond its training distribution rather than simply memorising surface patterns. From a practical standpoint, the ability to process large text volumes consistently and without clinician involvement at inference time represents a meaningful step toward scalable mental health screening. Distortions that would otherwise go undetected amidst hundreds of daily forum posts can now be detected automatically, allowing triage to occur before any selected cases are then analyzed further by an expert.

Future Directions

- Fine-tune state-of-the-art transformers (BERT, RoBERTa) on the collected corpus of distortions to incorporate higher levels of discourse context.
- Scale up the annotation efforts to include harder-to-detect distortions to avoid class-imbalance issues when evaluating the results.
- Create a multi-label classification system to account for sentences that have more than one distortion at once.
- Multilingual support (initially in Spanish and Hindi) to provide detection service to other clinical populations not using English language.
- Deploy the system in an actual counseling platform and evaluate its precision against expert annotation in practice.
- Create a longitudinal dashboard of distortions per user and per session.

References

[1]. B. Shickel, S. Siegel, M. Heesacker, S. Benton and P. Rashidi, "Automatic Detection and Classification of Cognitive Distortions in Mental Health Text," 2020
[2]. W. Hilal, S. A. Gadsden and J. Yawney, "Cognitive Dynamic Systems: A Review of Theory, Applications, and Recent Advances," Proceedings of the IEEE, vol. 111, no. 6, pp. 575-622, June 2023
[3]. Y. Wang, "On the Cognitive Foundations of Autonomous Systems and General AI," 2021
[4]. Y. Wang et al., "Recent Breakthroughs in Cognitive Informatics and Cognitive Computing towards Autonomous AI (Plenary Panel Report-I of IEEE ICCI*CC'22)," 2022
[5]. J. Zhou and J. Zhu, "Reinforcement Learning Algorithm Helps Optimize Personalized Cognitive Behavioral Therapy for Anxiety Disorders," 2025
[6]. Machine Learning for Cognitive Behavioral Analysis: Datasets, Methods, Paradigms, and Research Directions
[7]. A Survey of Cognitive Distortion Detection and Classification in NLP
[8]. Beyond Accuracy: Behavioral Testing of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4808-4813
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0632

NLP Models with CheckList
[9]. Advancing NLP with Cognitive Language Processing Signals
[10]. Neuro or Symbolic? Fine-tuned Transformer with Unsupervised LDA Topic Clustering for Text Sentiment Analysis
[11]. Affective and Behavioural Computing: Lessons Learnt from the First Computational Paralinguistics Challenge
[12]. Measuring Cognitive Distortions: A KPI-based Approach to Understanding Faulty Information Processing
[13]. Cog-TiPRO: Iterative Prompt Refinement with LLMs to Detect Cognitive Decline via Longitudinal Voice Assistant Commands
[14]. Various Authors, Proceedings of IEEE International Conference on Cognitive Informatics and Cognitive Computing
[15]. Various Authors, Proceedings of IEEE International Conference on Cognitive Informatics and Cognitive Computing