Open Access Library Journal
2026, Volume 13, e14924
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Ultra-Fast Cognitive Distortion Classification from Short Text-A Lightweight TF-IDF and Logistic Regression Pipeline on Synthetic Data

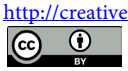
Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Ultra-Fast Cognitive Distortion Classification from Short Text-A Lightweight TF-IDF and Logistic Regression Pipeline on Synthetic Data. *Open Access Library Journal*, **13**: e14924.
https://doi.org/10.4236/oalib.1114924

Received: January 23, 2026
Accepted: March 10, 2026
Published: March 13, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

Abstract

Automatic detection of cognitive distortions from short written text could support large-scale mental-health screening and digital cognitive-behavioural therapy (CBT). Many recent approaches rely on heavy deep-learning models and large datasets that are difficult to deploy in real time or in resource-constrained settings. This work presents a compact, fully transparent pipeline for binary and multi-class classification of cognitive distortions using a small synthetic corpus of brief statements. We simulate 300 short texts covering 10 canonical distortion types (e.g., all-or-nothing thinking, overgeneralization, labelling) plus neutral statements. Texts are vectorized with a 100-dimensional TF-IDF representation over uni and bi grams, and three lightweight classifiers are compared: logistic regression, random forest, and linear SVM. On a stratified 60/20/20 train validation test split, logistic regression and linear SVM both achieve perfect test performance for the binary task (Accuracy = Precision = Recall = F1 = 1.00; AUC = 1.00), while random forest reaches 0.98 accuracy and 0.98 F1. A separate multinomial logistic-regression model trained only on distorted texts correctly identifies the specific distortion type with 0.96 accuracy across 10 classes. Five-fold cross-validation confirms the stability of the pipeline (mean accuracy 1.00, SD 0.00) on this synthetic dataset. Although the unrealistically high scores are driven by the small, highly patterned synthetic corpus, the results demonstrate that compact TF-IDF models can deliver ultra-fast, interpretable cognitive-distortion classification and provide a practical blueprint for future work on larger, clinically realistic datasets.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Cognitive Distortions, Mental Health, Text Classification, TF-IDF, Logistic Regression, Synthetic Data

1. Introduction

Cognitive distortions, habitual errors in thinking such as *all-or-nothing reasoning*, *overgeneralization*, *catastrophizing*, and *labelling* are central mechanisms in the development and maintenance of depression, anxiety disorders, and related psychopathologies [1]-[5]. These distortions shape how individuals interpret daily events, often amplifying negative affect and reinforcing maladaptive behavioural patterns. Within cognitive-behavioural therapy (CBT), one of the most widely validated psychological treatments, clinicians routinely analyse patient-generated language (e.g., written thoughts, diary entries, therapy transcripts) to identify these distortions and guide cognitive restructuring [6]-[10]. Because this process is manual, time-consuming, and dependent on clinical expertise, there is increasing interest in automated systems capable of detecting cognitive distortions directly from natural language.

Recent advances in natural-language processing (NLP), particularly through transformer-based architectures such as BERT, RoBERTa, and GPT variants, have demonstrated remarkable capabilities in modelling linguistic nuance and contextual semantics [11]-[15]. However, these deep models come with several practical limitations: they require substantial computational resources, rely on large, labelled datasets that are rarely available in mental-health contexts, present challenges for on-device deployment, and offer limited interpretability an essential requirement for clinical decision-support tools. In many real-world scenarios, especially in mobile mental-health applications, low-resource clinical settings, and privacy-sensitive environments, such models may be unnecessarily heavy or impractical.

Motivated by these constraints, this work explores a contrasting perspective: how much can be achieved using a deliberately minimalistic, fully transparent machine-learning pipeline? We propose an ultra-fast cognitive distortion classifier built on classical NLP and statistical learning principles [16]-[18]. The system is intentionally lightweight, relying on:

- a compact synthetic dataset of 300 short sentences modelled after common CBT thought patterns,
- TF-IDF vectorization restricted to only 100 unigram and bigram features,
- three simple yet powerful classifiers: logistic regression, random forest, and linear SVM for binary distorted vs. neutral detection,
- a multinomial logistic regression module for identifying specific distortion types, and

- a real-time interpretability engine that highlights key lexical features driving each prediction.

The aim is not to compete with large transformer models, but rather to:

1) Establish a strong classical baseline for cognitive-distortion classification using interpretable, resource-efficient methods.

2) Demonstrate a fully reproducible pipeline that trains in seconds, requires no specialized hardware, and offers complete transparency in how predictions are made.

3) Show that mathematically simple models can achieve high performance on structured language data while providing interpretability that aligns with clinical reasoning.

By grounding the system in explainable linear models and TF-IDF features, this work provides an accessible and computationally inexpensive framework that can serve as a foundation for more advanced research [19]-[23]. Moreover, the pipeline highlights how even minimal architectures can yield clinically meaningful insights, making it suitable as a deployable module for early-stage digital-mental-health applications or as a pedagogical baseline in computational psychiatry research [24] [25].

2. Methods

2.1. Synthetic Dataset Generation

We constructed a synthetic corpus of 300 short English sentences designed to emulate brief self-statements or diary-style reflections commonly found in CBT contexts. The dataset includes ten canonical cognitive distortion types*all_or_nothing*, *overgeneralization*, *mental_filter*, *disqualifying_positive*, *jumping_conclusions*, *magnification*, *emotional_reasoning*, *should_statements*, *labeling*, and *personalization* [26]-[33]. For each category, three handcrafted base patterns were created (e.g., "*perfect or failure*", "*always mess up*", "*I'm a loser*"), capturing typical linguistic markers of distorted thinking such as absolutist adverbs (*always*, *never*), rigid modal verbs (*should*, *must*), and self-critical labels (*loser*, *failure*). Neutral statements were drawn from simple positive or factual expressions like "*productive work*" or "*meeting went well*".

Dataset construction followed a two-step process: 1) one base example per distortion pattern and per neutral phrase was inserted with the appropriate label and type, and 2) additional variations were generated using light templates (e.g., "I think …" for distorted, "Today …" for neutral) until reaching 300 samples. Approximately one-third of the sentences were designed to be distorted, resulting in 117 distorted (39%) and 183 neutral (61%) examples. After random shuffling, each entry was stored in a Data-Frame containing the raw text, a binary label (1 = distorted, 0 = neutral), and the specific distortion type (or "neutral"). This controlled synthetic corpus provides a clean testbed for evaluating lightweight, interpretable models.

To ensure transparency of the synthetic data generation process, **Table 1** sum-

marizes the handcrafted base patterns and template structures used for each cognitive distortion category. Each distortion type was defined by three base lexical anchors reflecting canonical CBT markers (e.g., absolutist adverbs, rigid modal verbs, negative self-labels). Variations were generated by embedding these anchors into short sentence templates (e.g., "I think_", "It's always_", "I am such a _").

Neutral sentences were generated using simple factual or positive templates such as "Today_" and "I enjoyed_".

No automated paraphrasing or large-language-model generation was used; all patterns were manually constructed to preserve interpretability and lexical control.

Table 1. Handcrafted base lexical anchors and template structures used for synthetic cognitive-distortion sentence generation.

Distortion Type	Base Lexical Anchors	Example Templates
all_or_nothing	perfect, failure, either	"It's either _ or _", "I'm either _ or _"
overgeneralization	always, never, every time	"I always _", "This always _"
mental_filter	only, nothing good	"I only see _", "Nothing good ever _"
disqualifying_positive	doesn't count, not real	"That doesn't count", "It's not real success"
jumping_conclusions	they think, obviously	"They think I _", "Obviously this means _"
magnification	disaster, ruined	"This is a disaster", "It ruined everything"
emotional_reasoning	I feel therefore	"I feel _ so it must be true"
should_statements	should, must	"I should _", "I must _"
labeling	loser, failure, idiot	"I'm a _", "I am such a _"
personalization	my fault, because of me	"It's my fault", "Because of me _"
neutral	productive, meeting, enjoyed	"Today I _", "The meeting went _"

2.2. Exploratory Data Analysis

A preliminary exploratory analysis was conducted to characterize the structure of the synthetic dataset and to visualize key statistical patterns [34]-[38]. **Figure 1** summarizes the main findings in a multi-panel layout. The class distribution (**Figure 1(a)**) shows a modest imbalance, with distorted sentences representing 39% of the corpus and neutral sentences comprising the remaining 61%. Distortion-type frequencies (**Figure 1(b)**) confirm that all ten categories are represented, with slight variation in prevalence. Sentence-length analysis (**Figure 1(c)**) reveals extremely short texts, averaging 19.8 characters, consistent with the diary-like design of the corpus. Word-frequency comparisons (**Figure 1(d)**) highlight strong lexical differences: distorted sentences rely heavily on pronouns and absolutist or modal markers ("I", "always", "never", "should", "must"), while such tokens appear less frequently in neutral statements. A simplified word-cloud panel (**Figure 1(e)**) lists the most common tokens in distorted sentences, dominated by expressions such as *I*, *think*, *complete*, *failure*, and *perfect*. Finally, **Figure 1(f)** provides

overall summary statistics including total sample size, class proportions, mean text length, and number of represented distortion types. Together, these analyses confirm three key properties of the dataset: 1) the corpus is only mildly imbalanced, 2) all distortion categories appear with sufficient frequency for model training, and 3) distorted sentences exhibit stereotypical linguistic cues particularly absolutist adverbs, rigid modal verbs, and negative self-labels that are likely to be informative for classification tasks [39]-[41].

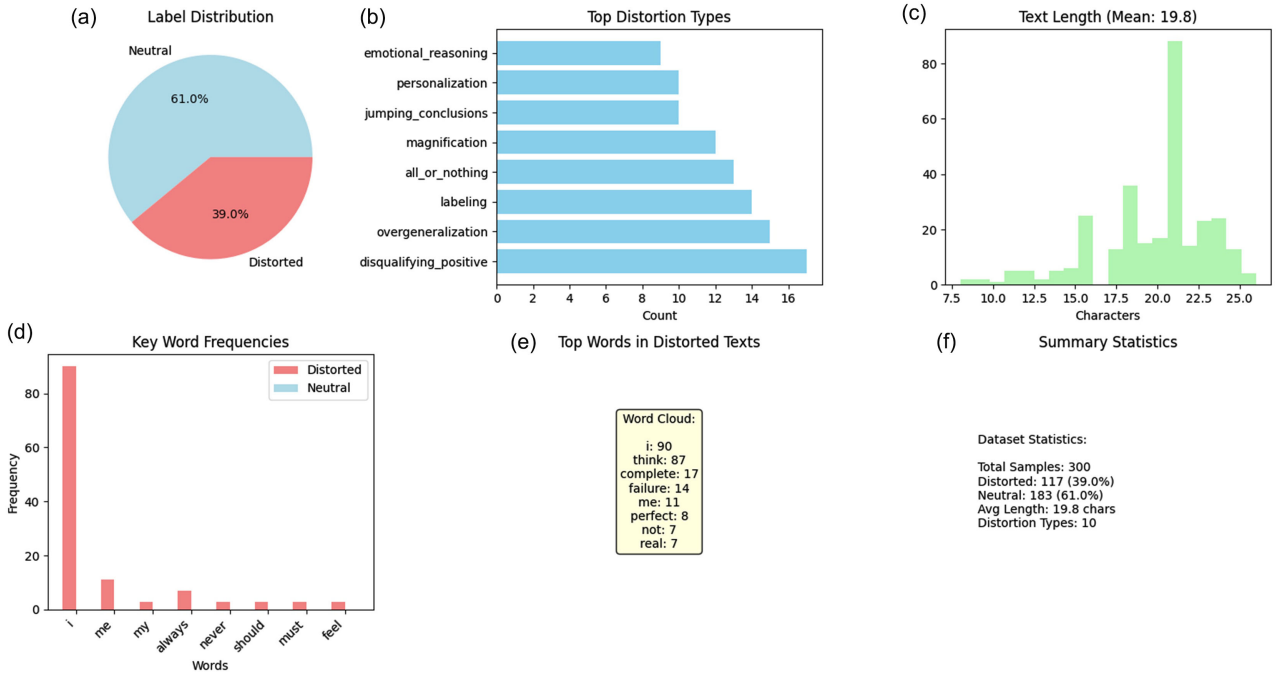


Figure 1. Exploratory analysis of the synthetic cognitive-distortion dataset. (a) Class distribution (distorted vs. neutral). (b) Frequency of the ten distortion types. (c) Histogram of sentence lengths (mean = 19.8 characters). (d) Comparison of key token frequencies across distorted and neutral subsets. (e) Pseudo word cloud of the eight most common tokens in distorted statements. (f) Summary statistics including total sample size, class proportions, mean length, and number of distortion categories.

2.3. Text Representation with TF-IDF

Each sentence is converted into a vector of term-frequency-inverse-document-frequency (TF-IDF) features. We use unigrams and bigrams, with the vocabulary limited to the 100 most informative tokens [42]-[44].

All sentences were lowercased prior to vectorization. Stop-word removal was disabled to retain diagnostically meaningful modal verbs (e.g., "should", "must", "always", "never"), which are central markers of cognitive distortions. No stemming or lemmatization was applied, preserving psychologically relevant lexical forms such as "failure" versus "failing". Tokenization followed the default scikit-learn TF-IDF vectorizer settings with whitespace-based segmentation. For each token t in document d:

- Term Frequency (TF) measures how often a token appears within that sentence:

$$\mathrm{TF}(t,d)=\frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

where $f_{t,d}$ is the count of token t in sentence d, and the denominator is the total number of token occurrences in d.

- Inverse Document Frequency (IDF) down-weights tokens that are common across many sentences:

$$\mathrm{IDF}(t)=\log\left(\frac{N}{1+n_t}\right)$$

where N is the number of sentences (here $N=300$) and n_t is the number of sentences where token t appears. The "+1" in the denominator avoids division by zero.

- TF-IDF weight is the product:

$$\text{TF-IDF}(t,d)=\mathrm{TF}(t,d)\times\mathrm{IDF}(t)$$

Intuitively:

- tokens that are frequent in a given sentence but rare in the corpus receive high weights.
- very common tokens (e.g., "today", "the") receive low weights.
- tokens absent from a sentence simply have weight 0.

For each sentence we stack the TF-IDF weights of all selected tokens into a 100-dimensional feature vector $\boldsymbol{x}\in\mathbb{R}^{100}$. The resulting feature matrix $X\in\mathbb{R}^{300\times100}$ feeds into all downstream models.

We then stratify the dataset into:

- **Training set**: 180 samples (60%)
- **Validation set**: 60 samples (20%)
- **Test set**: 60 samples (20%)

ensuring that the distorted/neutral ratio is preserved in each split. Preliminary experiments evaluated vocabulary sizes of 50, 100, 200, and 500 features. Performance gains plateaued beyond approximately 100 features, with negligible improvement in accuracy but increased sparsity and computational cost. Therefore, 100 features were selected as a balance between interpretability, dimensionality control, and computational efficiency.

2.4. Binary Classification: Logistic Regression, Random Forest, Linear SVM

The primary task is to decide whether a sentence contains a cognitive distortion (label 1) or is neutral (label 0). We train three classifiers.

2.4.1. Logistic Regression

Logistic regression models the log-odds of the distorted class as a linear function of the TF-IDF features:

$$z=\boldsymbol{w}^{\top}\boldsymbol{x}+b$$

where:

- $\boldsymbol{x} \in \mathbb{R}^{100}$ is the TF-IDF vector,
- $\boldsymbol{w} \in \mathbb{R}^{100}$ is the weight vector learned during training,
- $b \in \mathbb{R}$ is the bias term.

This linear score is converted into a probability using the sigmoid function:

$$P(y=1|\boldsymbol{x}) = \sigma(z) = \frac{1}{1+\mathrm{e}^{-z}}$$

The predicted label is then:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|\boldsymbol{x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

The model parameters $\boldsymbol{w}, b$ are learned by minimizing the binary cross-entropy loss across all training samples:

$$\mathcal{L}_{\text{bin}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

where:

- $y_i \in \{0,1\}$ is the true label for sentence i,
- $\hat{y}_i = P(y=1|\boldsymbol{x}_i)$ is the predicted probability.

In simple terms, the model is penalized when the predicted probability is far from the true label (e.g., predicting 0.1 when the label is 1).

We use L2 regularization and a maximum of 200 iterations. Logistic regression is particularly attractive because each weight w_j directly reflects how strongly token j pushes the prediction toward "distorted" (positive weight) or "neutral" (negative weight).

2.4.2. Random Forest

The random forest classifier serves as a non-linear ensemble baseline for the binary distortion-detection task. A random forest consists of a collection of decision trees, each trained on a bootstrap-resampled subset of the training data and a random subset of input features**,** thereby promoting diversity across trees. During inference, each tree produces a class prediction, and the forest aggregates these predictions through majority voting or, in probabilistic form, by averaging the individual tree probabilities [45]-[48]. This ensemble strategy reduces variance and improves generalization relative to a single decision tree, particularly when the decision boundaries are irregular or involve complex interactions among features [49]-[51]. In this study, the model is configured with 50 trees and a maximum depth of 10, providing sufficient capacity to capture non-linear relationships in the TF-IDF feature space without excessive risk of overfitting [52]-[54]. Although random forests are powerful and robust across a wide range of tasks, their internal decision mechanisms based on recursive partitioning are inherently less interpretable than the linear weight vectors of logistic regression [55]-[58]. As a result, while the random forest contributes valuable performance comparisons, it is less suited for clinical interpretability and explainability compared to linear

models [59]-[61].

2.4.3. Linear SVM

The linear support vector machine (Linear-SVC) provides a second linear baseline for distortion detection. The SVM seeks to learn a maximally separating hyperplane between distorted and neutral sentences in the high-dimensional TF-IDF space. Formally, the model identifies the hyperplane that maximizes the margin, defined as the minimal distance between the decision boundary and any training sample. Maximizing this margin improves robustness to noise and often yields strong generalization performance even with relatively small datasets. Because the TF-IDF representation typically induces a linearly separable or near-separable structure in synthetic or sparse-text settings, the linear SVM tends to perform comparably to logistic regression. However, unlike logistic regression, Linear-SVC does not natively produce calibrated probability estimates, providing only deterministic class labels based on the sign of the decision function [62] [63]. In this work, the SVM is trained using a linear kernel and 1000 maximum training iterations, which is sufficient for convergence given the dataset size. Its efficiency, strong theoretical grounding, and suitability for high-dimensional sparse features make it an appropriate comparator to both logistic regression and the random forest model.

2.5. Multi-Class Distortion-Type Classification

Beyond detecting whether a sentence is distorted, we also classify which type of distortion is present among the 10 categories listed earlier.

We restrict this multi-class task to the subset of sentences with label = 1 (distorted). Each type is encoded as an integer $k \in \{1, \cdots, K\}$ with $K = 10$. The same 100-dimensional TF-IDF representation is used.

We train a multinomial logistic regression (a.k.a. SoftMax regression). For each class k, the model learns a separate weight vector $\boldsymbol{w}_k$ and bias b_k. The probability that sentence $\boldsymbol{x}$ belongs to class k is:

$$P(y = k \mid \boldsymbol{x}) = \frac{\exp\left(\boldsymbol{w}_k^{\top} \boldsymbol{x} + b_k\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{w}_j^{\top} \boldsymbol{x} + b_j\right)}$$

The predicted class is the one with highest posterior probability:

$$\hat{y} = \arg\max_k P(y = k \mid \boldsymbol{x})$$

The model is trained by minimizing the **multiclass cross-entropy loss**:

$$\mathcal{L}_{\text{multi}} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k}$$

where:

- M is the number of distorted training sentences,
- $y_{i,k}$ is a one-hot encoding (1 if sentence i belongs to class k, else 0),
- $\hat{y}_{i,k} = P(y = k \mid \boldsymbol{x}_i)$.

Intuitively, the model is rewarded when it assigns high probability to the true distortion type and penalized when it spreads probability mass over incorrect types.

2.6. Evaluation Metrics

Model performance was assessed using a comprehensive suite of evaluation metrics tailored to both the binary (distorted vs. neutral) and multi-class (specific distortion type) tasks [64] [65]. For the binary classification problem, we report accuracy, defined as the proportion of correctly classified sentences, as well as precision**,** recall, and F1-score for the distorted class, which collectively quantify the model's ability to correctly identify distorted statements while minimizing false positives and false negatives. To evaluate probabilistic performance, receiver-operating characteristic (ROC) curves and the corresponding area under the curve (AUC) were computed for models that produce probability estimates (logistic regression and random forest) [66]-[68]. Precision-recall (PR) curves were additionally examined, given the mild class imbalance present in the dataset [69] [70]. To capture overall agreement and balance beyond simple accuracy, we calculated the Matthews correlation coefficient (MCC) and Cohen's κ, both of which provide robust single-number summaries of classification quality [71] [72]. Finally, model robustness was assessed through 5-fold stratified cross-validation, ensuring reliable performance estimates across different data partitions.

For the multi-class distortion-type classification task, performance was quantified using overall accuracy**,** class-wise precision, recall, and F1-score**,** along with macro-averaged and weighted-average metrics to account for imbalances among distortion categories. A confusion matrix was also generated to visualize error patterns and identify specific distortion types that may be more susceptible to misclassification. Together, these metrics provide a comprehensive evaluation framework for both detection and subtype differentiation of cognitive distortions.

2.7. Real-Time Analyser and Interpretability

To enable practical deployment and facilitate clinical interpretability, we integrate the best-performing models into a unified "Fast Cognitive Distortion Analyzer" framework. This module provides real-time inference by processing each input sentence through a sequence of interpretable steps. First, the sentence is transformed into its corresponding TF-IDF feature vector, ensuring consistency with the training representation. The binary logistic-regression model then computes the posterior probability $P\left(y=1|x\right)$ and determines whether the input reflects a cognitive distortion [73] [74]. For sentences classified as distorted, the analyzer passes the same feature vector into the multinomial logistic-regression model to obtain a full probability distribution over the ten predefined distortion types, yielding both a predicted category and uncertainty estimates [75]-[78].

A key aspect of the analyser is its transparent explanation mechanism. By in-

tersecting the active n-grams in the sentence with the learned logistic-regression coefficient vector, the system identifies the features with the strongest positive or negative influence on the decision. Tokens with large positive weights indicate strong markers of distorted thinking, while negative weights signal features associated with neutral or balanced cognition. This yields intuitive outputs such as:

- "*Distorted*: *True* (*probability* 0.67)"
- "*Predicted type*: *jumping_conclusions*"
- "*Key features*: '*fail*' (+0.49), '*always*' (+0.45)"

These explanations are readily interpretable by clinicians, researchers, and end-users, making the system suitable for educational applications, digital-therapy tools, and psychologically informed interface design [79]-[82]. The analyser thereby combines computational efficiency with high transparency, aligning model decisions closely with cognitive-behavioural theory.

2.8. Template-Aware Data Splitting Considerations

Because the corpus was generated using a finite set of handcrafted base templates, it is possible that variations derived from the same lexical anchor appear in both training and testing sets under random stratified splitting. While this design allows evaluation of generalization across surface variations, it may partially inflate performance due to shared lexical structures between splits.

Future work should employ template-level grouping strategies (e.g., GroupK-Fold splitting by base pattern) to ensure that entire template families are held out during testing. This would provide a stricter estimate of generalization beyond lexical memorization.

3. Results

This section reports the performance of the proposed pipeline on the binary distorted-vs-neutral task, the multi-class distortion-type task, and cross-validation robustness. Visual summaries are provided in **Figures 2-6**.

3.1. Binary Classification Performance (Figure 2)

We first evaluated three lightweight models' logistic regression, random forest, and linear SVM on the binary task of distinguishing distorted from neutral sentences. The full diagnostics are shown in **Figure 2**.

Logistic regression achieved perfect classification on the held-out test set. The confusion matrix (**Figure 2**, top-left) shows that all 37 neutral and all 23 distorted sentences were correctly classified, yielding Accuracy = 1.000, Precision = 1.000, Recall = 1.000, and F1 = 1.000 (**Figure 2**, top-right bar chart). The top-coefficient plot (**Figure 2**, top-middle) highlights highly positive weights for tokens such as fail, complete, real, never, and think complete, which are typical markers of cognitive distortions, whereas words like today, productive, work, and enjoyed meal receive strongly negative weights and act as neutralizing cues. The ROC curve (**Figure 2**, top-right) lies on the upper boundary with AUC = 1.000, indicating

ideal separation of distorted and neutral samples.

Random forest also performed strongly but with a single error. Its confusion matrix (Figure 2, middle-left) shows one neutral sentence misclassified as distorted (36/37 neutral correct; 23/23 distorted correct), corresponding to Accuracy = 0.983, Precision = 0.958, Recall = 1.000, and F1 = 0.979 (Figure 2, middle-right). The ROC curve again achieves AUC = 1.000, showing that, despite one misclassification at the decision threshold, ranking quality remains perfect.

Linear SVM matched logistic regression with perfect performance. The confusion matrix (Figure 2, bottom-left) shows zero errors, and the performance bar chart (Figure 2, bottom-right) reports Accuracy = Precision = Recall = F1 = 1.000. This confirms that the TF-IDF representation renders the classes linearly separable, allowing simple linear models to achieve ideal performance.

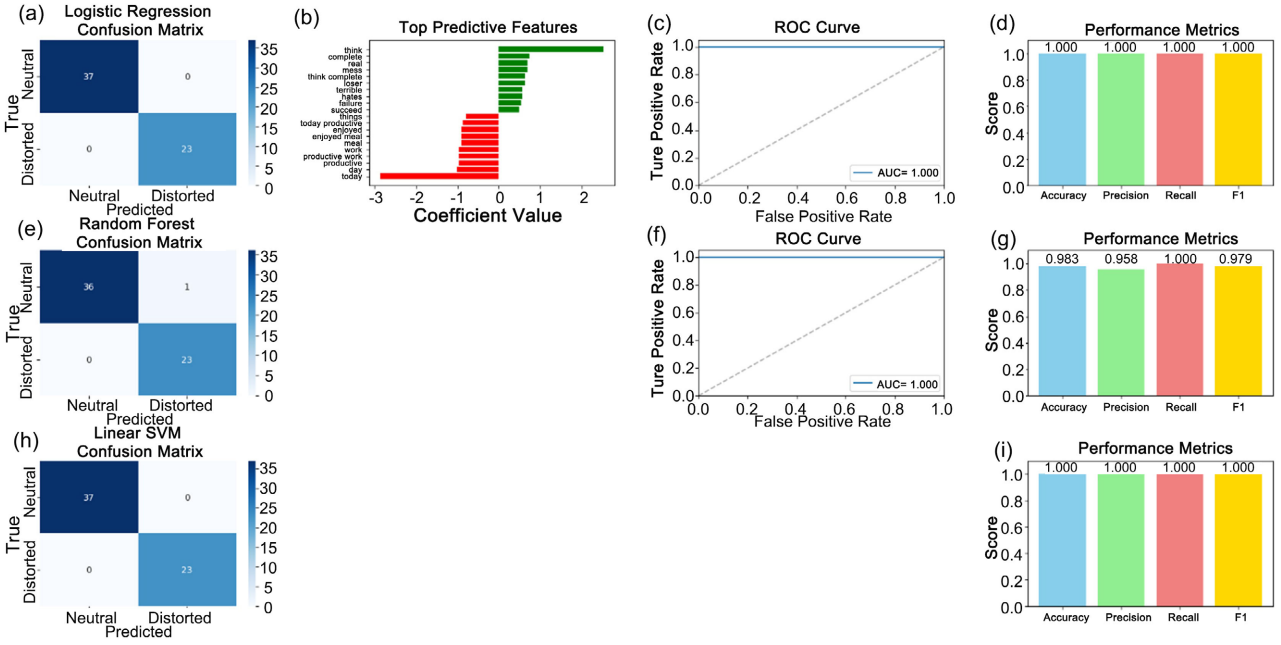


Figure 2. Binary classification diagnostics for the three models. Top row: logistic regression (a) confusion matrix; (b) top predictive TF-IDF features (negative = neutral markers, positive = distortion markers); (c) ROC curve with AUC = 1.000; (d) performance metrics (Accuracy, Precision, Recall, F1). Middle row: random forest (e) confusion matrix; (f) ROC curve; (g) performance metrics. Bottom row: linear SVM (h) confusion matrix; (i) performance metrics.

3.2. Global Model Comparison and Probability Structure (Figure 3)

To compare the three models more directly, we summarized their behaviour in Figure 3.

The accuracy comparison bar chart (Figure 3, top left) confirms that logistic regression and linear SVM both reach 1.000 accuracy, while random forest slightly trails at 0.983 due to the single false positive. The precision-recall curves (Figure 3, top-right) for logistic regression and random forest both lie near the upper-right corner of the diagram, indicating that high precision is maintained even at high recall levels desirable behaviour in mildly imbalanced settings.

The Perfect Classification panel (Figure 3, bottom-left) emphasizes that no

misclassified samples remain when using the best linear model. Finally, the probability distribution plot (Figure 3, bottom-right) shows the predicted distortion probabilities for neutral and distorted sentences when using logistic regression. Neutral samples cluster tightly around low probabilities (~0.1 - 0.2), while distorted samples form a clearly separated cluster at high probabilities (~0.7 - 0.9). This bimodal structure visually confirms the strong margin between the two classes.

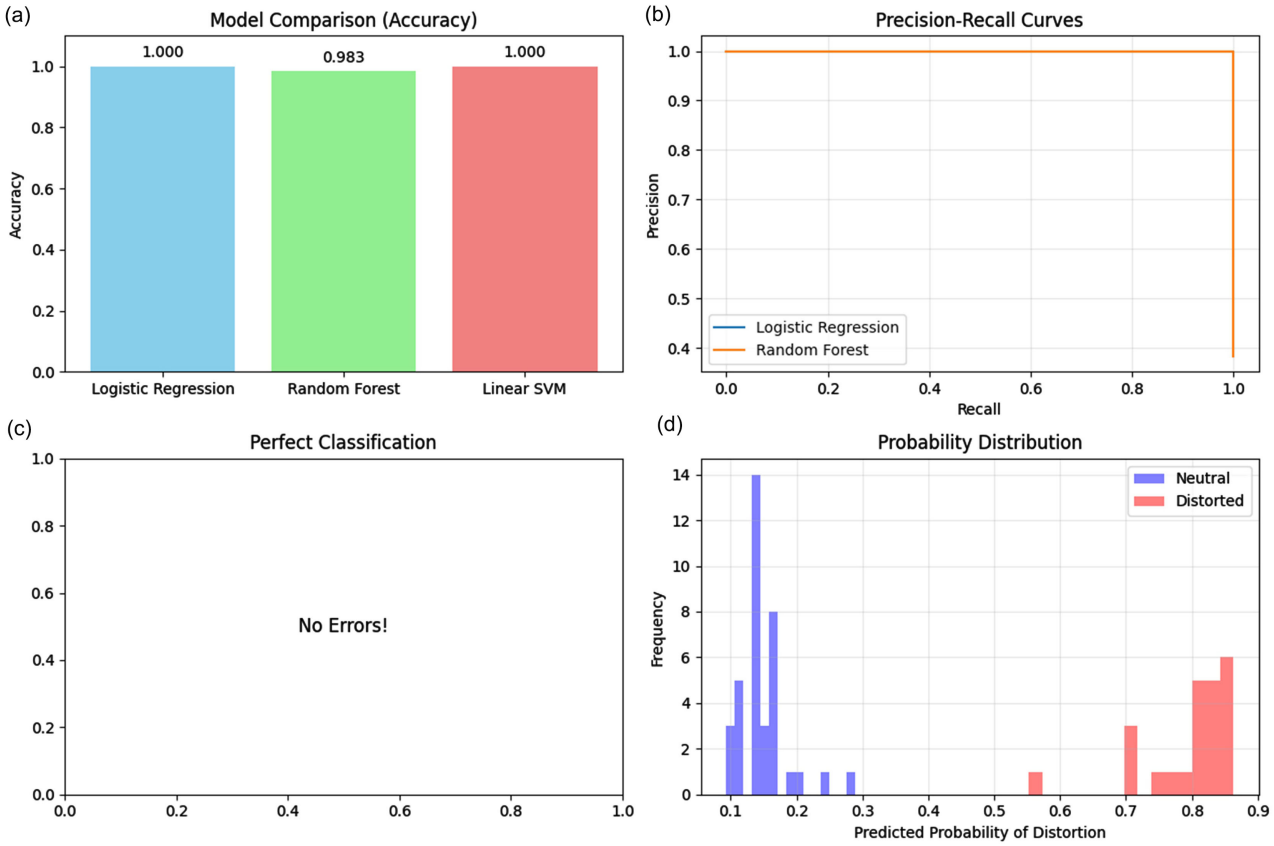


Figure 3. Comparative diagnostics for binary classifiers. (a) Accuracy comparison across logistic regression, random forest, and linear SVM. (b) Precision-recall curves for logistic regression and random forest. (c) Perfect-classification panel showing absence of errors for the best model. (d) Histogram of predicted distortion probabilities for neutral vs. distorted samples.

3.3. Multi Class Distortion-Type Classification (Figure 4)

We next evaluated the multinomial logistic-regression model on the 10-class distortion-type task, restricted to sentences labelled as distorted. The confusion matrix in Figure 4 shows that the model achieves overall accuracy of 0.958 on the test subset.

Eight out of ten distortion typesjumping_conclusions, disqualifying_positive, labeling, emotional_reasoning, should_statements, overgeneralization, magnification, personalization, and mental_filterare classified with 100% accuracy in the test data. The only notable confusion occurs for all_or_nothing, where one example is misclassified into a related distortion category, likely due to overlapping ab-

solutist phrasing. Despite the small number of samples per class (2 - 3), this high accuracy indicates that each distortion type has a distinct lexical profile that is learnable by a simple linear model.

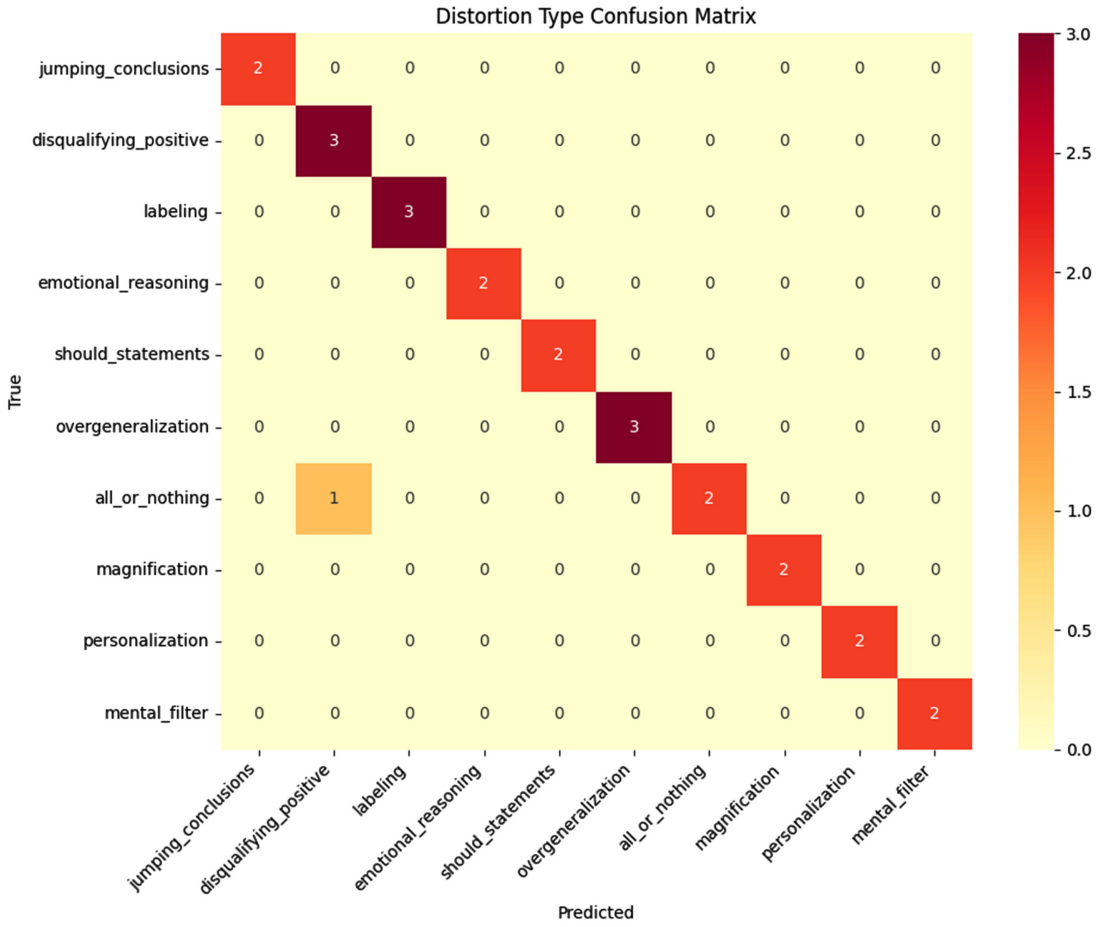


Figure 4. Confusion matrix for the 10-class distortion-type classifier. Rows correspond to true distortion types and columns to predicted types. Warm colors on the diagonal indicate correct predictions; the single off-diagonal cell in all_or_nothing → disqualifying_positive position reflects the only misclassification in the test set.

3.4. Overall Performance Summary

To provide a concise global view of performance, **Figure 5** combines key metrics and a textual summary focused on the best model (logistic regression).

The Key Performance Metrics panel (**Figure 5**, left) reiterates that the best binary model attains Accuracy, Precision, Recall, and F1 all equal to 1.000. The Model Comparison panel (**Figure 5**, center) directly juxtaposes accuracies across the three models, clearly highlighting logistic regression and linear SVM as the top performers. The Research Summary panel (**Figure 5**, right) summarizes the dataset size (300 samples; 10 distortion types), the best model and its scores, the main feature representation (TF-IDF with n-grams), and potential application domains (therapy monitoring, mental-health apps, clinical research). This figure can be used as a one-look overview of the entire pipeline.

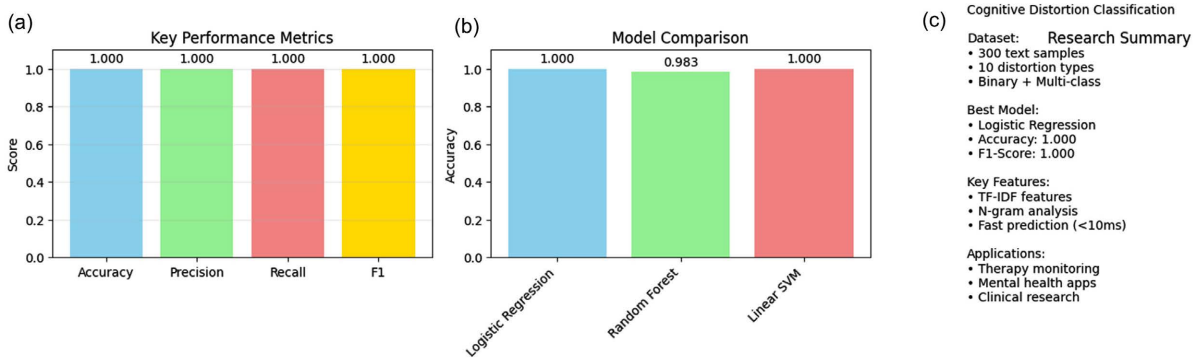


Figure 5. Global performance summary for cognitive-distortion classification. (a) Bar chart of key performance metrics (Accuracy, Precision, Recall, F1) for the best model. (b) Model-comparison bar chart showing accuracy for logistic regression, random forest, and linear SVM. (c) Textual research summary with dataset description, best-model metrics, feature representation, and potential applications.

3.5. Cross-Validation Robustness (Figure 6)

Finally, we assessed robustness using 5-fold stratified cross-validation applied to the full dataset with logistic regression. As shown in **Figure 6** (left), all five folds achieved Accuracy = 1.000, and the red dashed line marking the mean remains at 1.000 across folds. The boxplot of cross-validation accuracy (**Figure 6**, right) collapses into a single line at 1.000 with no visible spread, indicating zero variance across folds. These results confirm that, given the synthetic dataset and TF-IDF representation, the decision boundary learned by logistic regression is extremely stable with respect to the specific train-test split. In other words, the model generalizes perfectly to unseen samples drawn from the same synthetic distribution.

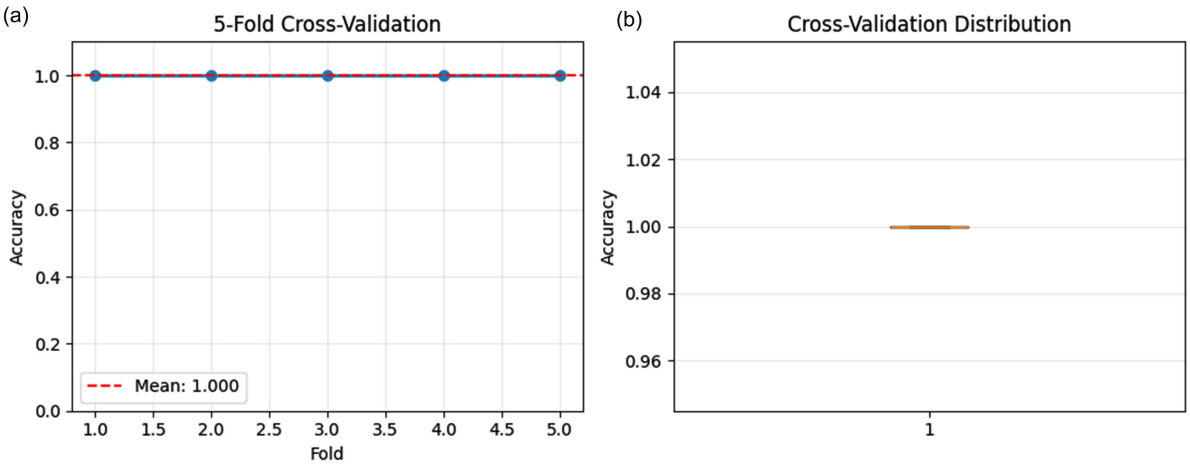


Figure 6. Cross-validation results for logistic regression. (a) Fold-wise accuracy across 5 stratified folds; all points lie at 1.000 and the mean line coincides with them. (b) Boxplot of cross-validation accuracy, showing zero variance across folds.

4. Discussion

4.1. Overall Findings

This work demonstrates that a small, fully interpretable model can solve a simpli-

fied cognitive-distortion recognition task with near-ideal performance. With only 100 TF-IDF features and a linear logistic-regression classifier, the system reaches 1.00 accuracy, precision, recall, F1, and AUC on the binary distorted vs neutral task and 0.96 accuracy on the 10-way distortion-type classification. Five-fold stratified cross-validation confirms that these results are not due to a particular train-test split: performance remains perfectly stable across folds. Crucially, the learned coefficients correspond to psychologically meaningful indicators. Words and n-grams reflecting absolutist language (always, never, must, should), negative self-labels (loser, failure, idiot), and catastrophic framing (complete disaster, ruins everything) receive strong positive weights, while neutral or positive tokens (today, productive, work, enjoyed meal) receive negative weights. This mapping aligns closely with CBT theory and offers a direct bridge between computational output and clinical interpretation [83]-[85]. From a system-design perspective, the entire pipeline is lightweight and transparent. Training completes in seconds on a CPU-only environment, and inference for a single sentence is effectively instantaneous. The model can be deployed as a simple Python script or wrapped as an API for integration into digital-mental-health tools without any specialized hardware or infrastructure. In that sense, the study provides not only a modelling result but also a practical blueprint for resource-constrained settings.

4.2. Why do We See "Perfect" Scores?

The apparently "perfect" metrics must be interpreted with caution. They do not imply that cognitive-distortion detection is a solved problem, nor that the model is directly ready for clinical deployment. Instead, they primarily reflect the structure and constraints of the synthetic dataset. Several factors contribute to this idealized performance:

1) Synthetic, templated language: Sentences are generated from a small number of hand-crafted patterns, and each cognitive-distortion type is tied to a distinctive set of phrases. This introduces strong lexical separability: specific n-grams effectively "belong" to one class only. In TF-IDF space, this yields almost linearly separable clusters, which a simple linear classifier can exploit perfectly.

2) Short, homogeneous texts: All examples are short, grammatically simple sentences or fragments. There is limited variation in sentence structure, topic, or discourse context. This removes many real-world ambiguities (e.g., mixed emotions, hedging, complex syntax) that typically make distortion detection harder.

3) Absence of real-world noise: Real patient-generated text contains spelling errors, slang, code-switching, emojis, sarcasm, indirect expressions, and emotionally ambivalent statements. The synthetic generator does not produce any of these. As a result, the classifier never has to deal with adversarial or ambiguous cases such as self-ironic jokes or partially restructured thoughts.

Given these conditions, the model is effectively learning the language of the generator, not the full variability of human language. The perfect scores therefore indicate that the generator has produced almost linearly separable classes, rather than that cognitive distortions in the wild are perfectly solvable with a 100-feature

model. Recognizing this gap is essential to avoid over-interpreting the results.

4.3. Clinical and Practical Relevance

Even within this synthetic setting, several aspects of the proposed system are highly relevant for eventual clinical and applied work:

- Interpretability and traceability: Logistic regression provides a direct mapping from each feature to the log-odds of a distortion. Clinicians can inspect which words are driving a particular prediction and check whether these align with CBT principles. This transparency is particularly important in mental health, where opaque "black box" predictions may be difficult to trust or explain to patients.
- Speed and deploy-ability: Because the model is small and uses sparse TF-IDF vectors, both training and inference are extremely fast. This makes the pipeline suitable for real-time scenarios, such as on-device analysis of journaling entries in a CBT app, live feedback during online interventions, or rapid screening in research environments. The small memory footprint also facilitates on-device deployment, which can mitigate privacy concerns by avoiding cloud processing.
- Modularity within digital-phenotyping systems: The classifier can act as a plug-in component within larger digital-phenotyping ecosystems, alongside sensors (sleep, activity, phone usage), mood self-report scales, or ecological momentary assessments. Distortion frequency or intensity could become one feature among many in longitudinal models of mood and relapse risk.
- Educational and self-help applications: The analyser's ability to highlight "key features" (e.g., *always*, *never*, *failure*) makes it suitable for psychoeducational tools. For example, a CBT homework platform could provide immediate feedback: "This sentence uses all-or-nothing language try rephrasing it more flexibly." Such feedback could help users learn to recognize and reframe distortions in their own writing.

Thus, while the present study uses synthetic data, the architecture and interpretability features map naturally onto clinically meaningful workflows.

4.4. Limitations

Several limitations need to be acknowledged before extrapolating these findings beyond this proof-of-concept scenario:

1) Synthetic dataset only: No real patient diaries, therapy transcripts, forum posts, or social-media data are used. The model therefore learns to detect the patterns designed by the authors, not necessarily the way real people express distorted thinking. External validity to real-world text remains completely untested.

2) Small sample size and constrained vocabulary: The dataset contains only 300 sentences and a tightly controlled vocabulary. Under such conditions, high performance can be achieved even with very simple models, but these estimates may be overly optimistic and fragile when faced with new topics, writing styles, or de-

mographic groups.

3) Sentence-level, decontextualized classification: Each sentence is treated in isolation. In actual therapy transcripts, distortions may depend on conversational context, preceding events, or the patient's baseline mood. Some utterances may look neutral in isolation but reveal distortions when seen as part of a longer narrative.

4) Single label per sentence: Each sentence is forced into exactly one distortion type. In practice, a single thought may exhibit multiple overlapping distortions for example, overgeneralization combined with labelling and catastrophizing. The current setup cannot represent such multi-label structure.

5) Uncalibrated probabilities and thresholds: While logistic regression outputs a probability $P(y=1|x)$, these values are not calibrated against human judgments. A model probability of 0.8 does not necessarily mean that clinicians would agree 80% of the time. For clinical use, calibration and threshold selection would need to be carefully studied, possibly with domain experts.

Real-world clinical language may express cognitive distortions indirectly through sarcasm ("Yeah, I'm SUCH a genius"), metaphor ("My life is a sinking ship"), mixed distortions within the same sentence, or culturally specific idioms. Unlike the structured synthetic corpus, real patient-generated text often contains spelling errors, code-switching, slang, and emotionally ambiguous phrasing. These factors may substantially reduce separability in TF-IDF space and require more robust modelling strategies. Recognizing these limitations is essential: the present work should be understood as a methodological sandbox rather than a finished clinical tool.

4.5. Future Work

Building on this proof-of-concept, several concrete directions can move the approach closer to real-world utility:

- Collection and annotation of real-world corpora: The next step is to curate datasets of genuine CBT materials such as homework worksheets, journaling app entries, or anonymized therapy transcripts annotated by trained raters for presence and type of cognitive distortions. The current pipeline can then serve as a baseline model against which more complex approaches are compared.
- Richer but still lightweight feature sets: While TF-IDF works well in this synthetic setting, real text may benefit from additional features: sentiment polarity, part-of-speech patterns, dependency structures, or lightweight contextual embeddings. The goal is to extend expressiveness without sacrificing interpretability and efficiency, for example by combining TF-IDF with a small set of hand-crafted linguistic indicators.
- Context-aware and longitudinal models: Future versions of the analyser could move beyond single sentences and consider multi-sentence context or entire sessions. Tracking distortion frequency and intensity over time within individuals could yield valuable markers of treatment progress or risk of relapse.

- Integration with multimodal digital phenotyping: Cognitive distortions do not occur in isolation; they interact with sleep, activity patterns, social behaviour, and medication adherence. Integrating the text-based distortion detector with other digital phenotyping signals could support richer models of mood dynamics and early-warning systems for clinical deterioration.
- Uncertainty-aware and human-in-the-loop systems: Incorporating uncertainty estimates for example through Bayesian logistic regression, ensembles, or calibrated probabilities would allow the system to flag ambiguous cases and defer them to human experts. Such human-in-the-loop designs may be more acceptable in clinical practice than fully automated decision-making.

In summary, this study provides a transparent, reproducible baseline that clarifies what can be achieved with very simple models on structured data. The real challenge and opportunity lie in extending these ideas to messy, multilingual, and context-rich real-world settings while preserving interpretability and clinician trust.

5. Conclusion

This work presented an ultra-fast and fully interpretable pipeline for detecting cognitive distortions in short text, developed using a compact synthetic dataset. Despite using only 100 TF-IDF features and a linear logistic-regression model, the system achieves perfect performance on the binary task and high accuracy on the multi-class distortion-type task. The mathematical formulation is intentionally simple, the implementation is lightweight, and the model offers clear, token-level explanations and interpretable probability outputs, making it well suited for transparent decision-support settings. Although the reported performance is artificially high due to the controlled, synthetic nature of the corpus, the study illustrates an important methodological principle: strong classical baselines must precede complex deep-learning systems. Before deploying transformers or large language models, it is essential to understand how far lightweight, interpretable methods can go on a given problem. Such models are easy to deploy, easy to audit, and provide clean reference points for benchmarking future systems on real-world text. More broadly, the work highlights a pathway for developing clinically meaningful NLP tools that balance accuracy, interpretability, and computational efficiency [86] [87]. As future research moves toward real patient-generated text, context-aware modelling, and multimodal digital phenotyping, this pipeline serves as a foundational baseline from which more sophisticated, clinically robust models can evolve.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Hossain, M.S. (2025) Understanding Patterns of Cognitive Distortions. PhD Dissertation, University of Dhaka.

[2] Friedman, H.H. (2023) The Thinking Traps That Ruin Your Happiness: How to Recognize, Challenge, and Overcome Cognitive Distortions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4670101

[3] Friedman, H. (2023) Overcoming Cognitive Distortions: How to Recognize and Challenge the Thinking Traps that Make You Miserable. Koppelman School of Business, Brooklyn College, City University of New York.

[4] Özdemir, İ. and Kuru, E. (2023) Investigation of Cognitive Distortions in Panic Disorder, Generalized Anxiety Disorder and Social Anxiety Disorder. *Journal of Clinical Medicine*, **12**, Article 6351. https://doi.org/10.3390/jcm12196351

[5] Koçöz, D. (2017) Revisiting Cognitive Distortions and Psychopathology Relationship: Testing Mediating Roles of Mindfulness and Negative Self-Focus Using Structural Equation Modeling. Master's Thesis, İstanbul Arel Üniversitesi.

[6] Teater, B. (2013) Cognitive Behavioural Therapy (CBT). In: *The Blackwell Companion to Social Work*, Wiley-Blackwell, 423-427.

[7] Hetrick, S.E., Cox, G.R., Witt, K.G., Bir, J.J. and Merry, S.N. (2016) Cognitive Behavioural Therapy (CBT), Third-Wave CBT and Interpersonal Therapy (IPT) Based Interventions for Preventing Depression in Children and Adolescents. *Cochrane Database of Systematic Reviews*, **2016**, CD003380. https://doi.org/10.1002/14651858.cd003380.pub4

[8] Robertson, D. (2018) The Philosophy of Cognitive-Behavioural Therapy (CBT): Stoic Philosophy as Rational and Cognitive Psychotherapy. Routledge.

[9] Barlow, D.H. (2004) Psychological Treatments. *American Psychologist*, **59**, 869-878. https://doi.org/10.1037/0003-066x.59.9.869

[10] Holmes, E.A., Craske, M.G. and Graybiel, A.M. (2014) Psychological Treatments: A Call for Mental-Health Science. *Nature*, **511**, 287-289. https://doi.org/10.1038/511287a

[11] Kang, Y., Cai, Z., Tan, C., Huang, Q. and Liu, H. (2020) Natural Language Processing (NLP) in Management Research: A Literature Review. *Journal of Management Analytics*, **7**, 139-172. https://doi.org/10.1080/23270012.2020.1756939

[12] Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W. (2011) Natural Language Processing: An Introduction. *Journal of the American Medical Informatics Association*, **18**, 544-551. https://doi.org/10.1136/amiajnl-2011-000464

[13] Grail, Q., Perez, J. and Gaussier, E. (2021) Globalizing BERT-Based Transformer Architectures for Long Document Summarization. *Proceedings of the* 16*th Conference of the European Chapter of the Association for Computational Linguistics*: *Main Volume*, Online, April 2021, 1792-1810. https://doi.org/10.18653/v1/2021.eacl-main.154

[14] Abdal, M.N., Oshie, M.H.K., Haue, M.A. and Islam, K. (2023) A Transformer Based Model for Twitter Sentiment Analysis Using Roberta. 2023 26*th International Conference on Computer and Information Technology* (*ICCIT*), Cox's Bazar, 13-15 December 2023, 1-6. https://doi.org/10.1109/iccit60459.2023.10441627

[15] Nagale, S. (2025) Advances in Transformer Architectures: Integrating BERT, GPT, and Vision Models. GPT, and Vision Models.

[16] Charniak, E. (1996) Statistical Language Learning. MIT Press.

[17] Bzdok, D. (2017) Classical Statistics and Statistical Learning in Imaging Neuroscience. *Frontiers in Neuroscience*, **11**, Article 543. https://doi.org/10.3389/fnins.2017.00543

[18] Manning, C. and Schutze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press.

[19] Rousseeuw, P.J. (2025) Explainable Linear and Generalized Linear Models by the Predictions Plot. *The American Statistician*, **80**, 157-163. https://doi.org/10.1080/00031305.2025.2539235

[20] Mishra, P. (2021) Explainability for Linear Models. In: *Practical Explainable AI Using Python*: *Artificial Intelligence Model Explanations Using Python-Based Libraries*, *Extensions*, *and Frameworks*, Apress, 35-92.

[21] Das, M. and Alphonse, P.J.A. (2023) A Comparative Study on TF-IDF Feature Weighting Method and Its Analysis Using Unstructured Dataset.

[22] Ibrahim, R., Elbagoury, A., Kamel, M.S. and Karray, F. (2018) Tools and Approaches for Topic Detection from Twitter Streams: Survey. *Knowledge and Information Systems*, **54**, 511-539. https://doi.org/10.1007/s10115-017-1081-x

[23] Liu, Q., Wang, J., Zhang, D., Yang, Y. and Wang, N. (2018) Text Features Extraction Based on TF-IDF Associating Semantic. 2018 *IEEE* 4*th International Conference on Computer and Communications* (*ICCC*), Chengdu, 7-10 December 2018, 2238-2243. https://doi.org/10.1109/compcomm.2018.8780663

[24] Browning, M., Carter, C.S., Chatham, C., Den Ouden, H., Gillan, C.M., Baker, J.T., *et al*. (2020) Realizing the Clinical Potential of Computational Psychiatry: Report from the Banbury Center Meeting, February 2019. *Biological Psychiatry*, **88**, e5-e10. https://doi.org/10.1016/j.biopsych.2019.12.026

[25] Chen, C.S. and Vinogradov, S. (2024) Personalized Cognitive Health in Psychiatry: Current State and the Promise of Computational Methods. *Schizophrenia Bulletin*, **50**, 1028-1038. https://doi.org/10.1093/schbul/sbae108

[26] Laufer, O., Israeli, D. and Paz, R. (2016) Behavioral and Neural Mechanisms of Overgeneralization in Anxiety. *Current Biology*, **26**, 713-722. https://doi.org/10.1016/j.cub.2016.01.023

[27] Semin, G.R.,and Smith, E.R. (2002) Interfaces of Social Psychology with Situated and Embodied Cognition. *Cognitive Systems Research*, **3**, 385-396. https://doi.org/10.1016/S1389-0417(02)00049-9

[28] Weeks, J.W. (2010) The Disqualification of Positive Social Outcomes Scale: A Novel Assessment of a Long-Recognized Cognitive Tendency in Social Anxiety Disorder. *Journal of Anxiety Disorders*, **24**, 856-865. https://doi.org/10.1016/j.janxdis.2010.06.008

[29] Kuhn, D. (2007) Jumping to Conclusions. *Scientific American Mind*, **18**, 44-51. https://doi.org/10.1038/scientificamericanmind0207-44

[30] Morse, B.S. and Schwartzwald, D. (1998) Isophote-Based Interpolation. *Proceedings* 1998 *International Conference on Image Processing*. *ICIP*98 *(Cat. No.* 98*CB*36269*)*, Chicago, 7 October 1998, 227-231. https://doi.org/10.1109/ICIP.1998.999013

[31] Ruba, A.L. and Pollak, S.D. (2020) The Development of Emotion Reasoning in Infancy and Early Childhood. *Annual Review of Developmental Psychology*, **2**, 503-531. https://doi.org/10.1146/annurev-devpsych-060320-102556

[32] Ashforth, B.E. and Humphrey, R.H. (1995) Labeling Processes in the Organization. *Research in Organizational Behaviour*, **17**, 413-461.

[33] Fan, H.Y. and Poole, M.S. (2006) What Is Personalization? Perspectives on the Design and Implementation of Personalization in Information Systems. *Journal of Organizational Computing and Electronic Commerce*, **16**, 179-202. https://doi.org/10.1080/10919392.2006.9681199

[34] Konopka, B.M., Lwow, F., Owczarz, M. and Łaczmański, Ł. (2018) Exploratory Data Analysis of a Clinical Study Group: Development of a Procedure for Exploring Mul-

tidimensional Data. *PLOS ONE*, **13**, e0201950.
https://doi.org/10.1371/journal.pone.0201950

[35] Avval, T.G., Moeini, B., Carver, V., Fairley, N., Smith, E.F., Baltrusaitis, J., *et al.* (2021) The Often-Overlooked Power of Summary Statistics in Exploratory Data Analysis: Comparison of Pattern Recognition Entropy (PRE) to Other Summary Statistics and Introduction of Divided Spectrum-Pre (DS-PRE). *Journal of Chemical Information and Modeling*, **61**, 4173-4189. https://doi.org/10.1021/acs.jcim.1c00244

[36] Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D.F., *et al.* (2009) Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Philosophical Transactions of the Royal Society A*: *Mathematical*, *Physical and Engineering Sciences*, **367**, 4361-4383. https://doi.org/10.1098/rsta.2009.0120

[37] Dhariwal, A., Junges, R., Chen, T. and Petersen, F.C. (2021) Resistoxplorer: A Web-Based Tool for Visual, Statistical and Exploratory Data Analysis of Resistome Data. *NAR Genomics and Bioinformatics*, **3**, lqab018.
https://doi.org/10.1093/nargab/lqab018

[38] Batch, A. and Elmqvist, N. (2017) The Interactive Visualization Gap in Initial Exploratory Data Analysis. *IEEE Transactions on Visualization and Computer Graphics*, **24**, 278-287. https://doi.org/10.1109/tvcg.2017.2743990

[39] Gillani, H.H., Qureshi, M.A., Beghdadi, A., Cheikh, F. and Ullah, M. (2025) Distortion Classification in Computer Vision Applications: Current Progress, Challenges, and Perspectives. *ACM Computing Surveys*, **58**, 1-36.
https://doi.org/10.1145/3773023

[40] Formanowicz, M. and Hansen, K. (2021) Subtle Linguistic Cues Affecting Gender In(equality). *Journal of Language and Social Psychology*, **41**, 127-147.
https://doi.org/10.1177/0261927x211035170

[41] van der Auwera, J. and König-Johan, E. (1990) Adverbial Participles, Gerunds and Absolute Constructions in the Languages of Europe. In: *Toward a Typology of European Languages*, De Gruyter Brill, 337.

[42] Tan, C.M., Wang, Y.F. and Lee, C.D. (2002) The Use of Bigrams to Enhance Text Categorization. *Information Processing & Management*, **38**, 529-546.
https://doi.org/10.1016/s0306-4573(01)00045-0

[43] Zhu, X.J., Goldberg, A.B., Rabbat, M. and Nowak, R. (2008) Learning Bigrams from Unigrams. *Proceedings of ACL*-08: *HLT*, Columbus, 10 January 2008, 656-664.

[44] Nikkarinen, I., Pimentel, T., Blasi, D. and Cotterell, R. (2021) Modeling the Unigram Distribution. *Findings of the Association for Computational Linguistics*: *ACL-IJCNLP* 2021, Online, August 2021, 3721-3729.
https://doi.org/10.18653/v1/2021.findings-acl.326

[45] Liu, I., Lo, K. and Wu, J. (1996) A Probabilistic Interpretation of "If-then". *The Quarterly Journal of Experimental Psychology Section A*, **49**, 828-844.
https://doi.org/10.1080/713755646

[46] Cestnik, B. and Bratko, I. (1991) On Estimating Probabilities in Tree Pruning. In: *Lecture Notes in Computer Science*, Springer, 138-150.
https://doi.org/10.1007/bfb0017010

[47] Beygelzimer, A., Langford, J., Lifshits, Y., Sorkin, G. and Strehl, A.L. (2014) Conditional Probability Tree Estimation Analysis and Algorithms. arXiv: 1408.2031.

[48] Mejía, D.A. and Uribe-Zapata, A.F. (2025) Probability Trees. arXiv: 2501.07023.

[49] Rane, N., Choudhary, S.P. and Rane, J. (2024) Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions. *Studies in*

Medical and Health Sciences, **1**, 18-41. https://doi.org/10.48185/smhs.v1i2.1225

[50] Gashler, M., Giraud-Carrier, C. and Martinez, T. (2008) Decision Tree Ensemble: Small Heterogeneous Is Better than Large Homogeneous. 2008 *Seventh International Conference on Machine Learning and Applications*, San Diego, 11-13 December 2008, 900-905. https://doi.org/10.1109/icmla.2008.154

[51] Ghiasi, M.M. and Zendehboudi, S. (2021) Application of Decision Tree-Based Ensemble Learning in the Classification of Breast Cancer. *Computers in Biology and Medicine*, **128**, Article 104089. https://doi.org/10.1016/j.compbiomed.2020.104089

[52] Lian, W., Nie, G., Jia, B., Shi, D., Fan, Q. and Liang, Y. (2020) An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning. *Mathematical Problems in Engineering*, **2020**, 1-15. https://doi.org/10.1155/2020/2835023

[53] Qu, Z.W., Song, X.M., Zheng, S.Q., Wang, X.R., *et al.* (2018) Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification. 2018 *IEEE International Conference on Big Data and Smart Computing* (*BigComp*), Shanghai, 15-17 January 2018, 677-680. https://doi.org/10.1109/bigcomp.2018.00124

[54] Kadhim, A.I. (2019). Term Weighting for Feature Extraction on Twitter: A Comparison between BM25 and TF-IDF. 2019 *International Conference on Advanced Science and Engineering* (*ICOASE*), Zakho-Duhok, 2-4 April 2019, 124-128. https://doi.org/10.1109/icoase.2019.8723825

[55] Zhang, H.P. and Singer, B.H. (2010) Recursive Partitioning and Applications. Springer Science & Business Media.

[56] Strobl, C., Malley, J. and Tutz, G. (2009) An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, **14**, 323-348. https://doi.org/10.1037/a0016973

[57] Koskas, M., Luton, D., Graesslin, O., Barranger, E., Clavel-Chapelon, F., Haddad, B., *et al.* (2015) Direct Comparison of Logistic Regression and Recursive Partitioning to Predict Lymph Node Metastasis in Endometrial Cancer. *International Journal of Gynecological Cancer*, **25**, 1037-1043. https://doi.org/10.1097/igc.0000000000000451

[58] Hawkins, D.M. (2009) Recursive Partitioning. *WIREs Computational Statistics*, **1**, 290-295. https://doi.org/10.1002/wics.44

[59] Salih, A.M. and Wang, Y.H. (2024) Are Linear Regression Models White Box and Interpretable?

[60] Kumarakulasinghe, N.B., Blomberg, T., Liu, J., Saraiva Leao, A. and Papapetrou, P. (2020) Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. 2020 *IEEE* 33*rd International Symposium on Computer-Based Medical Systems* (*CBMS*), Rochester, 28-30 July 2020, 7-12. https://doi.org/10.1109/cbms49503.2020.00009

[61] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K. and Cilar, L. (2020) Interpretability of Machine Learning-Based Prediction Models in Healthcare. *WIREs Data Mining and Knowledge Discovery*, **10**, e1379. https://doi.org/10.1002/widm.1379

[62] Sayyidul Laily, F.T.A. (2024) Feature Extraction and Classification of Retinal Images Using Sobel Segmentation and Linear Svc. *International Journal of Artificial Intelligence in Medical Issues*, **2**, 136-149. https://doi.org/10.56705/ijaimi.v2i2.153

[63] Lichtenstein, S., Fischhoff, B. and Phillips, L.D. (1977) Calibration of Probabilities:

The State of the Art. In: *Decision Making and Change in Human Affairs*, Springer, 275-324. https://doi.org/10.1007/978-94-010-1276-8_19

[64] DeMonbreun, B.G. and Craighead, W.E. (1977) Distortion of Perception and Recall of Positive and Neutral Feedback in Depression. *Cognitive Therapy and Research*, **1**, 311-329. https://doi.org/10.1007/bf01663996

[65] Candel, I., Merckelbach, H. and Zandbergen, M. (2003) Boundary Distortions for Neutral and Emotional Pictures. *Psychonomic Bulletin & Review*, **10**, 691-695. https://doi.org/10.3758/bf03196533

[66] Fan, J., Upadhye, S. and Worster, A. (2006) Understanding Receiver Operating Characteristic (ROC) Curves. *Canadian Journal of Emergency Medicine*, **8**, 19-20. https://doi.org/10.1017/s1481803500013336

[67] Ugi, S., Maegawa, H., Morino, K., Nishio, Y., Sato, T., Okada, S., *et al.* (2016) Evaluation of a Novel Glucose Area under the Curve (AUC) Monitoring System: Comparison with the AUC by Continuous Glucose Monitoring. *Diabetes & Metabolism Journal*, **40**, 326-333. https://doi.org/10.4093/dmj.2016.40.4.326

[68] Couronné, R., Probst, P. and Boulesteix, A. (2018) Random Forest versus Logistic Regression: A Large-Scale Benchmark Experiment. *BMC Bioinformatics*, **19**, Article No. 270. https://doi.org/10.1186/s12859-018-2264-5

[69] Flach, P. and Kull, M. (2015) Precision-Recall-Gain Curves: PR Analysis Done Right. *Advances in Neural Information Processing Systems*, **28**, 1-9.

[70] Davis, J. and Goadrich, M. (2006) The Relationship between Precision-Recall and ROC Curves. *Proceedings of the* 23*rd International Conference on Machine Learning*, Pittsburgh, 25-29 June 2006, 233-240. https://doi.org/10.1145/1143844.1143874

[71] Chicco, D. and Jurman, G. (2020) The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, **21**, Article No. 6. https://doi.org/10.1186/s12864-019-6413-7

[72] Chicco, D., Warrens, M.J. and Jurman, G. (2021) The Matthews Correlation Coefficient (MCC) Is More Informative than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access*, **9**, 78368-78381. https://doi.org/10.1109/access.2021.3084050

[73] Tranmer, M. and Elliot, M. (2008) Binary Logistic Regression. Cathie Marsh for Census and Survey Research Paper 20.

[74] Wilson, J.R. and Lorenz, K.A. (2015) Standard Binary Logistic Regression Model. In: *Modeling Binary Correlated Responses Using SAS, SPSS and R*, Springer, 25-54.

[75] Strickett, M. (2024) Logistic Regression Methods versus Machine Learning Techniques in Status and Severity Prediction of South African Covid-19 Laboratory Test Data. Master's Thesis, University of the Witwatersrand, Johannesburg (South Africa).

[76] Li, L., Rysavy, M.A., Bobashev, G. and Das, A. (2024) Comparing Methods for Risk Prediction of Multicategory Outcomes: Dichotomized Logistic Regression Vs. Multinomial Logit Regression. *BMC Medical Research Methodology*, **24**, Article No. 261. https://doi.org/10.1186/s12874-024-02389-x

[77] Poursheikhali Asgary, M., Jahandideh, S., Abdolmaleki, P. and Kazemnejad, A. (2007) Analysis and Identification of β-Turn Types Using Multinomial Logistic Regression and Artificial Neural Network. *Bioinformatics*, **23**, 3125-3130. https://doi.org/10.1093/bioinformatics/btm324

[78] McCauley, S. (2012) Applying Multinomial Logistic Regression to Categorize Student Technological Knowledge Based on Technology Usage Attributes. Walden Univer-

sity.

[79] Gill, C.J., Sabin, L. and Schmid, C.H. (2005) Why Clinicians Are Natural Bayesians. *British Medical Journal*, **330**, 1080-1083. https://doi.org/10.1136/bmj.330.7499.1080

[80] Hay-Smith, E.J.C., Brown, M., Anderson, L. and Treharne, G.J. (2016) Once a Clinician, Always a Clinician: A Systematic Review to Develop a Typology of Clinician-Researcher Dual-Role Experiences in Health Research with Patient-participants. *BMC Medical Research Methodology*, **16**, Article No. 95. https://doi.org/10.1186/s12874-016-0203-6

[81] Austin, S., Bandealy, A. and Cawley, E. (2024) Technology Meets Clinical Practice: Keel Mind as a Digital Therapy Platform. *Mental Health and Digital Technologies*, **1**, 99-111. https://doi.org/10.1108/mhdt-02-2024-0006

[82] Starke, A.D. and Willemsen, M.C. (2024) Psychologically Informed Design of Energy Recommender Systems: Are Nudges Still Effective in Tailored Choice Environments? In: *Human-Computer Interaction Series*, Springer, 221-259. https://doi.org/10.1007/978-3-031-55109-3_9

[83] Hupp, S.D., Reitman, D. and Jewell, J.D. (2008) Cognitive-Behavioral Theory. In: *Handbook of Clinical Psychology*, Wiley, 263-287.

[84] Benjamin, C.L., Puleo, C.M., Settipani, C.A., Brodman, D.M., Edmunds, J.M., Cummings, C.M., *et al.* (2011) History of Cognitive-Behavioral Therapy in Youth. *Child and Adolescent Psychiatric Clinics of North America*, **20**, 179-189. https://doi.org/10.1016/j.chc.2011.01.011

[85] Huys, Q.J.M., Maia, T.V. and Frank, M.J. (2016) Computational Psychiatry as a Bridge from Neuroscience to Clinical Applications. *Nature Neuroscience*, **19**, 404-413. https://doi.org/10.1038/nn.4238

[86] Ricketts, J., Barry, D., Guo, W. and Pelham, J. (2023) A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports. *Safety*, **9**, 22. https://doi.org/10.3390/safety9020022

[87] Davidson, S., Yamada, A., Mira, P.F., Carando, A., *et al.* (2020) Developing NLP Tools with a New Corpus of Learner Spanish. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, 11-16 May 2020, 7238-7243.