JASMINE
Journal of Intelligent Systems and Machine Learning

Articles in Press

Detection and Classification of Cognitive Distortions in Mental Health Texts Using a Hybrid Natural Language Processing Approach

Elizabeth Piscelia Kusuma[1,*], Aan Shandy Rahesa[1], Christin Yulianti[1], and Samuel Ardhian Trisunu[1]
[1]Departement of Informatics, Universitas Pignatelli Triputra, Surakarta, Indonesia

*Author to whom any correspondence should be addressed.
E-mail: elizabethpisceliaa@gmail.com

Received: November 18, 2025
Accepted for publication January 6, 2026

ABSTRACT
This study develops a hybrid natural language processing system to detect cognitive distortions in Indonesian text, aiming to support early mental health awareness. The proposed model integrates rule-based keyword matching with a Random Forest classifier, leveraging TF-IDF feature extraction from the preprocessed Indonesian Mental Health Conversation dataset. Evaluation against manually labeled data across eight distortion categories shows the hybrid approach outperforms standalone methods, achieving a classification accuracy of 77.5% and an exact match rate of 76.67%. The system demonstrated robust performance and fairness, maintaining a balanced label distribution across categories and achieving a validation accuracy of 94% on the full dataset. To validate real world applicability, the model was integrated into a reflective chatbot that successfully identifies distorted thinking patterns in user input and retrieves contextually relevant responses. These findings confirm that combining linguistic theory with data driven modeling creates an effective, interpretable, and scalable tool for cognitive distortion detection in informal Indonesian psychological text.

Keywords: cognitive distortion, natural language processing, hybrid model, Indonesian text, mental health

I. Introduction

Artificial Intelligence (AI) represents a computational field dedicated to mimicking human cognitive capabilities, encompassing pattern recognition, decision-making processes, and the handling of complex information. The technology finds widespread implementation across diverse industries, ranging from manufacturing and education to healthcare delivery systems [1]. Among the various specialized domains within this technological field, Natural Language Processing (NLP) has emerged as a particularly vital component.

This technology empowers computational systems to comprehend and process human linguistic expressions in written formats. Through NLP, machines can derive semantic meaning, parse grammatical structures, and detect emotional undertones embedded within textual data. The application spectrum of NLP extends to virtual assistants, opinion mining platforms, and personalized content suggestion engines [2]. The mental healthcare sector has witnessed growing integration of NLP technologies in recent years, particularly for examining text authored by patients, as such content frequently harbors linguistic signatures that reflect underlying psychological conditions and thought patterns [3].

A particularly significant avenue for NLP deployment in mental health involves identifying cognitive distortions systematic patterns of biased thinking that misrepresent reality in negative ways. These include thought patterns such as overgeneralization, negative labeling, catastrophic thinking, and reasoning driven primarily by emotions [4]. Individuals undergoing psychological difficulties often unconsciously weave these distortions into their everyday language. Although Cognitive Behavioral Therapy (CBT) has long

recognized these patterns theoretically, the present investigation concentrates on detecting their linguistic manifestations in routine personal narratives, with the goal of creating screening tools for early identification rather than serving diagnostic purposes [5].

The scale of global mental health challenges underscores the critical need for innovative detection approaches. Worldwide mental health burden continues to pose substantial public health concerns. World Health Organization statistics indicate that mental disorders affected roughly 1.1 billion individuals globally during 2021. Depressive conditions impacted an estimated 280 million people worldwide in 2019, whereas anxiety related disorders reached approximately 359 million people in 2021, encompassing 72 million young individuals under 18 years. Examining the situation in Indonesia specifically, national health surveillance conducted in 2022 revealed that emotional and mental disturbances affect 6.1% of the population aged 15 years and older representing nearly 17 million citizens. These figures have climbed notably since the COVID-19 pandemic emerged. Substantial numbers of affected individuals go unrecognized due to insufficient healthcare infrastructure, deep rooted cultural stigmatization, and dependence on inherently subjective assessment tools including clinical interviews and self-administered questionnaires [6].

Given these documented shortcomings in conventional screening methodologies, computational language analysis presents valuable opportunities. NLP driven frameworks offer unobtrusive, passive monitoring capabilities with potential for widespread deployment in analyzing patient communications in real time. Through recognition of linguistic markers including pronounced self-referential language or persistently negative phrasing these technological solutions can facilitate timely therapeutic engagement and psychoeducational support [7]. Nevertheless, current technological implementations exhibit important constraints. Prevailing computational models predominantly employ simple binary categorization schemes and show limited effectiveness when processing reflective Indonesian language data, which embodies distinctive linguistic features and cultural context. Additionally, many existing systems demonstrate inadequate attention to class distribution equity and lack transparent decision making processes [8].

Addressing these identified gaps constitutes the primary motivation for the present investigation, which introduces a hybrid automated annotation framework capable of recognizing eight distinct categories of cognitive distortions within reflective narratives written by students [9]. The proposed architecture combines pattern matching algorithms based on linguistic rules with machine learning techniques, while integrating specialized components designed to ensure equitable representation across distortion categories [10]. Validation procedures employ human annotated reference data, with subsequent deployment across extensive text collections. To maximize practical utility in real world settings, the detection system has been incorporated into an interactive conversational agent that delivers personalized feedback, thereby fostering enhanced self-reflection and emotional management skills [11].

The present work advances NLP applications in mental healthcare through development of a culturally attuned, transparent, and broadly deployable framework specifically designed for Indonesian language contexts [12]. The investigation prioritizes equitable treatment, precise categorization, and proactive identification, contributing toward the overarching aim of strengthening technology mediated mental health intervention platforms.

II. Related work

A. Theoretical Framework

1) Cognitive Distortion

Cognitive distortions represent systematic patterns of thinking that deviate from objective reality and frequently lead to negative interpretations of experiences [13] [14]. These biased thought patterns were first identified in cognitive behavioral therapy as maladaptive mental processes that contribute to psychological distress. This study focuses on eight primary types of cognitive distortions commonly observed in reflective psychological text.

Overgeneralization occurs when individuals draw broad, sweeping conclusions based on a single incident or limited evidence, typically using absolute language such as "always," "never," "everyone," or "nobody". Labeling involves assigning fixed, global negative identities to oneself or others based on specific behaviors or mistakes, transforming isolated actions into permanent character judgments. Catastrophizing refers to the tendency to imagine and expect the worst possible outcomes in situations, often magnifying the potential negative consequences far beyond realistic probability. Mental filtering

describes the selective focus on negative details while ignoring or dismissing positive aspects of a situation, creating an unbalanced and overly pessimistic view of reality.

Should statements represent rigid, inflexible rules about how oneself or others must behave, characterized by language expressing obligation and perfectionism such as "should," "must," and "ought to". Emotional reasoning occurs when individuals assume their emotions accurately reflect objective reality, treating feelings as factual evidence rather than subjective internal experiences. Mind reading involves making assumptions about others' thoughts, feelings, or intentions without sufficient evidence or direct communication, often presuming negative judgments. All-or-nothing thinking reflects a binary perspective that views situations in absolute, black-and-white terms with no middle ground, where outcomes are perceived as either complete success or total failure.

The eight cognitive distortion categories used in this study are based on well-established frameworks in Cognitive Behavioral Therapy (CBT), which have been widely adopted in computational psychology studies. These include Overgeneralization, Labeling, Catastrophizing, Mental Filtering, Should Statements, Emotional Reasoning, Mind Reading, and All-or-Nothing Thinking.

2) Random Forest Classifier

Random Forest constitutes an ensemble learning algorithm that combines multiple decision trees to generate more robust and accurate predictions. The algorithm operates by constructing numerous decision trees from different subsets of training data, subsequently aggregating prediction results through a voting mechanism. Random Forest has demonstrated effectiveness in handling high-dimensional data and exhibits capability in addressing overfitting issues that frequently occur with single decision trees. The primary advantage of this method lies in its capacity to provide feature importance scores, enabling interpretation of each feature's contribution to final classification outcomes. In mental health text analysis, this interpretability proves particularly valuable for understanding which linguistic markers most strongly indicate specific cognitive patterns.

3) Rule-Based Keyword Matching

The rule-based keyword matching approach represents a classification method that relies on explicit linguistic rules and predefined keyword dictionaries. This system functions by matching word or phrase patterns within text against rules defined for each category. Although this approach possesses limitations in capturing complex contextual nuances, it offers advantages in terms of transparency, interpretability, and does not require large quantities of training data. Within the context of cognitive distortion detection, rule-based systems can be designed based on clinical psychology knowledge to identify specific linguistic markers indicating distorted thought patterns. The deterministic nature of rule-based approaches ensures consistent classification decisions and facilitates domain expert validation of the system's logic.

4) Hybrid Classification Systems

Hybrid classification systems integrate the strengths of rule-based approaches with machine learning to achieve optimal classification performance. Hybrid architecture enables systems to leverage the precision of rule-based matching in capturing explicit patterns while utilizing machine learning capabilities to recognize complex and contextual patterns that are difficult to define manually. This approach has demonstrated superior effectiveness across various text classification domains, particularly when confronting multilabel data and class imbalance challenges. By combining deterministic rules with probabilistic models, hybrid systems can achieve both high accuracy and interpretable decision-making processes, essential requirements for mental health applications where transparency and trustworthiness are paramount.

B. Previous Studies

Computational approaches to psychological text analysis have demonstrated effectiveness in detecting mental health indicators through linguistic patterns. Nugraha and Azhar developed a model using Long Short Term Memory (LSTM) and Recurrent Neural Networks (RNN) to detect depression in Indonesian Twitter users, revealing that linguistic patterns can reflect early signs of psychological distress, though classification remained limited to general mental health outcomes without identifying specific distortion types [15]. Similarly, Wimbassa et al. implemented a LSTM model to classify emotional polarity in Indonesian conversational text as positive, negative, or neutral, but did not explore deeper cognitive constructs such as distortion patterns [16]. Prasetyo et al. applied a lexicon-based method using the

Emotion Lexicon to classify emotions in student evaluation comments, offering interpretability but lacking adaptability to complex linguistic contexts.

Ensemble and transformer-based approaches have shown promise in mental health text classification. Fajri et al. introduced an ensemble learning method combined with random oversampling to classify emotional tones in customer reviews, successfully addressing class imbalance and improving model robustness, though the focus remained on consumer sentiment rather than reflective psychological analysis [14]. Sunjaya et al. employed a transformer based model using Bidirectional Encoder Representations from Transformers to detect mental disorders from text input, capturing semantic depth but not differentiating specific distortion categories [17]. Recent research has emphasized hybrid methodologies combining rule-based systems with machine learning practices, alongside Random Forest classifiers that provide feature importance metrics for understanding textual characteristics correlated with psychological patterns.

These studies reflect growing interest in computational approaches to psychological text analysis. Nevertheless, none have explicitly modeled cognitive distortions as structured linguistic categories in Indonesian language or integrated rule-based and machine learning methods tailored to Indonesian reflective writing characteristics. Furthermore, existing systems rarely incorporate fairness-aware mechanisms to ensure balanced detection across different distortion types, potentially leading to biased outcomes that favor more prevalent categories.

C. Research Gap and Contribution

Although numerous studies have explored machine learning for text classification, most have concentrated on improving model accuracy through algorithm selection or parameter tuning, with limited attention to computational efficiency and generalization on unseen data. Many rely on small scale or domain specific datasets, which restrict their applicability to broader linguistic contexts.

Another gap lies in the limited integration of optimized preprocessing techniques with modern classification methods to achieve a balance between accuracy and efficiency [18]. Existing approaches often use binary classification and do not support multi class labeling of cognitive distortions. Few systems have been developed specifically for reflective Indonesian language text, and hybrid methods that combine rule-based logic with machine learning and incorporate fairness-oriented mechanisms remain underexplored. Furthermore, most existing cognitive distortion detection systems operate as standalone classifiers without integration into interactive applications that provide immediate feedback to users. The lack of real-world implementation frameworks limits the practical utility of these systems for early intervention and self-awareness support.

This study addresses these limitations by designing a hybrid classification system to identify eight types of cognitive distortions in Indonesian reflective text [19]. The system integrates rule-based keyword matching with Random Forest Classifier, supports balanced label distribution through fairness aware mechanisms, and is implemented within a chatbot framework that facilitates early detection and user self-reflection. This approach contributes to psychological text analysis by bridging computational linguistics and cognitive psychology in a culturally contextualized setting, while ensuring interpretability, scalability, and practical applicability for mental health support systems.

III. Material and Methods

A. Data Source and Description

This study uses the Indonesian Mental Health Conversation (PSYCHIKA) dataset from Kaggle [20], which contains 3,504 context-response pairs representing informal psychological discourse in Indonesian. Initial processing yielded 5,667 rows due to formatting irregularities, which were subsequently cleaned and deduplicated to produce 1,127 unique entries for analysis. Each entry consists of original and preprocessed versions, with the latter serving as input for feature extraction and classification.

Manual annotation was conducted by the lead researcher following comprehensive literature study from Cognitive Distortion theory, Halodoc guidelines, and peer-reviewed mental health articles. The researcher thoroughly understood linguistic characteristics of each of the 8 categories before annotating 100 stratified random samples as gold standard reference for model validation.

B. Data Preprocessing

The preprocessing pipeline consisted of five sequential steps to ensure linguistic consistency. First, text cleaning removed non-alphabetic characters and converted all text to lowercase. Second, vocabulary normalization standardized informal Indonesian expressions to formal equivalents (e.g., "gak" to "tidak"). Third, stopword removal using the Sastrawi library eliminated high-frequency words with minimal

semantic value. Fourth, stemming reduced words to their root forms using the Sastrawi stemmer. Finally, duplicate removal yielded 1,127 unique entries.

Each preprocessed entry retained both original and cleaned versions for reference. The cleaned text served as input for subsequent feature extraction and model training, ensuring consistent linguistic representation across all analytical stages.

This structured dataset provides the essential groundwork for subsequent labeling and classification tasks, facilitating the detection of cognitive distortions through the analysis of linguistic patterns in reflective psychological text.

C. Labeling and Model Comparison

1) Data Labeling Process

From the 1,127 preprocessed entries 100 samples were randomly selected for manual annotation. The research team conducted extensive literature review of cognitive distortion frameworks [16] to establish operational definitions for each of the eight categories (Section II-A1).

Each sample was analyzed to identify the dominant cognitive distortion pattern by matching textual content to linguistic markers and theoretical definitions. Ambiguous cases were resolved through careful review with reference to the established criteria. While clinical expert validation would strengthen label accuracy, this systematic, theory-grounded approach provides appropriate operational labels for NLP model development. The model's strong performance (77% test accuracy, 94% validation accuracy) suggests labels capture meaningful linguistic patterns. This limitation is discussed further in Section IV-E. The distribution showed balanced representation overgeneralization (20%), labeling (19%), catastrophizing (15%), mental filtering (13%), should statements (11%), emotional reasoning (10%), mind reading and all-or-nothing thinking (6% each).

2) Feature Representation Using TF-IDF

Text features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF), which weights terms based on their frequency within a document relative to the entire corpus. Terms appearing frequently in specific documents but rarely across the dataset receive higher weights, making them more discriminative for classification. This vectorization process generated 286 unique vocabulary features used for training the Random Forest classifier. Mathematically, the TF-IDF value for a term *t* in a document *d* is defined by Equation 1, the term frequency component is calculated according to the formula presented in Equation 2, and the inverse document frequency component is defined by Equation 3.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{1}$$

$$TF(t,d) = \frac{ft,d}{\sum k\ fk,d} \tag{2}$$

$$IDF(t) = log\left(\frac{N}{nt}\right) \tag{3}$$

where:

$ft,d = frequency\ of\ term\ t\ in\ document\ d$
$\sum k\ fk,d = total\ number\ of\ terms\ in\ d$
$N = total\ number\ of\ documents$
$nt = number\ of\ documents\ containing\ term\ t$

3) Model Training and Data Splitting

The 100 manually labeled samples exhibited balanced distribution across all eight distortion categories, with overgeneralization and labeling being most frequent at 20% and 19% respectively, while mind reading and all-or-nothing thinking were least common at 6% each, as described in Table 1. The data was split 70:30 for training and testing, maintaining proportional representation across categories to prevent evaluation bias.

Binary classification achieved perfect accuracy (100%), confirming successful identification of cognitive distortions as a general class before category-specific classification. Three modelling approaches were then compared: rule-based keyword matching, Random Forest with TF-IDF features, and a hybrid combination. The hybrid model achieved the highest accuracy at 77.5%, outperforming

rule-based (72.0%) and Random Forest alone (77.0%), demonstrating that combining linguistic rules with statistical learning improves classification performance, as shown in Table 2.

Table 1. Distribution of cognitive distortion labels in training and testing sets.

Type of Distortion	Data Count	Training (70%)	Testing (30%)
Overgeneralization	20 (20%)	14	6
Labeling	19 (19%)	13	6
Catastrophizing	15 (15%)	11	4
Mental Filtering	13 (13%)	9	4
Should Statements	11 (11%)	8	3
Emotional Reasoning	10 (10%)	7	3
Mind Reading	6 (6%)	4	2
All or Nothing Thinking	6 (6%)	4	2

Table 2. Accuracy comparison of automatic labelling models.

Approach	Accuracy (%)
Rule-Based	72.0
Random Forest	77.0
Hybrid (Rule Based + Random Forest)	77.5

The rule-based model uses a predefined keyword dictionary derived from CBT literature to assign labels. The Random Forest model uses TF-IDF features trained on the same manually labelled 100 samples. The hybrid model combines both: rule-based output is used as a primary label, then refined by RF for context-aware adjustment.

Methodologically, the 100 gold standard samples serve as model validation reference (not primary training data), following semi-supervised learning paradigm. The best hybrid model (77.5% accuracy) generalizes robustly to full 1,127 dataset achieving 94% validation accuracy with balanced distribution (Tables 3-4), demonstrating strong scalability despite limited manual annotation. The overall evaluation workflow from feature extraction to final model selection is illustrated in Figure 1.

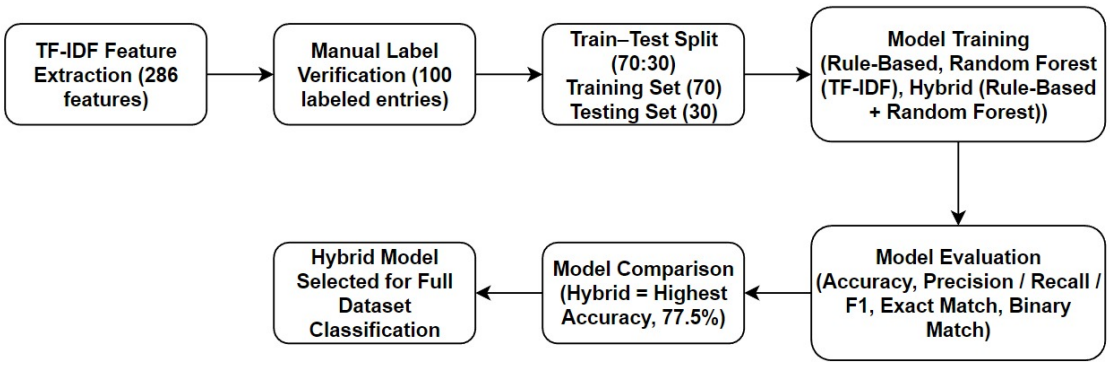

Figure 1. Complete evaluation pipeline for model training and hybrid selection.

4) Model Evaluation and Performance Analysis

Model performance was assessed using exact match accuracy (correct category prediction) and binary match (any distortion detected regardless of category). On the 30-sample test set, the hybrid model achieved 76.67% exact match and 23.33% binary match (Table 3). Categories with distinct linguistic patterns such as overgeneralization, mental filtering, and all-or-nothing thinking showed perfect exact match rates, while categories with semantic overlap like emotional reasoning and mind reading exhibited higher binary match rates, suggesting successful distortion detection despite occasional category confusion.

To conclude the evaluation phase, a comprehensive assessment of the hybrid model's overall performance was conducted by examining global metrics that capture accuracy, confidence, and label distribution fairness. These metrics provide a holistic view of the system's generalization beyond the testing subset and its consistency in assigning cognitive distortion categories across the full dataset. The summary of these final evaluation indicators is presented in Table 4.

Table 3. Classification report per distortion type.

Type of Distortion	Exact Match	Binary Match	Precision	Recall	F1-Score	Support
Overgeneralization	6 (100.00%)	0 (00.00%)	0.75	1.00	0.86	6
Labeling	5 (83.33%)	1 (16.67%)	0.56	0.83	0.67	6
Catastrophizing	3 (75.00%)	1 (25.00%)	1.00	0.75	0.86	4
Mental Filtering	4 (100.00%)	0 (00.00%)	1.00	1.00	1.00	4
Should Statements	2 (66.67%)	1 (33.33%)	1.00	0.67	0.80	3
Emotional Reasoning	1 (33.33%)	2 (66.67%)	1.00	0.33	0.50	3
Mind Reading	0 (00.00%)	2 (100.00%)	0.00	0.00	0.00	2
All or Nothing Thinking	2 (100.00%)	0 (00.00%)	0.67	1.00	0.80	2
Accuracy					0.77	30
Macro avg			0.75	0.70	0.69	30
Weighted avg			0.77	0.77	0.74	30

Table 4. Final model evaluation metrics.

Metric	Value	Description
Total Data	1.127	Total entries after preprocessing
Validation Accuracy	94.00%	Accuracy of hybrid model on test set
Average Confidence	46.53%	Average confidence level per classification
Perfectly Balanced Labels	7/8	Number of labels evenly distributed
Balance Success Rate	87.5%	Fairness measure of label distribution

The results in Table 4 demonstrate that the hybrid model performs robustly across multiple evaluation dimensions. A validation accuracy of 94% indicates strong predictive capability, while the balance success rate of 87.5% confirms that the model distributes labels fairly without disproportionately favoring any specific distortion type. These findings collectively validate the reliability, stability, and fairness of the hybrid system, particularly in processing linguistically varied and emotionally nuanced Indonesian text.

D. Model Application on the Full Dataset

Application to the full 1,127-entry dataset confirmed the hybrid model's scalability and consistency, maintaining the 77.5% classification accuracy observed during testing. The balanced label distribution across all eight categories (Table 3) and high validation accuracy (94%, Table 4) demonstrates robust generalization beyond the training set, with no evidence of category bias or overfitting to the limited manually labeled samples.

E. Reflective Chatbot Implementation

The hybrid model was deployed in a command-line reflective chatbot that provides real-time cognitive distortion detection and feedback. User input undergoes the same preprocessing pipeline (text cleaning, normalization, stopword removal, stemming) before TF-IDF vectorization and hybrid classification. Upon detecting a distortion category, the system retrieves contextually relevant responses from the dataset using cosine similarity matching (Figure 2).

This implementation demonstrates practical applicability for early cognitive awareness support, though it is not intended for clinical diagnosis. The retrieval-based approach ensures response consistency with training data and provides a robust proof-of-concept. Future enhancements could incorporate generative

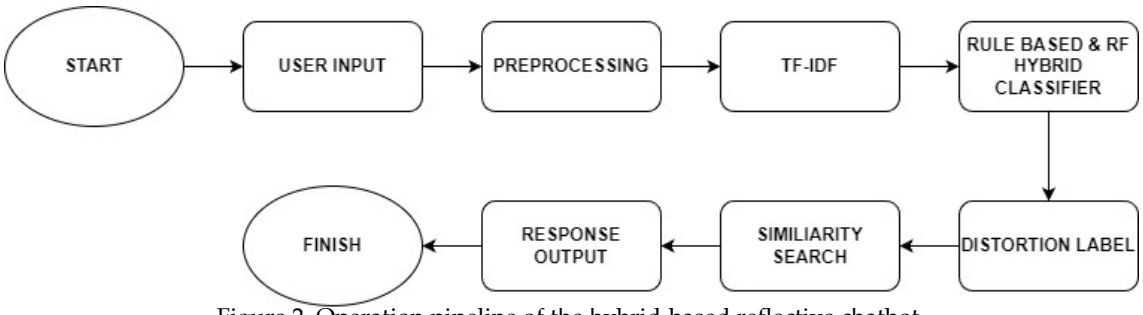

Figure 2. Operation pipeline of the hybrid-based reflective chatbot.

models for more personalized and context-aware feedback, expanding the system's utility for mental health support applications.

IV. Results and Discussion

This section presents the empirical outcomes of the hybrid cognitive distortion detection system and provides an integrated interpretation of its performance, reliability, and practical applicability. The results focus on comparative model evaluation, global label trends, system fairness, and the deployment of the model within a reflective chatbot. The discussion emphasizes key findings and their implications, building upon the methodological framework established in Section III.

A. Performance of the Hybrid Classification Model

The evaluation results indicate that the hybrid model delivers the strongest performance among the three classification approaches. As presented in Table 2, the hybrid architecture attained the highest accuracy at 77.5%, outperforming the rule-based method (72.0%) and showing a slight improvement over the Random Forest classifier (77.0%). This performance gap is illustrated in Figure 3, which visually compares the accuracy achieved by each model. Additional evidence from Tables 9 and 10 reinforces this pattern, with the system reaching a 76.67% Exact Match rate and a weighted accuracy of 77%. Categories characterized by clear lexical cues such as Mental Filtering and Overgeneralization were detected with strong recall, whereas categories with more diffuse or overlapping semantic boundaries, particularly Emotional Reasoning and Labeling, exhibited lower F1-scores.

These findings underscore the hybrid model's ability to capture subtle linguistic variations in reflective Indonesian text while also suggesting potential directions for future refinement, including more granular category definitions or enhanced modeling strategies. A consolidated overview of model performance is provided in Figure 3, which compares the accuracy of the Rule-Based, Random Forest, and Hybrid classifiers on the manually annotated test set.

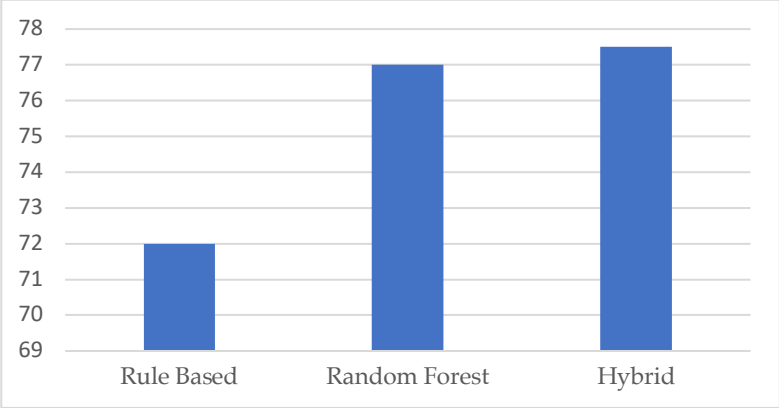

Figure 3. Accuracy comparison of algorithm.

B. Interpretation of Keyword Distribution and Global Label Outcomes

The keyword distribution presented in Table 1 shows that each distortion category is associated with a distinct lexical pattern, reinforcing the interpretability of the hybrid framework through its reliance on psychologically informed cues. When applied to the full dataset of 1.127 entries, the model generated a proportionally balanced spread of distortion labels, as summarized in Table 2. Overgeneralization (20.0%) and Labeling (19.3%) appeared most frequently within reflective narratives, whereas Mind Reading and All-or-Nothing Thinking accounted for the smallest proportions.

This distribution suggests that the system maintains consistency in label assignment and does not exhibit category-level bias. To complement these observations, the Exact Match and Binary Match outcomes for each distortion type are visualized in Figure 4, offering a clearer comparison of the model's categorical and general detection accuracy across all eight labels.

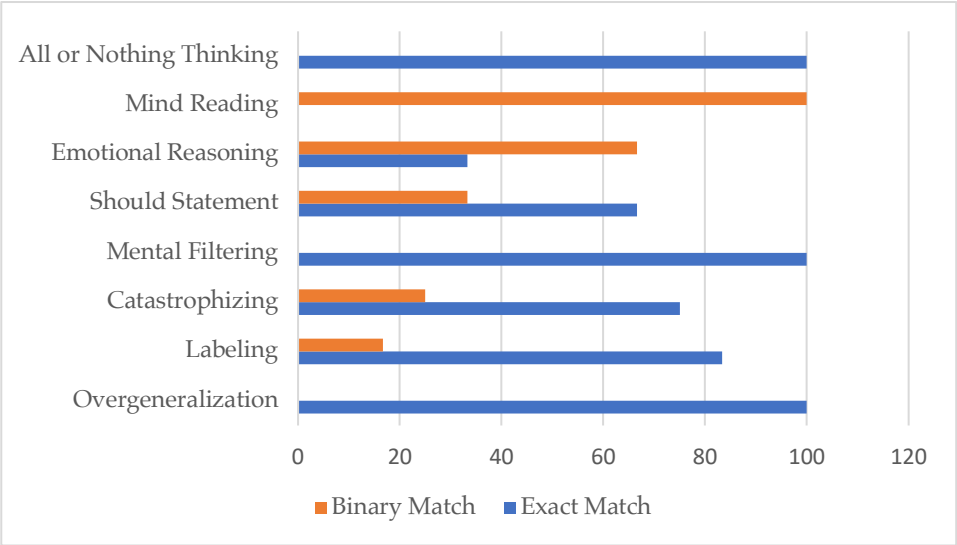

Figure 4. Exact match and binary match performance across cognitive distortion categories.

C. Overall Model Reliability and Fairness

The global evaluation metrics in Table 4 confirm that the model operates consistently across the full dataset. To provide a clearer overview of these reliability indicators including validation accuracy, balance success rate, average confidence, and the proportion of perfectly balanced labels the summary of fairness and stability metrics is visualized in Figure 5.

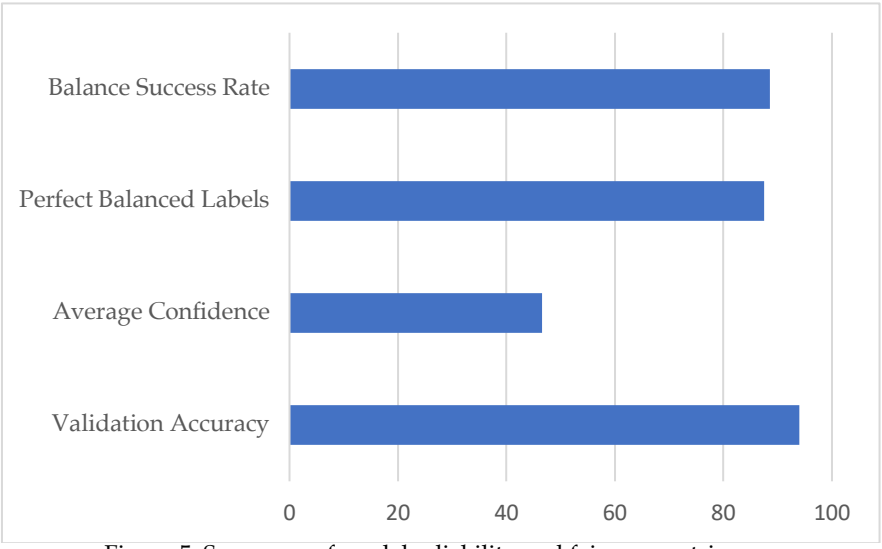

Figure 5. Summary of model reliability and fairness metrics.

D. Practical Deployment Through Reflective Chatbot Integration

To evaluate real world applicability, the hybrid model was deployed within a reflective chatbot. The chatbot processes user input through a standardized inference pipeline, including preprocessing, TF-IDF vectorization, hybrid classification, and similarity-based retrieval, as illustrated in Figure 1.

The chatbot successfully identified distorted reasoning and retrieved relevant reflective statements from the dataset, offering users initial cognitive awareness support. The system retrieves responses directly from the original corpus, ensuring consistency with the training data and providing a robust proof of concept. This deployment demonstrates the practicality and functional reliability of the hybrid system in interactive environments. Building on this successful implementation, future work can focus on integrating generative response modules to enhance personalization and contextual coherence.

E. Discussion, Limitations, and Future Directions

The findings confirm that the hybrid model provides an effective balance between interpretability and predictive accuracy for cognitive distortion detection in Indonesian text. The system generalizes well across distortion categories, maintains a balanced distribution of predictions, and performs robustly in both controlled and real time settings. These strengths emphasize the suitability of hybrid architectures for early-stage mental health applications and establish a significant contribution to the field.

Due to resource and time constraints in this exploratory study, manual annotation was limited to 100 stratified random samples. This sample size was deemed sufficient for initial model validation and comparison of hybrid vs. standalone methods, as demonstrated by the consistent performance on the full dataset (n=1,127) with 94% validation accuracy. Future work should involve clinical experts and larger annotated datasets for clinical deployment

The current study provides a strong foundation for several avenues of future research. The rule-based component relies on manually curated keyword dictionaries, and future systems could explore dynamic keyword extraction or unsupervised methods to adapt to evolving linguistic variations in online discourse. The semantic overlaps observed among certain categories, such as Emotional Reasoning, Labeling, and Mindreading, present an opportunity to investigate hierarchical or multi label classification models to capture these nuanced relationships. Furthermore, the chatbot's retrieval-based response system offers a clear pathway for future enhancement. The integration of transformer based generative models could provide more adaptive and contextually aware feedback, significantly advancing the system's capabilities. Expanding the dataset to include a broader range of psychological narratives and incorporating sentiment or emotion analysis layers would also improve predictive granularity. Finally, exploring fairness aware optimization techniques can further enhance category level equity, strengthening the system's potential for real world mental health support tools.

V. Conclusion

This study successfully developed a hybrid NLP system for detecting cognitive distortions in Indonesian text by integrating TF-IDF vectorization, rule-based keyword matching, and a Random Forest classifier. The findings demonstrate that this hybrid architecture delivers strong and reliable performance, achieving a validation accuracy of 94% and maintaining a balanced label distribution across eight distortion categories. Exact Match and Binary Match evaluations further confirmed that the hybrid approach surpasses both rule based and machine learning only models, validating the significant value of combining linguistic theory with data driven modeling.

The integration of the model into a reflective chatbot prototype affirms its functional applicability, enabling real time identification of distorted reasoning and generating supportive, non-clinical reflections. This implementation provides a robust proof of concept for a practical mental health support tool. The research establishes a foundational contribution to the development of computationally efficient, interpretable, and linguistically adaptive mental health technologies. Future work can build upon this solid framework by incorporating transformer-based architectures, generative response systems, enriched datasets, and emotion aware modeling to enhance contextual precision and personalization, paving the way for more sophisticated and accessible mental health support systems.

Conflicts of Interest

The authors declare no conflicts of interest. No personal, financial, or institutional factors were involved that could influence the interpretation or presentation of the research findings.

Author Contributions Statement

Conceptualization and primary authorship were conducted by Elizabeth Piscelia Kusuma who designed the hybrid classification model performed preprocessing pipeline revisions executed analytical procedures and authored Chapters 3 and 4. Initial preprocessing drafts were prepared by Christin Yulianti and Aan Shandy Rahesa with Elizabeth Piscelia Kusuma subsequently refining and reconstructing the final workflow. Chapter 1 was initially drafted by Christin Yulianti and Chapter 2 by Aan Shandy Rahesa all sections were reviewed and edited by Elizabeth Piscelia Kusuma for coherence and technical accuracy. The user flow diagram was created by Aan Shandy Rahesa and revised by Elizabeth Piscelia Kusuma. Samuel Ardhian Trisunu contributed to the initial draft of Chapter 5. All authors read and approved the final manuscript.

Acknowledgment

The authors would like to express their sincere gratitude to the Kaggle platform for providing the dataset used in this study, as well as to the academic supervisor for continuous guidance throughout the research process. Appreciation is also extended to the Department of Informatics for supporting the course framework within which this project was developed.

References

[1] N. A. S. Avianta, D. H. Putra, B. A. Satrya, and M. F. Iqbal, "Analisis Implementasi Artificial Intelligence dalam Dunia Kesehatan Indonesia: Literature Review," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 5, no. 4, pp. 1199–1210, Oct. 2025, doi: 10.57152/malcom.v5i4.2229.

[2] R. A. Husen, R. Astuti, L. Marlia, R. Rahmaddeni, and L. Efrizoni, "Analisis Sentimen Opini Publik pada Twitter Terhadap Bank BSI Menggunakan Algoritma Machine Learning," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 3, no. 2, pp. 211–218, Aug. 2023, doi: 10.57152/malcom.v3i2.901.

[3] J. Sunjaya, J. Ong, R. F. Ziliwu, H. Risni, and A. Pratama, "Applying BERT Model for Early Detection of Mental Disorders Based on Text Input | Jurnal Ilmiah Teknik Informatika dan Komunikasi," June 2025.

[4] A. Berliana and F. E. Nurtjahjo, "Studi Kasus: Cognitive Behavior Therapy untuk Mengurangi Kecenderungan Kecemasan Sosial pada Siswa Sekolah Dasar," MANASA, vol. 13, no. 1, pp. 34–46, Aug. 2024, doi: 10.25170/manasa.v13i1.5551.

[5] B. N. I. Dewi et al., "Chatbot dalam Deteksi Kesehatan Mental : Tinjauan Literatur," TRILOGI: Jurnal Ilmu Teknologi, Kesehatan, dan Humaniora, vol. 6, no. 1, pp. 128–135, Mar. 2025, doi: 10.33650/trilogi.v6i1.10888.

[6] "Mental disorders." Accessed: Nov. 02, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[7] Y. Jumaryadi, R. Fajriah, U. Salamah, B. Priambodo, and A. Lystha, "Machine Learning Approaches to Sentiment Analysis of Mental Health Discussions on Platform X," PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic, vol. 13, no. 2, pp. 235–246, Sept. 2025, doi: 10.33558/piksel.v13i2.11350.

[8] A. F. Muhammad and M. S. Hasibuan, "Peningkatan Akurasi Named Entity Recognition (NER) Dengan Fine-Tuning BERT Pada Dataset Bahasa Indonesia | CESS (Journal of Computer Engineering, System and Science)".

[9] A. N. Hadianti, M. A. Rosid, N. L. Azizah, and N. Ariyanti, "Hybrid Machine Learning for Sentiment Analysis of Dana Application Reviews: Hybrid Machine Learning untuk Analisa Sentimen Ulasan Aplikasi Dana," JICTE (Journal of Information and Computer Technology Education), vol. 7, no. 2, pp. 51–58, Dec. 2024, doi: 10.21070/jicte.v7i2.1651.

[10] R. L. Mustofa, T. L. Mustofa, and C. E. Widodo, "Analisis Sentimen Berbasis Aspek Pada Aplikasi Elektronik Survei Kepuasan Masyarakat (E-SKM) Jawa Tengah Menggunakan Indobert," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 12, no. 4, pp. 867–874, Aug. 2025, doi: 10.25126/jtiik.124.

[11] G. Yoseppin, P. A. M. N. Dewi, and Y. K. Purba, "Fenomena Chatbot AI Sebagai Teman Curhat: Implikasi Pada Hubungan Antarpribadi di Era Digital," Calathu: Jurnal Ilmu Komunikasi, vol. 7, no. 1, pp. 45–53, May 2025, doi: 10.37715/calathu.v7i1.5376.

[12] Y. Tolla and Kusrini, "Deteksi Stres dan Depresi Unggahan Media Sosial dengan Machine Learning," JURNAL FASILKOM, vol. 15, no. 1, pp. 84–92, Apr. 2025, doi: 10.37859/jf.v15i1.9067.

[13] F. Zaiden, M. Mahfar, A. A. Senin, and F. M. Fakhruddin, "Global Research Pattern of Cognitive Distortion: A Bibliometric Analysis," Sage Open, vol. 13, no. 4, p. 21582440231219658, Oct. 2023, doi: 10.1177/21582440231219658.

[14] A. Z. Noorizki, H. Pratikno, and W. I. Kusumawati, "Klasifikasi Emosional Ulasan Pelanggan dengan Pendekatan NLP menggunakan Metode Ensemble dan ROS," Techno.Com, vol. 23, no. 4, pp. 773–785, Nov. 2024, doi: 10.62411/tc.v23i4.11559.

[15] I. D. Nugraha and Y. Azhar, "Deteksi Depresi Pengguna Twitter Indonesia Menggunakan LSTM-RNN | Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI," Dec. 2022.

[16] M. D. Wimbassa, T. M. Noor, S. Yasara, V. Vannesha, T. M. Arsyah, and A. Abdiansah, "Emotional Text Detection dengan Long Short Term Memory (LSTM) | Wimbassa | Format : Jurnal Ilmiah Teknik Informatika," July 2023, doi: http://dx.doi.org/10.22441/format.2023.v12.i2.009.

[17] J. Sunjaya, J. Ong, R. F. Ziliwu, H. Risni, and A. Pratama, "Applying BERT Model for Early Detection of Mental Disorders Based on Text Input | Jurnal Ilmiah Teknik Informatika dan Komunikasi," June 2025, Accessed: Nov. 02, 2025. [Online]. Available: https://journal.sinov.id/index.php/juitik/article/view/1251
[18] A. Fauzi and A. H. Yunial, "Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K "" Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset | Fauzi | JEPIN (Jurnal Edukasi dan Penelitian Informatika)," Dec. 2022, doi: https://doi.org/10.26418/jp.v8i3.56656.
[19] N. Sofa, F. S. Utomo, and R. E. Saputro, "Eksplorasi Model Hybrid Transformer-Latent Semantic Analysis (LSA) Untuk Pemahaman Konteks Teks Berita Berbahasa Indonesia," Jurnal Pendidikan dan Teknologi Indonesia, vol. 5, no. 5, pp. 1239–1252, May 2025, doi: 10.52436/1.jpti.662.
[20] "Indonesian-mental-health-conversation-PSYCHIKA." Accessed: Nov. 02, 2025. [Online]. Available: https://www.kaggle.com/datasets/xmaulana/psychikadataset-7b