# Development and Evaluation of a Chatbot-Based System for Early Detection of Depression Indicators

Min Yang *, Makoto Oka , Hirohiko Mori *

Posted Date: 9 April 2026

doi: 10.20944/preprints202604.0668.v1

Keywords: early intervention; natural language processing; conversational support technology; linguistic feature extraction; sentiment analysis.

 doi:10.20944/preprints202604.0668.v1



*Article*

# Development and Evaluation of a Chatbot-Based System for Early Detection of Depression Indicators

**Min Yang *, Makoto Oka and Hirohiko Mori ***

Tokyo City University

* Correspondence: g2291404@tcu.ac.jp (M.Y.); hmori@tcu.ac.jp (H.M.)

## Highlights

**What are the main findings?**

- A chatbot-based system can effectively detect early signs of depression using linguistic features extracted from natural conversations.
- The proposed hybrid model with session-level aggregation improves detection stability and achieves reliable classification performance.

**What are the implications of the main findings?**

- The system enables non-intrusive and continuous mental health monitoring without relying on traditional clinical questionnaires.
- This approach provides a scalable framework for early intervention and supports future applications in digital mental health care.

## Abstract

In this study, we developed a chatbot-based system for detecting early signs of depression and verified its effectiveness through experimental evaluations and user surveys. Emphasizing that it does not rely on medical checklists, the system is designed to automatically extract three linguistic features associated with depression—frequent use of first-person pronouns, pessimistic expressions, and obsessive-compulsive writing styles—from natural user conversations. Multiple models were constructed for these features, and an ensemble layer integrates their outputs for a comprehensive judgment. The implemented system analyzes input sentences obtained through chat, extracts the three categories of features, calculates a final score through an ensemble layer, and visualizes potential signs of depression based on the total score. We conducted performed an evaluation experiment with 20 participants. In the test data evaluation, the system demonstrated over 76% accuracy in each of the three classification categories: first-person usage, pessimistic tendency, and obsessive-compulsive tendency.

**Keywords:** early intervention; natural language processing; conversational support technology; linguistic feature extraction; sentiment analysis

## 1. Introduction

### *1.1. Social Background*

Contemporary urban life, characterized by rapid change and intensifying competition, has led to a substantial increase in social pressure. Many individuals, often without conscious awareness, impose excessive demands on themselves, creating conditions in which psychological stress can easily accumulate. In particular, groups such as white-collar employees in corporate or organizational environments, as well as students facing academic advancement or job-seeking transitions, have been identified as being more vulnerable to the onset of depressive conditions [1].

 doi:10.20944/preprints202604.0668.v1



In recent years, depression has come to be widely recognized as a major societal issue in modern contexts [2]. The interaction between external demands and heightened internal expectations can disrupt psychological equilibrium, resulting in the gradual buildup of stress. Such circumstances are thought to contribute to the activation of underlying depressive tendencies, thereby elevating the likelihood of mental health disorders.

Early management of negative emotional states, including feelings of discomfort and anxiety, is essential for mitigating risk. When individuals seek professional support at an appropriate stage, it becomes possible to prevent the worsening of symptoms and facilitate recovery. Conversely, failure to address these conditions may lead to serious outcomes, including self-injurious behavior and suicide.

### *1.2. Current Status and Challenges of Depression*

According to the Patient Survey conducted every three years by the Ministry of Health, Labour and Welfare in Japan, the number of individuals diagnosed with mood disorders, including depression, has shown a marked increase over time. Specifically, the total number of patients rose from approximately 433,000 in 1996 to over 1.04 million in 2008, representing an increase of about 2.4 times within a 12-year period. In addition, the annual number of suicides exceeded 30,000 in 1998, after previously remaining in the low 20,000 range, and has since remained at a persistently high level for an extended period [3].

These statistics indicate that a substantial number of individuals may be experiencing depressive symptoms or psychological distress without receiving adequate support. In particular, systems for early detection and prevention of mental health disorders remain insufficient from both social and medical perspectives.

Although effective treatment and care for mental disorders require the involvement of trained professionals such as psychiatrists and counselors, there is a significant shortage of such specialists relative to the working population. Under these circumstances, it is essential to establish accessible approaches that enable individuals to manage stress and monitor their mental health in daily life. Such approaches can complement existing clinical services and contribute to reducing the burden on healthcare systems [4,5].

### *1.3. Research Objectives*

The objective of this study is to develop a system capable of detecting early signs of depression from everyday interactions with a chatbot and to examine its social acceptability among users.

In conventional clinical practice, depression is typically assessed using standardized medical checklists. However, applying such instruments in daily life may impose a psychological burden on individuals and create barriers to continuous use. To address this limitation, the proposed system does not rely on traditional diagnostic questionnaires. Instead, it focuses on identifying depressive tendencies based on linguistic patterns observed in casual conversations with a chatbot. By creating an interactive environment that provides users with a sense of being engaged and attended to, the system is expected to reduce psychological resistance, encourage sustained interaction, and facilitate the collection of higher-quality conversational data.

In this study, a depression early detection system (hereafter referred to as “the proposed system”) is implemented on WeChat, one of the most widely used social networking platforms. The final assessment is not determined at the level of individual utterances; rather, probabilistic outputs derived from each utterance are aggregated at the session level. This approach reduces the risk of false detections caused by isolated expressions and enables the system to capture consistent patterns of linguistic features over time. Furthermore, the study evaluates whether such a system can be adopted with minimal user burden by investigating its social acceptability.

### *1.4. Related Work*

 doi:10.20944/preprints202604.0668.v1



#### 1.4.1. Text Analysis Techniques for Early Detection of Depression

In recent years, increasing attention has been directed toward the application of text analysis techniques, particularly sentiment analysis, for identifying early signs of depression [6]. Exploring the relationship between linguistic features and mental health has become a key research focus in this domain. Pennebaker et al. (2003) [7] demonstrated that the frequency of first-person pronoun usage is significantly associated with depressive tendencies, suggesting a strong link between language use and psychological states.

Building upon this foundation, the present study adopts an approach centered on first-person pronoun detection as an important indicator for identifying depressive patterns. This perspective provides the theoretical basis for the construction of the first-person classification model used in this research.

#### 1.4.2. Sentiment Analysis and Detection of Obsessive Tendencies

Text-based analysis of obsessive tendencies has also been recognized as a crucial component in detecting early signs of depression. Prior research has reported that compulsive linguistic patterns—such as repetitive expressions and rigid or restrictive language—are characteristic of individuals exhibiting obsessive symptoms (Meyer et al., 2013) [8].

Various computational models have been proposed to extract such features from textual data, including methods based on TF-IDF for feature representation, as well as classification models utilizing RNNs and CNNs. In addition, visualization techniques have been introduced to enhance the interpretability of analytical results. Zhang and Zhao (2020) [9] proposed visualization approaches such as heatmaps and time-series graphs, enabling users to intuitively understand prediction outcomes.

In this study, the model for detecting obsessive tendencies focuses on specific linguistic cues, including restrictive expressions (e.g., “always,” “absolutely”) and repetitive patterns. Compared with existing approaches, the proposed method aims to capture more fine-grained linguistic characteristics and improve the accuracy of identifying obsessive tendencies as indicators of depression.

#### 1.4.3. Detection of Depressive Tendencies from Writing Style

Yang et al. (2023) [10] proposed a chatbot-based system designed to detect early signs of depression through analysis of users’ daily conversations. The system evaluates whether linguistic elements such as first-person expressions and pessimistic language are present in user inputs and uses these features to generate predictive judgments.

For first-person detection, TF-IDF-based methods were applied when subjects were explicitly expressed, whereas LSTM models were employed to handle cases in which subjects were implicit. Pessimistic tendencies were extracted using CNN-based models that capture contextual relationships within text. Additionally, expressions indicative of extreme or absolute thinking (e.g., “always,” “everything”) were incorporated as features to identify obsessive tendencies.

These prior studies highlight the effectiveness of combining multiple linguistic features and machine learning models for detecting depressive patterns. The present research extends these approaches by integrating multiple models into an ensemble framework and applying session-level aggregation to enhance robustness and reliability.

### *1.5. Contributions of This Study*

This study makes several notable contributions to the field of depression detection using natural language processing and conversational systems.

First, it proposes a novel framework for detecting early signs of depression based on linguistic patterns derived from everyday chatbot interactions. Unlike conventional approaches that depend

 doi:10.20944/preprints202604.0668.v1



on clinical questionnaires or self-reported assessments, the proposed method enables continuous and unobtrusive monitoring of users' mental states through natural conversations.

Second, the study introduces a multi-model architecture that captures diverse linguistic features associated with depressive tendencies, including first-person pronoun usage, pessimistic expressions, and obsessive-compulsive writing patterns. By integrating heterogeneous models such as LSTM, CNN, and TF-IDF-based classifiers, the system is able to analyze both contextual and surface-level characteristics of text.

Third, an ensemble-based decision mechanism is employed to combine the outputs of individual models, thereby improving the robustness and reliability of the final prediction. Furthermore, the system adopts a session-level aggregation strategy, which reduces false positives caused by isolated utterances and enables the identification of consistent linguistic patterns over time.

Finally, the proposed system is implemented on a widely used social platform and evaluated through user experiments, including an assessment of its social acceptability. This practical validation demonstrates the feasibility of deploying the system in real-world environments and highlights its potential as a tool for early intervention and continuous mental health monitoring.

## 2. Proposed Method and System Functionality Analysis

This section presents the proposed method for detecting early signs of depression and provides an analysis of the functionality of the developed system. The overall framework is designed to identify depressive tendencies from users' everyday conversational data by extracting and integrating multiple linguistic features.

### *2.1. Overall System Architecture and Analysis*

In this study, conversational data and utterance-level text data associated with individuals exhibiting depressive tendencies are collected from online sources. Based on these data, labeled datasets are constructed to train individual models within the proposed system. By leveraging these datasets, the system is trained to perform depression-oriented analysis, forming the foundation for a chatbot specifically designed to detect early signs of depression.

The proposed system operates by analyzing user-generated text obtained through natural interactions with the chatbot. During these interactions, the system continuously processes input utterances and evaluates whether they contain linguistic patterns indicative of depressive tendencies. Based on this analysis, the system determines the presence or absence of potential depressive signs and generates corresponding outputs, including feedback or recommendations when necessary.

The overall processing flow of the system consists of several stages, including data input, linguistic feature extraction, model-based classification, and final decision making. These processes are integrated into a unified framework that enables real-time analysis of conversational data. An overview of the system architecture and processing pipeline is illustrated in Figure 1.

The data used in this study primarily consist of textual content collected from online sources, including forums where individuals discuss experiences related to depression. Given the limited number of reliable data sources, careful selection and curation of the dataset were conducted. In particular, posts authored by individuals with depressive tendencies and those by non-depressed individuals were explicitly distinguished and organized into separate categories.

For model development, posts associated with depressive individuals were utilized to construct the training corpus, enabling the chatbot to learn linguistic patterns characteristic of depression. In addition, both depressive and non-depressive posts were incorporated into the dataset used for training and evaluation. This design allows the system to be assessed in terms of its ability to accurately distinguish between the presence and absence of depressive tendencies, thereby supporting a more reliable evaluation of predictive performance.

 doi:10.20944/preprints202604.0668.v1



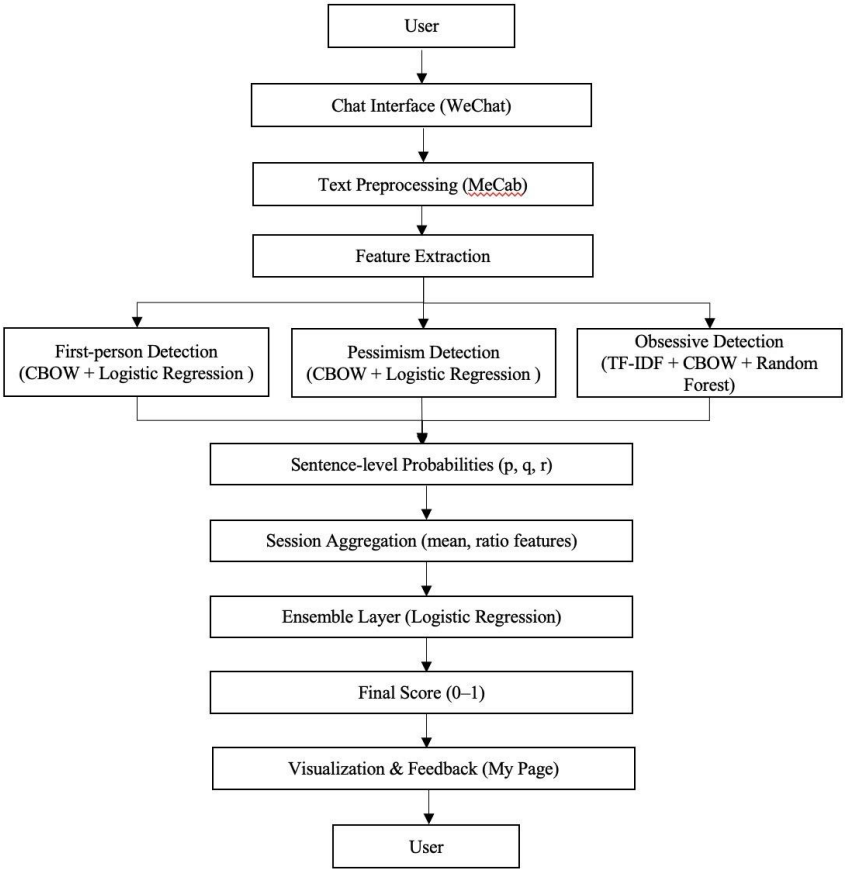


**Figure 1.** System workflow overview.

*2.2. Linguistic Characteristics Observed in Depressive Texts*

Language can generally be analyzed from two complementary perspectives: content and style. Content refers to the semantic aspects of language, including topics, meanings, and conveyed information, whereas style reflects how such content is expressed through linguistic choices and structures.

Previous studies on linguistic patterns associated with depression [11] have reported several distinctive tendencies in the language use of affected individuals. In particular, a higher frequency of first-person singular pronouns and an increased use of negative emotional vocabulary have been consistently observed. Typical examples of negative expressions include adjectives and adverbs such as “lonely,” “sad,” “miserable,” “unhappy,” and “hopeless.” In addition, first-person pronouns, such as “I” and their variants in different linguistic contexts, tend to appear more frequently in depressive texts [12].

In this study, the entire body of user-generated chat text is analyzed to extract these content-based features. Specifically, the system identifies and quantifies the occurrence of negative emotional terms and first-person pronouns, enabling the construction of feature representations grounded in semantic content.

From the perspective of linguistic style, individuals exhibiting depressive tendencies have also been reported to display patterns related to obsessive or rigid thinking [13]. This is often reflected in the frequent use of emphatic and absolutist expressions, such as “completely,” “always,” and “everyone,” which may indicate cognitive rigidity or distorted thinking patterns.

Based on these observations, three key indicators are considered to be strongly associated with early signs of depression: (1) high frequency of first-person singular pronouns, (2) frequent use of negative emotional expressions, and (3) increased presence of obsessive or absolutist language patterns. The proposed system automatically extracts these linguistic features and utilizes them as

 doi:10.20944/preprints202604.0668.v1



input for detecting depressive tendencies, aiming to support diagnostic processes through data-driven indicators.

### *2.3. Model Architecture of the Proposed System*

In this study, a hybrid architecture is designed to detect early signs of depression from textual data by combining multiple models with different analytical approaches. Each model independently performs training and inference, and their outputs are integrated at a later stage to improve overall classification performance.

Figure 2 illustrates the structure of the proposed system. In the input processing stage, user-generated text is first subjected to morphological analysis using MeCab, where each sentence is segmented into individual tokens. The tokenized text is then transformed into feature representations suitable for each model.

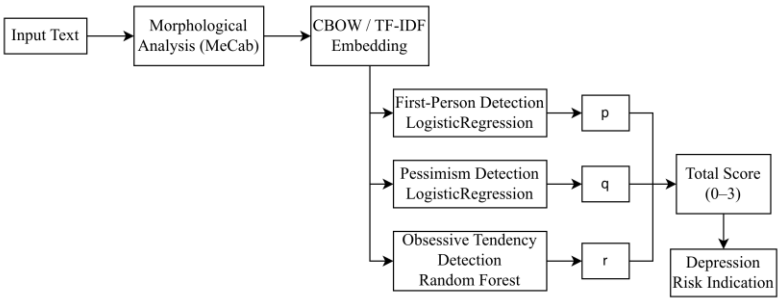


**Figure 2.** System architecture.

Specifically, the first-person detection model and the pessimistic tendency detection model employ CBOW-based distributed representations to capture semantic features of words within their contexts. In contrast, the obsessive tendency detection model utilizes a combination of TF-IDF and CBOW representations, allowing both term importance and contextual semantics to be reflected in the feature space. Through these feature transformation processes, the textual data are converted into formats that facilitate effective learning by each model.

Further details regarding the specific methodologies and configurations are provided in the subsequent subsections describing each individual model.

In the following subsections, the architectures of the first-person detection model, pessimistic tendency detection model, and obsessive tendency detection model are described in detail. Each model is explained in terms of its core components, including input processing, feature extraction, classification, and output generation.

#### 2.3.1. First-Person Pronoun Detection Model

In this study, a model is developed to capture linguistic patterns associated with the use of first-person expressions in Japanese, with the aim of detecting early signs of depression. Previous research has indicated that the frequency of personal pronoun usage can serve as a more reliable indicator of depressive tendencies than negative emotional vocabulary [13]. In particular, the frequent use of first-person pronouns has been identified as a salient linguistic feature of depressive language. Empirical findings suggest that expressions such as “I” and its variants (e.g., watashi, boku, ore, atashi, jibun) appear significantly more often in texts produced by individuals with depressive tendencies compared to non-depressive individuals.

However, accurately identifying first-person usage in Japanese presents unique challenges. Due to the frequent omission of subjects and pronouns, it is often difficult to determine the presence of first-person expressions based solely on explicit word occurrence. In many cases, the subject must be inferred from contextual cues, including verb forms, particles, and surrounding linguistic structures.

 doi:10.20944/preprints202604.0668.v1



Consequently, approaches that rely only on surface-level features, such as term frequency or TF-IDF scores, may fail to capture implicit first-person references.

Previous work by Yang et al. (2023) [10] employed TF-IDF-based methods for detecting first-person usage, assuming that frequently occurring pronouns would yield higher importance scores. While this approach is effective when pronouns are explicitly stated, it becomes less reliable in Japanese contexts where pronoun omission is common.

To address this limitation, the proposed model incorporates a Continuous Bag of Words (CBOW) framework to infer implicit first-person expressions from contextual information. By leveraging distributed word representations, the model captures semantic relationships within sentences and enables the estimation and reconstruction of omitted subjects. This approach allows for a more comprehensive and robust identification of first-person usage patterns, contributing to improved detection of self-focused and inward-oriented linguistic tendencies associated with depression.

Given a sentence $S=\{w1,w2,\ldots,wn\}$, each word is first mapped to a vector representation $v_i \in \mathbb{R}^d$ using the CBOW model. For positions where first-person expressions are inferred to be implicitly omitted based on contextual information, representative first-person pronoun vectors (e.g., corresponding to "I") are probabilistically inserted according to a maximum likelihood estimation.

Based on this augmentation, the sentence-level representation $v_S$ is recalculated as follows:

$$v_S=\frac{1}{n}\sum_{i=1}^{n} v_i'$$

Here, $v_i'$ denotes the word vector after the estimation-based augmentation, and $\mathbb{R}^d$ represents the $d$-dimensional vector space derived from Word2Vec (with $d$=200 in this study). The resulting sentence representation $v_S$ is then fed into a logistic regression model to estimate the probability $p$ that the sentence contains first-person expressions, as defined by the following equation:

$$p=\sigma(w_T v_S+b)$$

Here, $\sigma(x)=\frac{1}{1+e^{-x}}$ denotes the sigmoid function, $w$ is the weight vector, $v_S$ represents the sentence vector, and $b$ is the bias term. Through the sigmoid transformation, the output is mapped to a value between 0 and 1, corresponding to the probability that the sentence contains first-person expressions. A threshold of 0.5 is applied for classification: sentences with predicted probabilities greater than or equal to 0.5 are classified as containing first-person expressions, while those below 0.5 are classified as not containing them.

During training, the logistic regression model learns weights associated with each feature, reflecting their contribution to the classification task. Features with larger weights have a stronger influence on identifying first-person usage. The top-ranked features after training include terms such as "I" and its variants (e.g., watashi, jibun, boku, ore), as well as related expressions such as "think," "feel," and "do not understand." These results indicate that such words play a significant role in determining the presence of first-person expressions.

By combining CBOW-based word embeddings with logistic regression, the proposed approach achieves accurate prediction of first-person expression usage in Japanese text. Ultimately, each input sentence is classified as either "first-person present (1)" or "first-person absent (0)," and the resulting output serves as an important component in the detection of early signs of depression.

### 2.3.2. Pessimistic Tendency Detection Model

In this study, a model is constructed to identify sentences that contain pessimistic or negative emotional expressions as indicators of early depressive tendencies. Linguistic studies have shown that expressions such as "impossible," "useless," "it's over," and "there is no hope" are frequently observed in texts produced by individuals with depressive symptoms and serve as important markers of negative mental states [14]. The objective of this model is to capture the occurrence and distribution of such expressions and to quantitatively evaluate the degree of pessimistic tendency within a sentence.

 doi:10.20944/preprints202604.0668.v1



Previous work by Yang et al. (2023) [10] employed convolutional neural networks (CNNs) to detect pessimistic tendencies. CNNs, originally developed for image processing tasks, have been successfully applied to natural language processing by scanning input word sequences with sliding windows (n-grams) to extract local patterns. This enables the detection of characteristic phrase-level expressions such as "nothing is enjoyable," "it's hopeless," and "there is no hope," which are indicative of pessimistic sentiment.

In contrast to approaches that rely on sequential pattern extraction, the proposed method adopts a combination of CBOW-based word embeddings and logistic regression. By representing each word as a distributed vector and computing the average of these vectors, the model obtains a sentence-level representation that captures global semantic information without being strictly dependent on word order. This sentence vector is then input into a logistic regression classifier, which estimates the probability that the sentence expresses a pessimistic tendency. Through this approach, both semantic coherence and overall contextual meaning can be effectively incorporated into the classification process.

The semantic representation of a sentence is obtained by computing the average of the word vectors generated by the CBOW model. This averaged vector is then used as input to a logistic regression classifier, which predicts whether the sentence expresses a pessimistic tendency.

More specifically, the logistic regression model estimates a probability $q$ indicating the degree of pessimism based on the sentence-level vector representation. A threshold of 0.5 is applied for classification: if $q \geqslant 0.5$, the sentence is classified as "pessimistic," whereas if $q<0.5$, it is classified as "non-pessimistic."

Through this approach, pattern detection traditionally performed by CNNs—based on word order and local n-gram structures—is reinterpreted as a bias in the semantic vector space. As a result, such patterns can be approximated using a linear model operating on distributed representations, enabling efficient and interpretable classification.

The mathematical formulation of the proposed model is described as follows. Let an input sentence $S$ consist of $n$ words, denoted as $\{w1,w2,\ldots,wn\}$. Each word $w_i$ is transformed into a vector representation $v_i \in \mathbb{R}^d$ using the CBOW model.

The sentence-level representation $v_S$ is defined as the average of these word vectors:

$$v_S=\frac{1}{n}\sum_{i=1}^{n} v_i$$

where $v_i$ denotes the distributed representation of the word $wi$, and $\mathbb{R}^d$ represents the $d$-dimensional vector space produced by Word2Vec.

Next, logistic regression is applied to the sentence vector $vS$ to estimate the probability $q$ that the sentence exhibits a pessimistic tendency.

$$q = \sigma(w_{pess}^{T} \cdot v_S + b_{pess})$$

Here, $\sigma(x)$denotes the sigmoid function, defined as $\sigma(x)=\frac{1}{1+e^{-x}}$. The parameter $w_{pess}$represents the weight vector, and $b_{pess}$ is the bias term.

The computed probability $q$ is used for binary classification: if $q \geqslant 0.5$, the sentence is classified as "pessimistic," whereas if $q<0.5$, it is classified as "non-pessimistic."

During training, words associated with larger weights in the logistic regression model are identified as salient features. In other words, terms that exert a stronger influence on the predicted pessimistic probability $q$ of the sentence vector are regarded as important. Representative feature words emphasized by the model include expressions such as "impossible," "useless," "it's over," "I want to disappear," "anyway," "anxious," "the worst," "painful," "lonely," and "denial." These terms are assigned relatively high weights in the CBOW-based vector space and play a significant role in the classification process.

The training procedure is conducted using a labeled dataset consisting of pessimistic and neutral sentences. First, word embeddings are generated using the CBOW model, and sentence vectors are subsequently constructed by aggregating these embeddings. These sentence representations are then used to train the logistic regression classifier.

 doi:10.20944/preprints202604.0668.v1



Ultimately, the model is designed to capture negative and future-oriented pessimistic expressions embedded in text and to facilitate early detection of depressive tendencies. Given an input sentence, the trained model outputs a binary classification result: "pessimistic (1)" or "non-pessimistic (0)."

2.3.3. Obsessive Tendency Detection Model

In this study, an obsessive tendency detection model was designed to capture compulsive and repetitive patterns of thought and language as one of the early indicators of depression. Expressions such as "by all means," "without fail," "absolutely," "forever," "again and again," and "I keep overthinking" [15] are often regarded as linguistic signals of excessive fixation or difficulty in cognitive control. Because such lexical and structural patterns may reflect obsessive tendencies associated with depression, this study focuses on modeling these features in order to predict obsessive language tendencies.

Previous work by Yang et al. (2023) [10] employed TF-IDF and frequent-word extraction to identify obsessive tendencies. However, approaches based solely on word frequency are limited in their ability to capture semantic repetition and rigid linguistic patterns embedded in sentence meaning. To address this issue, the present study adopts a Random Forest-based approach, enabling the model to learn more fine-grained combinations of linguistic features and thereby achieve more accurate classification of obsessive tendencies.

The detailed structure of the model is described in the following subsection.

In the obsessive tendency detection model, the input sentence is first decomposed into word units through morphological analysis. Subsequently, TF-IDF is applied to compute the importance of each word within the document, allowing the model to emphasize terms associated with obsessive tendencies.

To further capture semantic similarity and contextual information, each word is transformed into a distributed representation using the Word2Vec CBOW model. A sentence-level vector is then constructed by aggregating the word vectors. The TF-IDF scores and CBOW-based representations are integrated to form the input feature vector for the Random Forest classifier.

Specifically, the contribution of each word vector $v_i$ is weighted by its corresponding TF-IDF score $t_i$, and the sentence vector $vS$ is computed as a weighted average:

$$v_S = \frac{1}{\sum_{i=1}^{n} t_i} \sum_{i=1}^{n} t_i \cdot v_i$$

This formulation ensures that words with higher TF-IDF scores exert a greater influence on the sentence representation. As a result, terms strongly associated with obsessive tendencies—such as "repeat" or "handwashing"—are emphasized, while less informative words, such as auxiliary verbs or function words, contribute minimally.

The resulting weighted sentence vector $vS$ is then used as the input feature for the Random Forest classifier, which performs the final classification of obsessive versus non-obsessive tendencies.

Subsequently, a Random Forest classifier, which consists of an ensemble of decision trees, is applied to classify the input feature vector $x$ as either obsessive or non-obsessive. The final prediction is determined by majority voting over the outputs of individual trees. Specifically, each decision tree $Tm(x)$ produces a classification result, and the final prediction $y$ is obtained as:

$$y = \text{majority}\{T_m(x)\} \quad , \quad (m = 1 \ldots M)$$

where $y$ denotes the final classification outcome, taking the value 1 for obsessive sentences and 0 for non-obsessive sentences.

Through this approach, the model effectively integrates both word-level importance and contextual semantic information to detect obsessive linguistic patterns. Furthermore, the Random Forest framework enables the estimation of feature importance, allowing the identification of words that contribute most significantly to the prediction. The most influential features identified during training include terms such as "checking," "handwashing," "anxiety," "perfection," "dirty,"

 doi:10.20944/preprints202604.0668.v1



"counting," "repetition," "compulsion," "must do," and "cannot relax." These expressions are characteristic of obsessive thoughts and behaviors and play a critical role in classification.

By combining TF-IDF-based weighting with CBOW-derived contextual representations, the proposed model achieves accurate detection of sentences exhibiting compulsive and repetitive linguistic patterns. This capability is particularly valuable for identifying depressive tendencies associated with obsessive cognition and has potential applications in clinical data analysis and mental health support systems.

### *2.4. Output Integration and Ensemble Prediction*

In the proposed system, three individual models—the first-person detection model, the pessimistic tendency detection model, and the obsessive tendency detection model—each produce probability scores at the sentence level. These outputs are defined as follows:

$p$: probability of self-focused expression derived from the first-person model

$q$: probability of negative or pessimistic attitude derived from the pessimistic model

$r$: probability of compulsive or repetitive tendency derived from the obsessive model

After obtaining these probabilities for each sentence, the results are aggregated at the session level, where a sequence of user utterances is considered as a unit. From this aggregation, the following statistical features are computed:

- Mean probabilities: $\bar{p},\bar{q},\bar{r}$
- Proportions of sentences exceeding predefined thresholds: $\pi_p,\pi_q,\pi_r$

These features are combined into a feature vector:

$$x_s=[\bar{p},\bar{q},\bar{r},\pi_p,\pi_q,\pi_r]$$

This vector is then input into an ensemble classifier based on logistic regression, which estimates the final depression risk score $y_s\hat{}\in[0,1]$ as follows:

$$\hat{y_s} = \sigma(W^T x_s + b)$$

where $\sigma(\cdot)$ denotes the sigmoid function, $W$ is the weight vector, and $b$ is the bias term.

A threshold $\gamma$(set to 0.5 in this study) is applied to determine the final prediction:

$y_s\hat{}\geqslant\gamma$: presence of depressive tendency

$y_s\hat{}<\gamma$: absence of depressive tendency

This session-level integration mitigates the limitations of single-sentence analysis and enables more stable and reliable detection based on continuous linguistic patterns.

### *2.5. System Configuration and Introduction of Evaluation Questionnaires*

Upon accessing the system, users are first presented with a login interface. After successful authentication, they are granted access to both the chatbot function and a personalized dashboard (My Page), which supports continuous interaction and data visualization.

In addition to the chatbot functionality, a questionnaire module is incorporated in this study to evaluate the diagnostic performance of the system. Responses collected from participants are used to compare and validate the automated predictions generated by the chatbot. It should be noted that this questionnaire module is employed solely as an evaluation tool for research purposes and is not intended to provide direct feedback to users.

Two widely used self-assessment scales for depression in Japan are adopted in the questionnaire: the Self-Rating Depression Scale (SDS) [16] and the Beck Depression Inventory (BDI) [17,18]. The SDS is a self-report instrument designed to assess the level of depressive symptoms and was originally developed by Zung at Duke University in 1965. The Japanese version was later translated and adapted by Fukuda (Tohoku University) and Kobayashi (University of Tsukuba), and it is commonly utilized in clinical settings, including psychosomatic medicine. In recent years, the SDS has also been applied as a tool for monitoring treatment progress and for mental health screening [13].

In addition to the SDS, the Beck Depression Inventory (BDI) is employed as a complementary assessment tool. The BDI is a widely recognized self-report questionnaire developed by Beck et al. to

 doi:10.20944/preprints202604.0668.v1



measure the severity of depressive symptoms. It consists of multiple items that evaluate various aspects of depression, including mood, cognitive patterns, and physical symptoms. Each item is scored on a graded scale, and the total score reflects the overall level of depressive severity.

The BDI has been extensively validated across different populations and is commonly used in both clinical and research settings due to its high reliability and sensitivity. Compared with the SDS, which primarily captures general depressive tendencies, the BDI provides a more detailed assessment of symptom severity and cognitive characteristics. In this study, the BDI is used to complement the SDS by offering a more comprehensive evaluation of participants' mental states, thereby enabling a more robust comparison with the system's prediction results.

#### 2.5.1. Chat Function

The chat interface enables real-time analysis of user input through natural daily conversations. During interaction, the system automatically extracts linguistic features associated with depressive tendencies from the user's utterances. Each input sentence is processed immediately, and the system performs classification based on key indicators, including the use of first-person expressions, the presence of pessimistic emotional vocabulary, and the occurrence of obsessive or repetitive language patterns.

Figure 3 presents a screenshot of the chat interface. This design allows users to engage in conversations without being explicitly aware of undergoing a diagnostic process, thereby encouraging more natural expression of their psychological state.

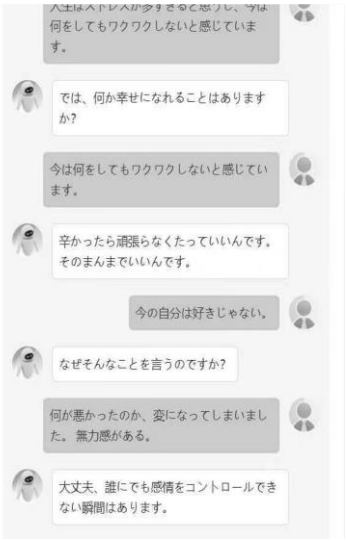


**Figure 3.** Chat page interface.

#### 2.5.2. My Page Function

The My Page interface allows users to review their past chat history as well as the results of the system's analysis. When potential depressive tendencies are detected, the system visualizes the results through scores and category-specific evaluations, enabling users to gain an objective understanding of their mental state.

In the evaluation experiment, the diagnostic results presented on the My Page were compared with the outcomes obtained from the SDS and BDI questionnaires. The degree of agreement between these results was used to assess the accuracy and effectiveness of the proposed system.

## 3. Evaluation and Results of the Proposed System

### *3.1. Experiment on Detection of Depressive Tendencies*

 doi:10.20944/preprints202604.0668.v1



To evaluate the effectiveness of the proposed system in detecting early signs of depression, an experimental study was conducted with 20 participants (N = 20; 13 males and 7 females). The primary objective of this experiment was to assess the feasibility and performance of the chatbot-based approach for identifying depressive tendencies.

During the experiment, participants were instructed to interact with the chatbot system for a specified period. After the interaction, they completed the Self-Rating Depression Scale (SDS) questionnaire. The results obtained from the chatbot-based detection were then compared with the SDS scores to evaluate the system's predictive performance.

Based on this comparison, evaluation metrics such as accuracy and recall were calculated to assess the classification performance of the proposed system. In addition, statistical analysis was conducted on the experimental results obtained from the participants.

The results are summarized in Table 1, where "○" indicates the presence of depressive tendencies and "×" indicates the absence of such tendencies.

**Table 1.** Experimental results.

| | Questionnaire | System Output | Ground Truth |
|---|---|---|---|
| **Subject 1** | × | ○ | × |
| **Subject 2** | × | × | × |
| **Subject 3** | ○ | ○ | ○ |
| **Subject 4** | × | × | × |
| **Subject 5** | ○ | × | ○ |
| **Subject 6** | × | × | × |
| **Subject 7** | ○ | ○ | ○ |
| **Subject 8** | × | ○ | × |
| **Subject 9** | ○ | ○ | ○ |
| **Subject 10** | × | × | × |
| **Subject 11** | ○ | × | ○ |
| **Subject 12** | × | × | × |
| **Subject 13** | × | × | × |
| **Subject 14** | × | ○ | × |
| **Subject 15** | × | × | × |
| **Subject 16** | ○ | ○ | ○ |
| **Subject 17** | × | × | × |
| **Subject 18** | × | ○ | × |
| **Subject 19** | × | × | × |
| **Subject 20** | ○ | ○ | ○ |

To compare the system predictions with the questionnaire results, standard evaluation metrics—including accuracy, precision, and recall—were calculated. The dataset consisted of 7 positive samples and 13 negative samples. Ideally, all positive samples should be correctly identified. In this experiment, the system predicted 8 samples as positive, among which 5 were true positives.

Accordingly, the confusion matrix is shown in Table 2, from which the performance metrics were derived. The computed accuracy, precision, and recall values are summarized in Table 3.

**Table 2.** Confusion matrix.

| | Predicted Positive | Predicted Negative |
|---|---|---|
| **Actual Positive** | 5 (TP) | 2 (FN) |

 doi:10.20944/preprints202604.0668.v1



| | | |
|---|---|---|
| **Actual Negative** | 4 (FP) | 9 (TN) |

**Table 3.** Evaluation metrics of the proposed system based on the confusion matrix.

| Metric | Formula | Value |
|---|---|---|
| Accuracy | (TP + TN) / (TP + FP + FN + TN) | 0.700 |
| Precision | TP / (TP + FP) | 0.556 |
| Recall | TP / (TP + FN) | 0.714 |

*3.2. Overall Evaluation of the Proposed System*

For the integrated output, each individual model was first trained using labeled datasets. The performance of the models was then evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The final integrated score, denoted as $S_{total}$, is used as an indicator to assess the overall depressive tendency of a user. This score is calculated by aggregating the outputs of the three individual models. The interpretation of $S_{total}$ is as follows:

$S_{total}$=0: All models indicate the absence of depressive tendencies.

$S_{total}$=3: All models indicate strong depressive tendencies.

Thus, the score ranges from 0 to 3, and the final evaluation is obtained by summing the outputs of the individual models. This score also reflects the confidence level of the classification results.

**Table 4.** Summary of evaluation results (Test data: 2,000 sentences per category).

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| First-person Detection | 0.865 | 0.876 | 0.850 | 0.863 |
| Pessimism Detection | 0.840 | 0.848 | 0.812 | 0.830 |
| Obsessive Tendency Detection | 0.815 | 0.814 | 0.769 | 0.791 |
| | **TP** | **FP** | **FN** | **TN** |
| First-person Detection | 850 | 120 | 150 | 800 |
| Pessimism Detection | 780 | 140 | 180 | 900 |
| Obsessive Tendency Detection | 700 | 160 | 210 | 930 |

In addition to evaluating classification performance, a user study was conducted to assess the acceptability and practical demand for the system. The results indicated that a majority of users recognized the necessity of such a system and provided constructive feedback regarding its functionality.

Based on these findings, several directions for future improvement were identified, including enhancing the naturalness and continuity of conversations, incorporating advisory and behavioral support functions, and designing notification mechanisms that take into account users' psychological burden.

## 4. Discussion

The results of this study demonstrate that the proposed chatbot-based system is effective in detecting early signs of depression through linguistic analysis of everyday conversations. The effectiveness of the system can be attributed to several key factors related to feature selection, model design, and system integration.

First, the use of linguistically grounded features—namely, first-person pronoun usage, pessimistic expressions, and obsessive-compulsive language patterns—provides a theoretically supported basis for detecting depressive tendencies. These features are closely associated with self-focused attention, negative cognitive bias, and rigid thinking patterns, all of which are well-documented characteristics of depression. By focusing on such features, the system is able to capture subtle changes in language that may not be easily identifiable through conventional questionnaire-based methods.

 doi:10.20944/preprints202604.0668.v1



Second, the adoption of a hybrid modeling approach enhances the robustness of the system. Each individual model captures a different aspect of depressive language: the first-person model reflects self-referential tendencies, the pessimistic model identifies negative semantic content, and the obsessive model detects repetitive and rigid expressions. By integrating these complementary perspectives through an ensemble framework, the system achieves a more comprehensive understanding of the user's linguistic behavior.

Third, the introduction of session-level aggregation contributes significantly to improving detection stability. Unlike approaches that rely on single utterances, the proposed method evaluates patterns across multiple conversational turns. This reduces the influence of noise and isolated expressions, enabling more reliable assessment of persistent psychological tendencies over time.

In addition, the use of a chatbot interface plays an important role in reducing user burden and encouraging natural interaction. Because users are not required to explicitly respond to diagnostic questionnaires, they can express their thoughts more freely, leading to more authentic linguistic data. This design improves both usability and data quality, which are critical factors for continuous monitoring systems.

However, several limitations should be noted. The sample size of the experimental evaluation was relatively small, which may limit the generalizability of the results. Furthermore, the system currently focuses primarily on textual features and does not incorporate other modalities such as speech, facial expressions, or behavioral data. Future work should address these limitations by expanding the dataset, incorporating multimodal analysis, and refining the interaction design to further enhance user engagement and diagnostic accuracy.

## 5. Conclusions

This study proposed a chatbot-based system for detecting early signs of depression through linguistic analysis of everyday conversations. By focusing on key linguistic features—first-person pronoun usage, pessimistic expressions, and obsessive-compulsive language patterns—the system enables non-intrusive and continuous monitoring of users' mental states.

A hybrid modeling approach combining CBOW-based representations, machine learning classifiers, and an ensemble framework was introduced to improve detection accuracy and robustness. In addition, session-level aggregation was employed to capture consistent linguistic patterns over time, reducing the impact of noise from individual utterances.

Experimental results demonstrated that the proposed system achieved reliable classification performance and showed consistency with established self-report measures. Furthermore, user evaluation indicated that the system was perceived as useful and acceptable, suggesting its potential applicability in real-world settings.

Overall, the proposed approach provides a practical and scalable framework for early detection of depressive tendencies. Future work will focus on expanding the dataset, improving model performance, and incorporating multimodal information to further enhance the effectiveness of the system.

## References

1. Lv, Y.; Zheng, X.; et al. Study on chai robot based on crowd-sourcing. Information Technology. 2017(4); p. 102–103＋109.https://doi.org/10.13274/j.cnki.hdzj.2017.04.026
2. Ministry of Health, Labour and Welfare (Japan). "Patient Survey (Kanja Chosa)." Available online: https://www.mhlw.go.jp/toukei/saikin/hw/kanja/ (accessed on 19 December 2024).
3. Ministry of Health, Labour and Welfare (Japan). "Summary of the Suicide and Depression Countermeasures Project Team (Jisatsu・Utsubyo Taisaku Project Team Torimatome)." Available online: https://www.mhlw.go.jp/seisaku/2010/07/03.html (accessed on 19 December 2024).
4. American Psychiatric Association.Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5). Washington, DC, American Psychiatric Publishing, 2013; p. 947.

 doi:10.20944/preprints202604.0668.v1




5. Inkster, B.; Sarda, S.; et al. Emotional support chatbot (Wysa) for mental health: Mixed methods study. JMIR mHealth and uHealth. 2018, 6(8); e12106.https://doi.org/10.2196/12106
6. Yu, Z.; Zhao, J. Detecting depression from social media posts using ensemble learning techniques. Journal of Affective Disorders. 2018, 238; p. 99–105.https://doi.org/10.1016/j.jad.2018.05.010
7. Mehl, M. R.; Pennebaker, J. W.; et al. G.Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology. 2003, 54; p. 547–577.https://doi.org/10.1146/annurev.psych.54.101601.145041
8. Meyer, B.; Marks, R. Cognitive-behavioral therapy for obsessive-compulsive disorder: A comprehensive guide to treatments and interventions. New York, Springer, 2013; p. 167–180.https://doi.org/10.1007/978-1-4614-6562-3
9. Zhao, J.; Zhang, Y. Text classification techniques: A survey. International Journal of Computer Science Issues. 2020, 17(3); p. 85–98.
10. Mori, H. H.; Yang, M. Detecting signs of depression using chatbots—extraction of the first person from Japanese. Human Interface and the Management of Information. Springer, 2023; p. 660–671. https://doi.org/10.1007/978-3-031-35132-7_48
11. Gortner, E. M.; Rude, S. S.; et al. Language use of depressed and depression-vulnerable college students. Cognition & Emotion. 2004, 18(8); p. 1121–1133.https://doi.org/10.1080/02699930441000030
12. Shinagawa Mental Clinic. “Behaviors, Language, and Facial Expressions Observed in Individuals with Depression (Utsubyo no Hito ga Toru Kōdō, Tsukau Kotoba, Kao no Hyōjō).” Shinagawa Mental Clinic Column. Available online: https://www.shinagawa-mental.com/column/psychosomatic/4words/ (accessed on 19 October 2025).
13. The Conversation. “People with depression use language differently—here’s how to spot it.” The Conversation. Available online: https://theconversation.com/people-with-depression-use-language-differently-heres-how-to-spot-it-90877 (accessed on 25 October 2025).
14. Lauer, M.; Smirnova, D.; et al. Language patterns discriminate mild depression from healthy controls. Frontiers in Psychiatry. 2018, 9; Article 105.https://doi.org/10.3389/fpsyt.2018.00105
15. Al-Mosaiwi, M.; Johnstone, T. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science. 2018, 6(4); p. 529–542.https://doi.org/10.1177/2167702617747074
16. Iidabashi Psychosomatic Medicine Clinic. “SDS: Self-Rating Depression Scale (Utsusei Jiko Hyōka Shakudo).” Available online: https://www.iidabashi-shinryounaika.jp/20170818150214 (accessed on 25 October 2025).
17. Zimmerman, M. Using scales to monitor symptoms and treatment of depression (measurement based care). In: Rose, B. D. (Ed.), UpToDate. Waltham, MA, UpToDate, Inc., 2011.
18. Hojat, M.; Shapurian, R.; et al. Psychometric properties of a Persian version of the short form of the Beck Depression Inventory for Iranian college students. Psychological Reports. 1986, 59(1); p. 331–338.