C-Journal: A Journaling Application for Detecting and Classifying Cognitive Distortions using Deep-Learning based on a Crowd-sourced Dataset

Anonymous submission

Abstract

Cognitive distortions are negatively biased thinking patterns and erroneous self-statements resulting from and leading to logical errors in one's own internal reasoning. Cognitive distortions have an adverse effect on mental health and can lead to mental health disorders in extreme cases. This paper belongs to a bigger project which aims to provide an application for detecting and classifying cognitive distortions in texts. As no public data sets were available for the task, the first contribution of the proposed work lies in providing an open-source labeled dataset of 14 cognitive distortions consisting of 34370 entries collected via crowd-sourcing, user questionnaires, and re-purposing emotions dataset from social media. The dataset is collected in cooperation with a licensed psychologist. We implemented a baseline model using Naïve Bayes and Count Vectorizer and different CNN, LSTM, and DNN classifiers to classify cognitive distortions based on the dataset. We investigated the usage of different word embeddings with the best-performing models. The best-performing model relied on a CNN with pre-trained Sentence-BERT embedding with an F1-score of 84% for classifying cognitive distortions. The best-performing model was built into C-Journal, a free journaling and mood-tracking mobile application that pinpoints potential thinking distortions to the users.

Keywords: Cognitive Distortions, Crowd-sourcing, Data Collection, Text Classification, Mental Health

1. Introduction

The mental health state is a critical part of well-being and is a main interest for researchers. According to Prince (Prince et al., 2007) et al., around 14% of the global burden of disease is because of mental disorders and their chronically disabling nature. Negative biases in thinking, also known as cognitive distortions, can increase negative well-being and are the main risk factors for depression and anxiety (Rnic et al., 2016). Cognitive distortions are self-statements that are either completely erroneous or only slightly misunderstood and most frequently followed by logical errors in one's internal reasoning. It is important to note that cognitive distortions can be experienced by healthy individuals and do not necessarily entail a mental disorder. However, they decrease mental well-being and can lead to mental disorders if exhibited frequently in the long run. Traditionally, one-on-one meetings and paper-and-pen diaries have been used for raising awareness and reducing cognitive distortions. These methods of monitoring and assessment require professional assistance. Given the number of individuals seeking help from a mental health professional, it is not easy for caregivers to give consistent mentoring for detecting and correcting negative thinking patterns. Traditional mental health monitoring and evaluation methods are insufficient today due to the fast-paced nature of living. In order to provide ease, anonymity, an introduction to caring, service to more individuals, consistency, 24-hour service, cost efficiency, and support, a digital mental health revolution was required. Despite the focus on cognitive distortions in cognitive behavioral theory and practice, little research has been done in the field from the perspective of Digital Psychology

This work is part of a bigger project aiming at developing applications to raise awareness of cognitive distortions while preserving user privacy. The aim is to provide new explanatory viewpoints on the statistical correlations between undesired thinking errors and well-being. These include self-help applications for individuals, as well as tools for aiding mental health caregivers, i.e., psychiatrists and life coaches. This paradigm has the potential to revolutionize psychology since it will help therapists and psychiatrists identify cognitive distortions by providing them with immediate feedback and making better use of their counseling sessions. Additionally, it could make mental healthcare treatments more easily accessible. This can be accomplished by applying and comparing various state-of-the-art techniques for classifying annotated data into two categories, distorted thinking and non-distorted thinking, then digging deeper to identify which type of distortion it is.

This paper's primary objective is to provide an application for identifying cognitive distortions from journaling texts in English natural language. Like most other natural language processing tasks, this is a heavily data-dependent task. A lack of sufficient data is one of the most prevalent impediments to employing AI for Clinical Psychology. Given that detecting cognitive distortion is a relatively new task with scarce resources and even scarcer open-source ones, we needed to collect and annotate our own dataset, maintaining all ethical aspects to be able to train models to detect cognitive distortions.

LREC-COLING 2024, pages 3224–3234

The main contributions of the proposed work are three-fold. The first is collecting and providing an open-source dataset of cognitive distortions collected via different modalities, including crowdsourcing, surveys, and re-purposing existing emotions datasets. The second is training baseline and main models for classifying whether sentences exhibit cognitive distortions or not. In the former case, the cognitive distortion is classified as one of the 14 included cognitive distortions: "Mental Filtering ", "Polarization", "Overgeneralization", "Catastrophizing", "Personalization", "Control fallacies ", "Fallacy of Fairness", "Blaming", "Shoulds", "Emotional Reasoning ", "Fallacy of Change", "Global Labeling", "Always Being Right", "Jumping to Conclusions ". The contribution lies in developing C-Journal, a mobile journaling and mood-tracking application built on the basis of the best-performing cognitive distortions classification model. The application identifies and raises awareness of potential cognitive distortions in the users' daily journaling texts.

The paper is organized as follows. Section 2 presents some related work for automatically detecting mental health issues and cognitive distortions from natural language. In Section 3, the data collection and labeling approaches are presented along with the data statistics. Section 4 explains the conducted experiments and the developed models for classifying cognitive distortions and their results. Section 5 gives an overview of the developed C-Journal app. Finally, Section 6 concludes the paper and gives suggestions for future extensions of this work.

2. Related Work

In this section, we provide an overview of the most relevant related work concerning collecting data for classification tasks related to human factors. We also discuss available work on ML techniques in relation to digital psychology tasks. In addition, we present some of the technological guides for developing application and technologies related to mental health.

2.1. Dataset Collection

Mental health is an important aspect of overall well-being and a subject of extensive research (Prince et al., 2007). A significant portion of the global illness burden, around 14%, is attributed to conditions such as depression and anxiety (Prince et al., 2007). Traditionally, monitoring and assessing mental health involves one-on-one sessions, written diaries, and paper-and-pencil tests. However, these methods have limitations, including limited availability, time consumption, and the inability to recall every instance. Given the fast-paced nature of life, there is a need for a digital revolution in mental health care that can provide convenience, anonymity, continuous support, and cost-effectiveness to a larger population.

Researchers have explored creative technology-based solutions to address the challenges in measuring and improving mental health. Analytical methodologies have focused on data collection through surveys, internet logs, and biometrics (Galderisi et al., 2015). Artificial neural networks and machine learning techniques like deep learning and transfer learning have been recommended for mental health status analysis. Several contributions were highlighted in a literature review spanning from 2012 to 2022. These include psychological data collection methods, data preprocessing, machine learning techniques, technological options for monitoring and enhancing mental health and addressing open gaps for future work.

One main approach for collecting datasets is through Social Media APIs. Social networking apps allow individuals to express themselves, and APIs provide a reliable way to acquire large amounts of data from these platforms. However, data collected through social media may have certain criteria biases and skewed representation. Studies have used filtered words related to mood states to collect tweets from Twitter APIs (Luo et al., 2018; Sadasivuni and Zhang, 2020; Chang et al., 2016; Deshpande and Rao, 2017). Other researchers have collected data from online blogs and communities, filtering them based on specific keywords like "depression" (Saha et al., 2016; Shickel et al., 2020b).

Some researchers have developed online systems to detect psychological problems, collecting data through test modules that users interact with (Zhu and He, 2021). Other studies have monitored research participants, including university staff and students, by collecting sensor data, questionnaire data, voice data, and test data (Sun et al., 2012; Liu et al., 2021; An et al., 2019; de Santos Sierra et al., 2011). Appointments made online via recruitment websites have also been used to detect mental illness and measure changes in mental health states (Arefin et al., 2021; Wongkoblap et al., 2017).

The use of IoT devices has revolutionized data collection, providing accurate data at higher rates. Studies have focused on analyzing the use of IoT devices such as sensors, security devices, and intelligent appliances. Environmental and biological data have been collected, along with physiological data, to correlate mental and physical health (Park et al., 2020; Maniyath et al., 2021; Li, 2021; Amate et al., 2021). Electrocardiogram, accelerometer, and galvanic response data have been measured for real-time stress detection (Sun et al., 2012; de Santos Sierra et al., 2011).

Data science has played a crucial role in understanding public mental health issues and developing solutions based on technological devices. Researchers have developed mobile applications to detect conditions like borderline personality disorder, psychological states, perinatal mental health problems, and general mental health issues. These studies have employed various data analysis techniques and machine learning algorithms (Khazbak et al., 2021; Dyriv et al., 2021; Wang et al., 2020; Kogilathota, 2021).

In summary, the field of mental health research has embraced technology-based solutions for monitoring and improving mental well-being. Creative data collection and analysis methods, along with the use of IoT devices and data science, have paved the way for advancements in this field.

2.2. Machine Learning Techniques

We reviewed many publications that examine how machine learning is applied in the field of psychology and provide the most innovative methods. The use of machine learning in the field of mental health has considerable advantages for diagnosis, treatment, and support. The majority of the research focused on diagnosing mental health problems and providing mental health care. We talk about machine learning strategies, possibilities, unresolved problems, and state-of-the-art techniques.

Existing techniques for stress detection include SVMs (Park et al., 2020; Sun et al., 2012) and fuzzy logic (de Santos Sierra et al., 2011). Jain et al.(Jain et al., 2020) implemented an OpenCV HAAR CASCADES approach and sentiment analysis to detect mental health states. They detected the face using webcams using TensorFlow Image Classifier. Experiments using machine learning algorithms to detect users' psychological states for the sake of mental health improvement are presented in (Crasto et al., 2021; BH et al., 2022; Wongkoblap et al., 2017; Katarya and Maan, 2020). Machine learning approaches for text classification for emotional valence and mental health issues prediction are presented in (Shickel et al., 2020b; Khazbak et al., 2021; Shickel et al., 2020a; An et al., 2019; Vajre et al., 2021; Laden et al., 2020; Wang et al., 2020; Kogilathota, 2021). The techniques used included logistic regression, deep transfer learning approach with BERT, LSTM, convolutional neural network, PsychBERT, Research Domain Criteria (RDoC), SVM, and artificial neural networks. Gao and Shi(Gao and Shi, 2021) experimented with a similar trajectory clustering algorithm to evaluate the mental health status of college students. They divided the data into 4 types, reaching accuracies between 88% to 96%. Li et al.(Li et al., 2021) implemented a combined mathematical model of artificial neural network and fuzzy math theory to evaluate the mental health state of college students. This improves the accuracy of the mathematical model. Sadasivuni et al.(Sadasivuni and Zhang, 2020) used the Autoregressive Integrated Moving Average ARIMA model to predict mental health state. ARIMA is a statistical analysis model for prediction using time series. ARIMA showed great performance as the number of the detected data has a similar pattern to World Health Organization WHO reports. Larsen et al.(Larsen et al., 2015) used PCA to map the emotions of users by classifying emotional tweets in real time.

2.3. Technological Solutions and Wellbeing Guides

Currently, data science is widely used to analyze public mental health issues and develop solutions (Varnfield et al., 2019; Lashari et al., 2021; Khazbak et al., 2021; Roy et al., 2017; Cárcamo et al., 2021). However, there is still untapped potential in the field of mental health technology. Despite the prevalence of mental health concerns, there are challenges in accessing adequate technological support, including flawed systems, ethical dilemmas, and security concerns.

Varnfield et al. introduced a web platform called Health-e Minds, which supports individuals with mental health disorders by providing guidance, monitoring health measures, evaluating mental health state, and offering feedback (Varnfield et al., 2019). Dyriv et al., Aigner et al., Aljaaf et al., Williams et al., and Kogilathota et al. developed software applications for patient monitoring, mental health support, and psychological state detection, which can aid in medical diagnostics (Dyriv et al., 2021; Aigner et al., 2020; Aljaaf et al., 2018; Williams and Washington, 2018; Kogilathota, 2021).

Crasto et al. created a chatbot system named CareBot, designed to provide mental health support. The system utilizes questionnaires, surveys, and data analysis methods to evaluate mental health (Crasto et al., 2021). Iyer et al. proposed a virtual mental health assistant that addresses the challenges of in-person therapy, offering features such as psychological assessments, emotion detection, chat functionality, and personalized recommendations (Iyer et al., 2021). Lush et al. introduced an interactive augmented reality system for self-assessment and mental health awareness, providing information and interventions to at-risk users (Lush et al., 2019).

While data science is being leveraged to tackle mental health concerns, the field of mental health technology continues to face obstacles and opportunities for further exploration.

3. Cognitive Distortions Dataset

The collected and annotated dataset is in English and will be available on OSF (for Review, 2023a). It builds on a previous dataset by one of the authors ().To the best of our knowledge, it is the first publicly available dataset of labeled cognitive distortions. The data annotation was done in cooperation with a licensed psychologist to label unlabeled sentences and verify the already labeled ones. The dataset extends a much smaller dataset (2409 sentences including non-distorted entries) of two cognitive distortions ("Over-Generalization" and "Should"). Data was collected from three different sources: social media platforms, crowd-sourcing platforms, and distributed surveys. The total number of collected sentences is 34370, identifying 14 different cognitive distortions as well as neutral, non-distorted text.

3.1. Data Collection

We have relied on multiple sources in order to collect a versatile, diverse dataset: 1) crowd-sourcing, 2) conducting a survey, and 3) publicly available Facebook and Twitter datasets.

3.1.1. Existing Social Media Datasets

We made use of the following two existing emotion datasets resulting from Facebook and Twitter.

1. *Facebook Empathetic Data:* The empathetic dialogues dataset and the baseline for empathic dialogue generation (Rashkin et al., 2018) consists of 18k interactions that are based on emotional circumstances. The entries are labeled based on emotions.

2. *Twitter:* The Sentiment Analysis dataset (M, 2023) contains 2477 tweets classified based on positive and negative sentiment. Each row is marked as 1 for positive sentiment and 0 for negative sentiment.

As both datasets were labeled for sentiments and emotions, respectively, we had to label them for cognitive distortions ourselves. This was done with the help of a licensed psychologist, as will be described in more detail in Subsection 3.3

3.1.2. Crowd-Sourcing

A broad summary of cognitive distortion and its definition, which was presented as a paragraph at the outset for clarity, were included in the survey along with the questions. The two main questions were if the subject had ever encountered certain thought patterns or distorted thinking, and if so, to recollect and explain those events and how they made them feel.

We used Amazon Mechanical Turk (MTurk), which is a crowd-sourcing marketplace, to assign workers to different tasks. They were first presented with a short description of the types of thinking that the particular distortion exemplifies. Then, they were asked to tell us about an actual, specific time from their own life in which they exhibited the same type of described thinking. Workers were encouraged to provide approximately 2-3 sentences, but no limit was enforced. The total number of sentences collected was 9033.

3.2. Data Collection Ethics

We followed the proposed ethics guidelines in collecting, handling, and storing the data (Fadda et al., 2022; Grover et al., 2020). As the only aim was dataset collection, we were only collecting sentences without any context, so the possible breaches were limited. We maintained anonymity and confidentiality by anonymizing all the data entries. No personal information was collected in all three data collection sources. For data collected using surveys and MTurk, the participants' consent to reuse the data was acquired. The purpose of the study and the data collection were clearly explained and outlined in all calls, so participants knew what they were signing up for.

3.3. Data Labeling and Label Verification

As shown above, the collected dataset is heterogeneous in nature. Some sentences are already collected in a labeled manner, while others are collected as is and need to be labeled. Accordingly, we distinguish between two tasks: (t1) labeling a sentence and (t2) verifying the label of already labeled sentences.

In both (t1) and (t2), the labeling consists of two phases: (l1) distinguishing between distorted and non-distorted sentences and (l2) identifying which of the 14 cognitive distortions the sentence exhibits in case it is showing a distortion.

Two annotators/verifiers were involved in the process: a certified psychologist and one of the paper's authors with a psychology background.

For (t1), each collected sentence was reviewed for relevance, and the first annotator provided a label. The labeled sentences are then verified by one verifier. In this task, there were only a handful of cases where the annotator and verifier's labels were not aligned. These cases were resolved through a discussion, and a final verdict was reached.

For (t2), each collected labeled sentence was reviewed for relevance, and the provided label resulting from the data collection phase was verified by one verifier. The involved annotators are one of the authors with a psychology background and the certified psychologist. In case the verifier's label did not align with the collected label, the second verifier was involved in resolving the label. This happened more frequently in the (l2) labeling phase. In the very rare case where the two annotators could not reach a consensus, a psychology expert professor was consulted.

3.4. Data Statistics

The dataset contains text passages labeled into one of 14 cognitive distortions such as ["Mental filtering ", "Polarization", "Overgeneralization", "Catastrophizing", "Personalization", "Control fallacies ", "Fallacy of Fairness", "Blaming", "Shoulds", "Emotional Reasoning ", "Fallacy of Change", "Global Labeling", "Always Being Right", "Jumping to Conclusions "]. A summarization of the dataset is provided in Table 1. These dataset statistics are summarized in Fig. 1.

Table 1: Dataset Collection Resources

Source	Number of Sentences
Facebook Empathetic Data	18509
Twitter Data	2477
Crowd-Sourcing	13384

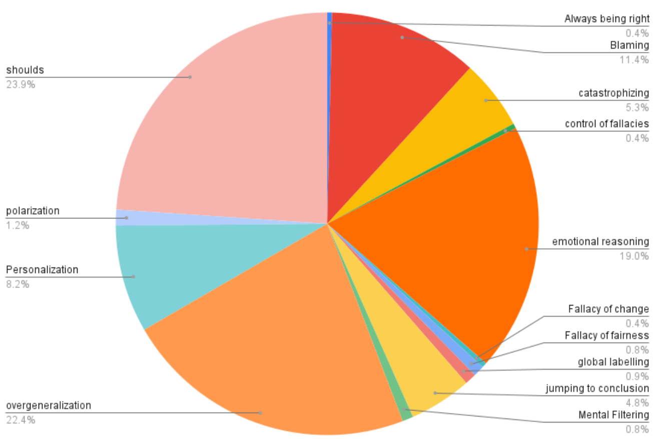

Figure 1: Percentage of sentences per cognitive distortion.

3.5. Data Pre-processing

We used Scikit-learn as it offers a single interface for all conversion stages and the final output, making it the most practical method of building a model. To enable better classification of the sentences, we applied the customary data preparation steps, including converting all text to lowercase, removing special characters, numerals, patterns, and punctuation, and removing the stopwords. Stopwords are words in any language that contribute little meaning to a phrase and can safely be ignored without sacrificing the meaning of the sentence. Additionally, we performed stemming on the words using the Porter Stemmer in the NLTK package.

4. Experiments and Results

Natural language processing (NLP) has recently gained much attention for computationally representing and analyzing human language. It has spread its applications in various fields such as machine translation, email spam detection, information extraction, summarization, medical, and question answering. In this paper, we used different state-of-the-art techniques to create several models for language classification. In our created models, vectorizers were used for the machine learning models, which utilized GloVe, BERT, and Flair word embeddings. Word embeddings are projections of words into a continuous space with the goal of maintaining their semantic and grammatical similarity. Word embeddings highly enhance the performance of the classification models. Classical embedding reflects words into a vector space where each word in vocabulary has only one vector representation. This vector representation stays the same no matter the word's context. This static vector has two drawbacks. Firstly, it fails to model polysemous words. The second issue is that it cannot predict word embedding for out-of-vocabulary words. Contextual word embedding, on the other hand, solves the drawbacks of classical embedding. It can compute a vector representation of a word depending on its context. Furthermore, it has a stronger word representation that carries more meaning than classical word embedding.

The first baseline is implemented using Naive Bayes with a count vectorizer. The second baseline is created using SVM architecture with TF-IDF.

The main models were implemented using Convolutional Neural Network, LSTM, DNN, RNN, with different embedding combinations. All our created models were tested with different vectorizers to collect features from different n-gram-ranges. We then split the data into a training set (80%) and a test set (20%), setting a consistent integer for the random state to ensure the separation of sets. Model hyperparameters were adjusted, and deep learning models such as CNN and LSTM were used with 100 or 300 GloVe or BERT embeddings. Since pretrained word embedding was used, training the embedding layer was avoided to prevent the overfitting of models. We set the random state to a constant and split the data into 80% for training and 20% for testing for every model. The CNN architecture (Fig. 2) includes three max-pooled convolution layers with 512 filters and 3-5 filter windows, followed by concatenated and flattened layers, a dropout layer, and a fully connected layer.

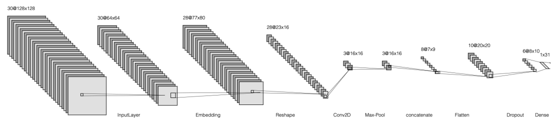

Figure 2: CNN Diagram

The LSTM model architecture is illustrated in Fig. 3, To prevent the model from being overfit, a spatial dropout layer with a drop rate of 0.20 is inserted. This layer functions by discarding inputs at random to limit the network's size and the likelihood of interdependent learning between neurons. In addition, the LSTM layer has a dropout rate of 0.2, which subsamples the random outputs entering the fully connected layer as shown in Fig. **??**

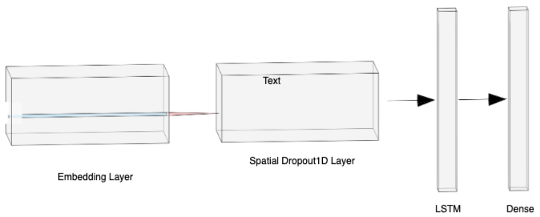

Figure 3: LSTM Diagram

4.1. Results

The data used to assess the models consisted of 28,250 training sentences and 7,062 testing sentences. In all models, the same testing data was used. Our models' performance was tested using accuracy, precision, recall, and F1-score (F1). Table 2 shows all the different results using different metrics. We can see that pre-trained BERT and Glove-300 showed significantly better results than other embedding words. BERT embeddings with the CNN model achieved the best accuracy, which is 85.64. The LSTM model with GLOVE-300 achieved an accuracy of 80.4 %. We start the hyper-parameters tuning process for the CNN model by changing the batch sizes between 5, 10, 15, and 20. We also changed the activation function between Softmax, Softplus, Softsign, and Relu. For the optimization functions, the list of functions we tested Adam, Adamax, Nadam, Adagard, RMSProp, and Hard-sigmoid. The CNN outperformed model included the number of epochs, batch size, activation functions, and optimization functions. We manually adjusted one hyperparameter at a time to find the best result, then fine-tuned all other hyperparameters. It shows the best results with batch size = 300

```
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            [(None, 300)]        0           []

embedding (Embedding)           (None, 300, 300)     2982900     ['input_1[0][0]']

reshape (Reshape)               (None, 300, 300, 1)  0           ['embedding[0][0]']

conv2d (Conv2D)                 (None, 298, 1, 512)  461312      ['reshape[0][0]']

conv2d_1 (Conv2D)               (None, 297, 1, 512)  614912      ['reshape[0][0]']

conv2d_2 (Conv2D)               (None, 296, 1, 512)  768512      ['reshape[0][0]']

max_pooling2d (MaxPooling2D)    (None, 1, 1, 512)    0           ['conv2d[0][0]']

max_pooling2d_1 (MaxPooling2D)  (None, 1, 1, 512)    0           ['conv2d_1[0][0]']

max_pooling2d_2 (MaxPooling2D)  (None, 1, 1, 512)    0           ['conv2d_2[0][0]']

concatenate (Concatenate)       (None, 3, 1, 512)    0           ['max_pooling2d[0][0]',
                                                                  'max_pooling2d_1[0][0]',
                                                                  'max_pooling2d_2[0][0]']

flatten (Flatten)               (None, 1536)         0           ['concatenate[0][0]']

dropout (Dropout)               (None, 1536)         0           ['flatten[0][0]']

dense (Dense)                   (None, 15)           23055       ['dropout[0][0]']

==================================================================================================
```

Figure 4: CNN Architecture

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 100, 300)          5187600

 spatial_dropout1d (SpatialD  (None, 100, 300)         0
 ropout1D)

 lstm (LSTM)                 (None, 100)               160400

 dense (Dense)               (None, 15)                1515

=================================================================
Total params: 5,349,515
Trainable params: 161,915
Non-trainable params: 5,187,600
_________________________________________________________________
```

Figure 5: LSTM Architecture

epochs =20, and the activation function is 'softmax'. An overview of the CNN architecture is shown in Fig.4.

The outperforming LSTM model included the number of epochs, batch size, activation functions, and optimization functions. We manually adjusted one hyperparameter at a time to find the best result, then fine-tuned all other hyperparameters. It shows the best results with batch size = 20 epochs=15, the activation function is 'softmax', and the optimizer is 'adam'.

The best-performing DNN model was produced by utilizing the BERT to give an accuracy of 78.13 % While SVM and RNN models were utilized using the TF-IDF vectorizer, RNN performed a little better than the SVM, meanwhile, Word-Embeddings with RNN showed less accuracy. While Naive Bayes models performed almost equal as RNN with Word-Embeddings using count vectorizer.

5. The C-Journal Mobile Application

The main aim of the work is to provide an aiding tool for individuals to track their own thinking patterns and for life coaches to better track and help large groups of patients. The collected datasets and cognitive distortion classification models are used as a basis for developing the C-Journal app. C-Journal is deployed on both the Apple App Store (for Review, 2023b) and the Android Playstore (for Review, 2023c). C-Journal serves as a day-to-day

Table 2: Best Performing Models

Model	Vectorizer	Accuracy	Recall	Precision	F1 Score
CNN	Bert	85.64 %	77.1 %	89 %	**84** %
Lstm	Glove 300	80.4 %	72.1 %	86.4 %	75.12 %
DNN	Bert	78.13 %	69.1 %	83.7 %	77.54 %
RNN	TF-IDF	69.77 %	61.06 %	78.2 %	68.3
SVM	TF-IDF	65 %	97.2 %	67.2 %	79.3
RNN	Word-Embeddings	61.39 %	61.06 %	61.38 %	62.3
Naive bayes	count vectorizer	61 %	87.7 %	70 %	78.4

app to track thinking patterns, improvements, and changes over time. Our thoughts and feelings are recorded in journals. Keeping a journal is a way to document daily life. Discovering our triggers for happiness and/or upset.

C-Journal was implemented in three phases: research, design, and assessment. A basic literature review followed by a brainstorming session was done. Relevant state-of-the-art applications were found, assessed, and categorized. This phase produced a thorough concept for a user-friendly mobile app that will identify and categorize distorted thought patterns in journal entries. The mobile app would be a vital data source, designed to gather data directly from the user, making data collecting easier. Because the system interacts directly with the user, we made certain that the user interface (UI) is clear and straightforward, has a modern style, and is fully featured and simple to use on mobile clients. The suggested architecture uses the user-friendly interface design paradigm of human-computer interaction. This framework's primary goals are to promote and enhance psychological care, reduce potentially harmful mental influences, and provide immediate feedback to notify people and assist them in monitoring their mental states. For the application, privacy and security considerations become more critical. Again, suggested guidelines proposed in (Fadda et al., 2022; Grover et al., 2020) were followed. In the future, we aim to improve this part further by incorporating stronger privacy preservation techniques and suggestions in (Kaplan et al., 2023).

5.1. C-Journal Features

The initial brainstorming session with a focus group yielded the expected functional requirements of the app. The requirements and the associated affected actors that were derived as a result of the participant feedback are shown in Table 3.

The **Journaling Entries Creation** feature includes a free-text input section that users are encouraged to use to record their perceived feelings, activities, experiences, and interactions with other people from the previous day, which forms the basis of the journaling process. It is also feasible to use images in journal entries to create a more vivid memory of that specific day. Journaling is usually a daily activity, however, there are no guidelines forcing users to add daily entries.

Table 3: Functional Requirements

Req. #	Description
R01	sign up with mail and password
R02	sign in with mail and password
R03	Create a new journal entry for today
R04	Create a new journal entry for any previous date
R05	add a specific journal entry to favorite
R06	view all journal entries
R07	view a specific journal entry
R08	Record mood data
R09	view mood data
R10	view the detected distorted thought
R11	view all Statistics

The **Mood Tracking** feature prompts the user to record their mood at regular intervals, which is supposed to be daily, the same as the journaling entry. This is meant to pinpoint patterns in how moods change over time and in response to various events. When the user writes down his journal, the application pops up a question, "how do you feel?" with a slider that uses a scale to evaluate their mood between bad, mixed, alright, and good. For an overall evaluation, an average of all entered data is generated and shown in the report section.

The purpose of maintaining a journal is to assist in reinforcing self-awareness while reviewing previous entries and to foster self-reflection. Users are encouraged to record both happy and bad incidents so that they may reflect on and learn from their prior experiences. All the journaling texts can be viewed later in a timeline and can be added to the favorite section.

The **Detecting Potential Cognitive Distortions** feature is the main app feature, which gives users instant feedback and a detailed report about their saved journaling texts. The user can view whether there's a detected potentially distorted thought or not.

The **General Report** feature of C-Journal also provides the user with detailed feedback about their journaling history and the average mood for the month. Finally, the app highlights recurring distorted thinking patterns.

5.2. C-Journal Technical Details

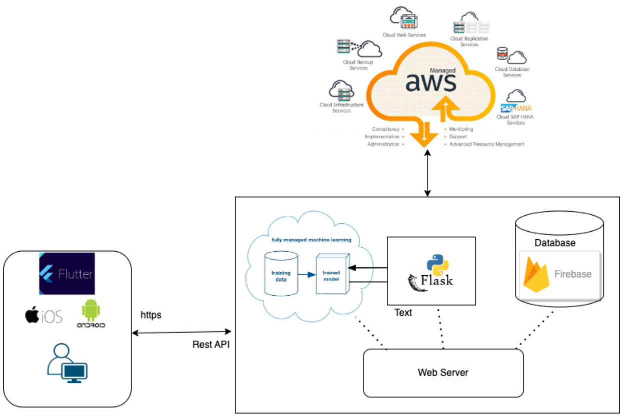

Figure 6: Mobile Application Workflow

The architecture of C-Journal is shown in Fig. 6. The mobile application, developed using Flutter, allows data tracking and is compatible with both iOS and Android phones. The data is transmitted to the backend server, where the Cognitive Distortions Detection model processes it. The backend then delivers real-time prediction results to the application. The backend server is created with Python using the Flask web development library. The server receives input data from the application through the /post route. It loads the saved Convolutional Neural Network TensorFlow Model, parses the received JSON object parameters, and processes the data to generate a prediction class. Additionally, the server sends a detailed report with various metrics. The gathered dataset is stored in Firebase. Amazon Web Services (AWS) is used for infrastructure purposes. The application utilizes a T3 microprocessor from Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3) to store the machine learning model, data, and scripts. Amazon CloudFront is used to ensure a quick and smooth process. The deployment script, based on Infrastructure as a Service (IaaS), sets up the EC2 machine, ensures the availability of the S3 bucket, and connects CloudFront to the server. The building stage involves several steps, including AWS connection setups, server code deployment, and Ansible utilization to upload code and the model to the S3 bucket, install dependencies, and run the code. Smoke tests are performed to validate the expected response and ensure everything is functioning properly. If the tests fail, past builds are deleted. In summary, C-Journal's architecture involves a Flutter-based mobile application for data tracking, a Python-based Flask backend server for processing and prediction, and AWS for infrastructure and storage purposes. The deployment process includes setting up the necessary components and conducting smoke tests for validation.

5.3. Initial Testing

A small pilot study to evaluate the usability of C-Journal was conducted on a small sample of 10 individuals. The results were used to update the prototype and remove the non-usable features. The recommendations included adding daily reminder prompts, which 9 out of the 10 users requested. The users were given a short version of the System Usability Scale (Brooke et al., 1996), which showed the following: All users found the application highly usable and the features easy to use. 8 out of the 10 said they would recommend it to others and would use it on a daily basis. In the future, more rigorous testing of the usability and design features of the application needs to be conducted.

6. Conclusion and Future Work

The paper presents the first open-source labeled corpus for detecting and classifying cognitive distortions, a baseline model (Naive Bayes and SVM), and primary models (CNN, LSTM, DNN, RNN), with different embeddings (Bert, Glove 300, TF-IDF). Sentences with cognitive distortions are classified with an accuracy of 84%. The best-performing model is used as a basis for C-Journal, a mobile app for aiding users to identify their own thinking patterns by highlighting potentially distorted thinking patterns in private journaling texts.

This work can be extended from the perspective of the dataset, cognitive distortion detection models, and C-Journal app, respectively, as follows:

- The dataset is constantly being extended with more sentences. In the future, we aim to include more languages in the dataset to increase its usability. This is a highly social task that would require employing more cooperative and crowd-sourcing techniques.

- The implemented classification models should be further enhanced to achieve better accuracy. We are currently comparing different autoencoders such RoBerta, BioClinical_Bert shown in (Guo et al., 2022).

- The C-Journal app is to be tested in a real-life scenario through a longitudinal study. One application mode to be further worked out in the future is a networking mode where life coaches and psychiatrists can track their patients' input and interact with them via the app. The app can then help notify caregivers in case of the occurrence of certain distorted patterns.

7. Bibliographical References

Christoph Aigner, Matthias Eder, René Baranyi, and Thomas Grechenig. 2020. recoverapp - a mobile health solution to support people in stationary rehabilitation. In *2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–6.

Ahmed J. Aljaaf, Dhiya Al-Jumeily, Thaaer kh. Asman, Abir J. Hussain, Thar Baker, Mohamed Alloghani, and Jamila Mustafina. 2018. H-diary: Mobile application for headache diary and remote patient monitoring. In *2018 11th International Conference on Developments in eSystems Engineering (DeSE)*, pages 18–22.

Nakul Amate, Sagar Patil, Pranav Jojan, and Sahil Morankar. 2021. Use of social media and smartwatch data analytics for mental health diagnosis. In *2021 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6.

Haizhen An, Xiaoyong Lu, Daimin Shi, Jingyi Yuan, Renjun Li, and Tao Pan. 2019. Mental health detection from speech signal: A convolution neural networks approach. In *2019 International Joint Conference on Information, Media and Engineering (IJCIME)*, pages 436–439.

Kazi Zawad Arefin, Kazi Shafiul Alam Shuvo, Masud Rabbani, Peter Dobbs, Leah Jepson, Amy Leventhal, Amy Vaughan Van Heeke, and Sheikh Iqbal Ahamed. 2021. Towards developing an emr in mental health care for children's mental health development among the underserved communities in usa. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 688–693.

Sujal BH, Neelima K, Deepanjali C, Bhuvanashree P, Kavitha Duraipandian, Sharanya Rajan, and Mithileysh Sathiyanarayanan. 2022. Mental health analysis of employees using machine learning techniques. In *2022 14th International Conference on COMmunication Systems NETworkS (COMSNETS)*, pages 1–6.

John Brooke et al. 1996. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Chun-Hao Chang, Elvis Saravia, and Yi-Shin Chen. 2016. Subconscious crowdsourcing: A feasible data collection mechanism for mental disorder detection on social media. pages 374–379.

Reuben Crasto, Lance Dias, Dominic Miranda, and Deepali Kayande. 2021. Carebot: A mental health chatbot. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–5.

Andrés Cárcamo, Franco Kemper, and David Mauricio. 2021. Mondep app: Monitoring patients with depression using sentiment analysis of therapeutic diary entries. In *2021 IEEE Sciences and Humanities International Research Conference (SHIRCON)*, pages 1–4.

Alberto de Santos Sierra, Carmen Sánchez Ávila, Gonzalo Bailador del Pozo, and Javier Guerra Casanova. 2011. Stress detection by means of stress physiological template. In *2011 Third World Congress on Nature and Biologically Inspired Computing*, pages 131–136.

Mandar Deshpande and Vignesh Rao. 2017. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 858–862.

Andrii Dyriv, Vasyl Andrunyk, Yevhen Burov, Ihor Karpov, and Lyubomyr Chyrun. 2021. The user's psychological state identification based on big data analysis for person's electronic diary. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 1, pages 101–112.

Marta Fadda, Martin Sykora, Suzanne Elayan, Milo A Puhan, John A Naslund, Stephen J Mooney, Emiliano Albanese, Rosalba Morese, and Oliver Gruebner. 2022. Ethical issues of collecting, storing, and analyzing geo-referenced tweets for mental health research. *Digital Health*, 8:20552076221092539.

Bilnd for Review. 2023a. Cognitive distortions dataset.

Blind for Review. 2023b. Application on appstore link.

Blind for Review. 2023c. Application on playstore link.

Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. 2015. Toward a new definition of mental health. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, 14:231.

Jiahui Gao and Guopu Shi. 2021. Mental health evaluation of college students based on similar trajectory clustering algorithm. In *2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 828–831.

Sandeep Grover, Siddharth Sarkar, and Rahul Gupta. 2020. Data handling for e-mental health professionals. *Indian journal of psychological medicine*, 42(5_suppl):85S–91S.

Yuting Guo, Yao Ge, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker. 2022. Comparison of pretraining models and strategies for health-related social media text classification. In *Healthcare*, volume 10, page 1478. MDPI.

Sreevidya Iyer, Dhanashree Shetty, Purva Badgujar, Anjali Nambiar, and Kajal Jewani. 2021. A proposal for virtual mental health assistant. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 953–957.

Manav Pradeep Jain, Soumyaprakash Sribash Dasmohapatra, and Stevina Correia. 2020. Mental health state detection using open cv and sentimental analysis. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 465–470.

Brent A Kaplan, Shawn P Gilroy, W Brady DeHart, Jeremiah M Brown, and Mikahil N Koffarnus. 2023. Data handling: ethical principles, guidelines, and recommended practices. In *Research Ethics in Behavior Analysis*, pages 191–214. Elsevier.

Rahul Katarya and Saurav Maan. 2020. Predicting mental health disorders using machine learning for employees in technical and non-technical companies. In *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)*, pages 1–5.

Mohamed Khazbak, Zeyad Wael, Zahwa Ehab, Maria Gerorge, and Essam Eliwa. 2021. Mindtime: Deep learning approach for borderline personality disorder detection. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 337–344.

Vivek Kogilathota. 2021. iponder: A novel approach to multimodal mental health diagnosis and therapy for teens via deep transfer learning and computer vision. In *2021 IEEE International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 20–24.

Daniel Laden, Shyaman Jayasundara, and Indika Kahanda. 2020. Rdocer: Information retrieval and sentence extraction for mental health using research domain criteria. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 226–229.

Mark E. Larsen, Tjeerd W. Boonstra, Philip J. Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. 2015. We feel: Mapping emotion on twitter. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1246–1252.

Tahira Lashari, Ebaad Amin, Sana Lashari, Murtaja Ali, and Saima Lashari. 2021. Development of a web portal 'ikigai' to assess the psychological well-being of university students.

Jing Li. 2021. Optimal modeling of college students' mental health based on brain-computer interface and imaging sensing. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 772–775.

Na Li, Haiting Zhai, Tamizharasi Seetharam, and Shanthini Arumugam. 2021. Psychological health analysis based on fuzzy assisted neural network model for sports person. *Journal of Intelligent Fuzzy Systems*, pages 1–15.

Jinping Liu, Fang Xia, Yanyin Cui, and Zixu Hao. 2021. The promotion effect of innovation and entrepreneurship education on medical students' mental health based on stepwise regression. In *2021 International Conference on Public Health and Data Science (ICPHDS)*, pages 33–37.

Jianhong Luo, Jingcheng Du, Cui Tao, Hua Xu, and Yaoyun Zhang. 2018. Exploring temporal patterns of suicidal behavior on twitter. In *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pages 55–56.

Victoria Lush, Christopher Buckingham, Stephen Wileman, Suzanne Edwards, and Ulysses Bernardet. 2019. Augmented reality for accessible digital mental healthcare. In *2019 5th Experiment International Conference (exp.at'19)*, pages 274–275.

S H I N I G A M. 2023. Twitter kaggle dataset.

Shima Ramesh Maniyath, Thrupthi S M, Uma M C, Thrupthi K R, and Apeksha Vasanth L. 2021. Evaluation of mental health using iot based wearables. In *2021 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, pages 239–242.

Kyounghyun Park, Min Jung Kim, Jungsook Kim, Oh Cheon Kwon, Daesub Yoon, and HyunSuk Kim. 2020. Requirements and design of mental health system for stress management of knowledge workers. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1829–1832.

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. 2007. No health without mental health.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy. *CoRR*, abs/1811.00207.

Katerina Rnic, David Dozois, and Rod Martin. 2016. Cognitive distortions, humor styles, and depression. *Europe's Journal of Psychology*, 12:348–362.

Romil Roy, Rajlakshmi Guha, Sangeeta Das Bhattacharya, and Jayanta Mukhopadhyay. 2017. Building a web based cognitive restructuring program for promoting resilience in a college campus. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, pages 520–524.

Sudha Tushara Sadasivuni and Yanqing Zhang. 2020. Using gradient methods to predict twitter users' mental health with both covid-19 growth patterns and tweets. In *2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI)*, pages 65–66.

Budhaditya Saha, Thin Nguyen, Dinh Phung, and Svetha Venkatesh. 2016. A framework for classifying online mental health-related communities with an interest in depression. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1008–1015.

Benjamin Shickel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020a. Automated emotional valence prediction in mental health text via deep transfer learning. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 269–274.

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

Feng-Tso Sun, Kuo Cynthia, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. 2012. Activity-aware mental stress detection using physiological sensors. volume 23.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. Psychbert: A mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082.

Marlien Varnfield, Kausthubram Rajesh, Christian Redd, Simon Gibson, Lisa Gwillim, and Stuart Polkinghorne. 2019. Health-e minds: a participatory personalised and gamified mhealth platform to support healthy living behaviours for people with mental illness. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6943–6947.

Ruyi Wang, Jiankun Wang, Yuan Liao, and Jinyu Wang. 2020. Supervised machine learning chatbots for perinatal mental healthcare. In *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, pages 378–383.

Lucretia Williams and Gloria Washington. 2018. Soul glow: An application for helping to understand and improve mental health care of hbcu students. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 358–360.

Akkapon Wongkoblap, Miguel A. Vadillo, and Vasa Curcin. 2017. Detecting and treating mental illness on social networks. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 330–330.

MengLi Zhu and ZhiFang He. 2021. Design and implementation of measurement system based on mental health education. In *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*, pages 680–683.