Cognify: A Modular Privacy-Conscious AI-Driven Mobile App for Mental Health Based on Cognitive Distortion Detection

Mariam Dawoud[1], Mohamad Rasmy[2] and Alia El Bolock[1]
[1]*Department of Computer Science and Engineering, The American University in Cairo, Cairo, Egypt*
[2]*Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt*

Keywords: Mental Health, Mobile App, Learning Model, Security, Cognitive Distortions.

Abstract: Cognitive distortions—irrational thought patterns contributing to emotional distress—are central to cognitive behavioral therapy (CBT), but early detection often depends on clinical assessments, limiting opportunities for timely self-reflection. Cognify is a cross-platform mobile application designed to help users detect and understand these distortions by analyzing daily journal entries using a fine-tuned NLP model that classifies entries into 14 cognitive distortion types. The app offers real-time feedback, weekly summaries highlighting recurring patterns, and an intuitive interface that promotes ongoing engagement. This paper presents Cognify's system architecture, AI model integration, and results from a pilot study, which demonstrated improved user awareness of cognitive patterns, high user satisfaction, and increased journaling consistency over time. The app's modular design also allows for optional integration of privacy-preserving features, ensuring flexibility to address evolving user needs. By combining AI-driven distortion detection with an adaptable journaling experience, Cognify offers a practical and engaging tool for enhancing cognitive awareness and supporting personal growth.

1 INTRODUCTION

Mental health applications have gained significant traction in recent years, providing accessible and effective support for users managing many psychological disorders and maintaining their emotional well-being. Due to the increased awareness on mental health, it is important to be able to detect early symptoms of mental disorders called cognitive distortions, which are thoughts that cause inaccurate perceptions of reality due to exaggerations or irrationality (Beck, 2022). These distortions can be analyzed on the basis of 15 characteristics(including neutral) to help guide mental health professionals (psychiatrists, psychologists, therapists, and others) to susceptible disorders. Most existing applications either lack a robust AI-based cognitive distortion detecting mechanism or fail to implement a privacy-preserving mechanism, causing drawbacks in participation due to concerns on trust and integrity. We propose a comprehensive approach by addressing challenges through a user-friendly cross-platform mobile application that can accurately detect cognitive distortions through an AI robust model using RoBERTa, as well as maintain privacy of sensitive data through a pluggable framework of integrated techniques. The user will be logging in their journal entry each day, and each entry will be classified with the relevant cognitive distortion, then the data will be stored safely on the database for retrieval when needed. We aim to ensure users are comfortable using the application and providing their journal data while trusting the application to provide meaningful results based on insights from their data.

At the core of the application is a robust AI-driven cognitive distortion detection module, which analyzes users' daily journal entries using a fine-tuned RoBERTa model. This AI model automatically classifies entries into one or more of 15 recognized cognitive distortion categories, providing users with clear, actionable feedback to help them recognize problematic thought patterns over time. The application is designed with modularity in mind, ensuring that the AI model, user interface, and data handling components are decoupled to allow for easy updates and customization. This modular design also leaves room for incorporating privacy-preserving technologies if needed, particularly to address user concerns when handling sensitive mental health data. By providing a user-friendly journaling interface combined with intelligent feedback, the application empowers users to

Dawoud, M., Rasmy, M. and El Bolock, A.
Cognify: A Modular Privacy-Conscious AI-Driven Mobile App for Mental Health Based on Cognitive Distortion Detection.
DOI: 10.5220/0013571400003964
In *Proceedings of the 20th International Conference on Software Technologies (ICSOFT 2025)*, pages 361-368
ISBN: 978-989-758-757-3; ISSN: 2184-2833

become more aware of their thinking patterns while offering valuable insights for both personal reflection and professional therapeutic support.

The paper presents the following contributions:

- develop a full-stack cross-platform journaling mobile application to securely detect cognitive distortions.
- construct a machine learning model for detecting cognitive distortions and their types.
- implement a privacy-preserving framework integrating ECC and Blockchain latest technologies.
- conduct a pilot study to ensure the usability of the app from the user's perspective.

By combining a powerful AI analysis engine with an intuitive and adaptable mobile interface, the application offers a comprehensive, flexible tool for cognitive distortion awareness and self-reflection, adaptable to both individual use and potential integration with therapeutic interventions.

The rest of the paper is designed as followed: Section 2 discusses existing literature in the fields of mental health mobile apps, AI detection models, and privacy-preserving algorithms. Section 3 discusses the system architecture and how the application flow works. Section 4 discusses the methodology in more detail, focusing on the technical implementations and evaluation metrics. Section 5 discusses results and what they infer in the field, followed by Section 6 to suggest future work based on the provided results.

2 RELATED WORK

This section discusses existing literature for mobile mental health applications specifically to survey available methodologies and identify gaps in research. We also look at AI detection models and briefly discuss the possibility of privacy preserving techniques.

2.1 Mobile Mental Health Applications

Awareness of mental health disorders has increased significantly in recent years, driving the development of mobile applications designed to support users through features like mood tracking, guided exercises, and assessments. A narrative review categorized mental health apps into distinct types such as mood trackers, mindfulness-based apps, self-care apps, and treatment apps. This categorization highlighted the need for mental health apps in widespread disorders such as anxiety, depression, and suicidal ideations, and the need for incorporation with psychotherapy for enhanced outcomes and interventions. Blending these applications with Cognitive Behavioral Therapy (CBT) techniques can be very efficient, but challenges such as low user engagement, inconsistent evaluation methodologies, and privacy concerns are emphasized. Moreover, the lack of standardized assessments and fragmentation issues of these apps limit the research in deducing a reliable conclusion (Diano et al., 2022).

Mindful Meadows is a mobile mental health app that utilizes self-assessments, mood tracking, music, yoga, a chatbots and therapist booking all in one framework to offer comprehensive support. However, the system raises some concerns as it lacks clear clinical validations and weak privacy protections (Gaikwad et al., 2024). Another app called Mindset provides similar comprehensive support through journaling, mood tracking, guided breathing, and anonymous peer support. Results and reviews prove its convenience in use, although no evaluation has been done on the long-term health benefits (Samuel and Shirley, 2023). MindWell Solace is a chatbot-driven application targeting student mental health that uses a classifier to detect disorders and inform counselors if needed. The model achieves a good accuracy for some symptoms but is depth-limited due to the nature of yes/no questions and symptom self-reporting (Bhave et al., 2024). Finally, an app developed during the COVID-19 pandemic offered remote counseling with mental health professionals, integrating their records and journal entries. While it proved effective for remote access, it is relatively limited in its scalability (Krisnanik et al., 2020).

2.2 Cognitive Distortion Detection Models

In the technical scope, detecting cognitive distortions has been done through gaming; ARCod is an augmented reality (AR) interactive game to assess cognitive distortions according to its 5 different levels (Tasnim and Eishita, 2022). Detection was also implemented using a machine learning-based text analysis of online blogs and journal entries, similar to what is implemented in this app (Shickel et al., 2020a).

The detection of cognitive distortions has been approached as a text classification problem. A major challenge in this task is obtaining a sufficiently large labeled dataset to train deep learning models effectively. The literature contains extensive efforts to utilize deep learning for cognitive distortion detection. Several works have focused on collecting and annotating datasets for this purpose, as well as training machine learning classification models on these datasets. Most of these datasets are in English, including those

compiled by (Shickel et al., 2020b; Elsharawi and El Bolock, 2024; Lim et al., 2024; Shreevastava and Foltz, 2021). Additionally, a few datasets have been developed in Chinese, such as those by (Wang et al., 2023; Qi et al., 2023; Na, 2024).

Our work specifically builds upon the foundational dataset introduced by (Mostafa et al., 2021), which was the first publicly available cognitive distortion detection dataset. This English-language dataset comprises 2,409 text entries categorized into two types of cognitive distortions (overgeneralization and should statements) along with non-distorted texts. The authors evaluated several machine learning and deep learning models employing pre-trained embeddings, ultimately identifying a tuned Long Short-Term Memory (LSTM) network with 300-dimensional GloVe embeddings as the most effective model.

Subsequently, (Elsharawi and El Bolock, 2024) expanded this dataset, creating the largest open-source collection available to date. This extended dataset contains 34,370 sentences annotated across 14 cognitive distortion categories, along with neutral (non-distorted) examples. Their highest-performing model was a Convolutional Neural Network (CNN) employing pre-trained BERT embeddings. This dataset was sourced from existing social media emotion datasets, annotated by one of the authors with a psychology background, and subsequently verified by a certified psychologist.

Given the limited availability of high-quality annotated datasets, recent studies have explored artificial dataset augmentation techniques. Notably, (Rasmy et al., 2024) proposed four data augmentation methods specifically to expand the dataset introduced by (Elsharawi and El Bolock, 2024), thereby enhancing training set size and improving model performance. They demonstrated an improvement of up to 5.9% in F1-score using a fine-tuned RoBERTa model trained on the augmented dataset, compared to the previously identified CNN-BERT model evaluated on the original non-augmented dataset. In our framework, we integrate this best-performing fine-tuned RoBERTa model trained on the augmented dataset.

2.3 Privacy Preservation in Mobile Apps

Privacy in mobile health apps remains a major concern due to excessive data collection, unclear policies, and security vulnerabilities that undermine user trust. Lightweight, transparent privacy frameworks are needed for sensitive health data. Privacy techniques in mental health mobile applications vary widely, from basic encryption methods like AES and RSA to more advanced approaches such as Homomorphic Encryption, differential privacy, and secure cloud storage frameworks (Vichare et al., 2017; Zhou et al., 2018; Inakollu et al., 2024; Latif et al., 2020). While these methods offer varying levels of data protection, many suffer from high computational costs, performance delays, reliance on external auditors, or reduced data accuracy, making them difficult to implement seamlessly in real-world mobile apps (Suguna and Shalinie, 2017; Vijay Sai et al., 2024; Whaiduzzaman et al., 2020). Although these techniques are not the core of our proposed framework, it is important to highlight them due to the highly sensitive nature of data collected in mental health applications, where strong privacy protections are critical to maintaining user trust and encouraging sustained app engagement (O'Loughlin et al., 2019; Parker et al., 2019).

3 COGNIFY OVERVIEW & FEATURES

We present the system features highlighting the use cases available on the app, then walk through the development process from reading in the user's journal entry to its storage in the database.

3.1 System Features

Cognify was designed to deliver real-time cognitive distortion feedback through a journaling experience at the frontend for an intuitive user interaction to a sophisticated backend processing and privacy maintenance. The app is a feature-rich, cross-platform mobile application designed to enhance self-reflection and mental well-being through AI-driven cognitive distortion detection. The system provides users with an user-friendly journaling experience while leveraging advanced NLP techniques to analyze thought patterns in real time. By automatically classifying journal entries into 15 cognitive distortion categories, the app offers immediate feedback to help users better understand their thought processes. Additionally, it generates structured weekly insights, allowing users to track recurring patterns over time and gain deeper self-awareness. Seamless data storage and retrieval and possible secure data handling, through encryption and account deactivation, provide accessibility and protection of user sensitive data. By combining AI-powered analysis with an adaptable and user-centric interface, Cognify aims to bridge the gap between self-guided mental health tools and professional ther-

apeutic interventions. In the section below, we take a look at the development of the application to fulfill these features seamlessly.

3.2 Data Storing

The single string of data is then immediately encrypted locally; we generate an encryption key that is unique per user and derived using a combination of device-stored keys protected via the Flutter Secure Storage module. This ensures that if the database is ever compromised, the entries remain unreadable without the user's local key. Upon encryption, the data is stored in a hybrid model of a blockchain on Firebase; this is described further in section 4.6.2.

3.3 Generating Analytics

Once the entry is saved safely on the database, we can view a scrollable view of all previous entries, and opening any entry card reveals the collected data: the title, date and time, mood and intensity, journal entry, and the classification result. This provides users an opportunity to view and reflect back on previous entries and emotions.

When the user views their profile, they are met with a couple of analytics and summaries. An average mood display calculates the user's average mood throughout the week, updating each day and restarting at the beginning of the new week. Moreover, a summary display of the week's entries is presented, with only the title visible under each day of the week. Figure 1b shows the interface developed for the profile page described.

3.4 Privacy & Account Deactivation

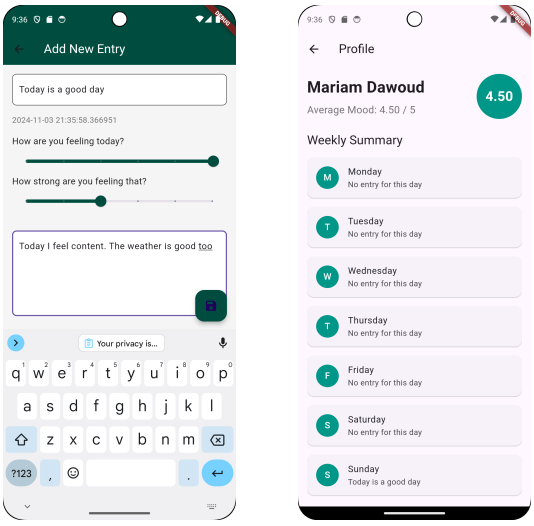
At any given time, the user can decide to deactivate or delete their account. Taking this action triggers a full data export process, permanently removing data from the database and exporting it in a ZIP archive file locally on the user's mobile device.

3.5 Application Prototyping

In the following two sections, we delve deeper into the model used and the privacy optional layer utilized.

3.5.1 Cognitive Distortion Detection Algorithm

In this study, we utilize the dataset provided by (Elsharawi and El Bolock, 2024), as it is the largest publicly available dataset in English for cognitive distortion detection. We enhance this dataset with



(a) Entry Editing Screen (b) Profile Page

Figure 1: Screenshots from the Cognify App.

augmentations provided by (Rasmy et al., 2024). The best-performing model identified by the authors for this dataset, which we adopt, is the fine-tuned RoBERTa model.

RoBERTa (Liu et al., 2019) is pre-trained on a large and diverse textual corpus, using a masked language modeling objective to predict missing tokens based on their surrounding context. This approach enables the model to develop a nuanced understanding of linguistic structures and meaning. It has demonstrated strong performance across a wide range of NLP tasks, particularly in text classification, which makes it a well-suited choice for identifying subtle linguistic patterns in cognitive distortions.

The authors in (Rasmy et al., 2024) experimented with four data augmentation techniques: synonym replacement (SR), random insertion (RI), word embedding substitution, and back-translation. Their results indicated that the highest performance was achieved using a combination of SR and RI. Consequently, we train our model using datasets augmented with these two techniques.

We follow the preprocessing, data splitting, and hyperparameter recommendations provided by the authors. The preprocessing steps include lowercasing text, removing punctuation, unrecognized symbols, and tags, and eliminating duplicate entries. The dataset is split into 70% training, 10% validation, and 20% testing. For fine-tuning, we employ the best-identified learning rate of 4e-5 and use early stopping to monitor validation loss instead of a fixed number of epochs. Additionally, we optimize the batch size to 32, as it yields the best results.

3.5.2 Privacy Preserving Algorithm

The development process of Cognify was guided by a modular framework design, ensuring that each core component—journal management, cognitive distortion detection, and data handling—can operate independently while still integrating seamlessly. This modularity allows for the optional inclusion of a privacy-preserving layer, which can be enabled or omitted depending on the sensitivity of the data being processed. For less sensitive data or during initial development phases, the app can function without encryption, maintaining flexibility for different deployment needs.

To validate the modular design and test the system's ability to handle encrypted data flows, we integrated a pluggable privacy-preserving layer using AES-based encryption. This test implementation ensures that the application can fetch, store, and process encrypted data without disrupting the journaling or cognitive distortion detection workflows. This privacy module remains open for further enhancement, allowing future contributors to replace or upgrade the encryption mechanism based on evolving privacy requirements or regulatory standards.

4 SYSTEM ARCHITECTURE

The system architecture of Cognify is designed to deliver a cross-platform mobile application that supports journaling, cognitive distortion detection, and optional privacy-preserving features. The architecture follows a modular client-server model, consisting of a mobile frontend, a cloud-hosted backend, and supporting services for storage, classification, and encryption. The mobile client, built using Flutter, ensures a unified experience across Android and iOS, simplifying maintenance and ensuring consistent functionality. Journal entries are processed by a cognitive distortion classifier deployed as a serverless function, activated only when new entries are submitted. The system is designed with modularity at its core, allowing each component—including data handling, AI processing, and privacy layers—to operate independently, enabling future upgrades or replacements without disrupting the overall system. Importantly, the privacy-preserving component, including encryption and secure storage, is implemented as a flexible, optional add-on that can be customized or replaced based on evolving privacy requirements. This architecture balances computational efficiency, platform flexibility, and extensibility to accommodate future enhancements, including integration with therapists' platforms or evolving security needs.

Figure 2: Cognify Architecture.

Figure 2 presents our proposed architecture for the Cognify app.

4.1 Development Process

The system is divided into the following key components:

- **Mobile Client:** responsible for the user interaction, journaling input, and encrypting the data.
- **Processing:** responsible for detecting the cognitive distortion characteristic; a model fine tuned on cognitive distortion datasets.
- **Data Storage:** handles encrypted data storage and retrieval.
- **Analytics Generation:** computing weekly summaries, average mood, and using entries to generate trends.
- **Deactivation & Data Export:** enables users to securely download their data before account deletion.

Figure 3 presents the basic processing layer of the Cognify system. Once the user is logged in, they can create their daily journal entry, where a simple interface is provided for a distraction-free entry editing, as seen in Figure 1a. The user titles the entry, records their mood on a scale from 1 to 5 (1 being not well and 5 being very well) and how intense are they feeling that emotion on a similar scale of 1 to 5. They then proceed to enter their thoughts in a form of journaling, which can be one or more sentences.

Once saved, this entry is fed into the trained BERT model to classify the entry into one of the 15 predefined cognitive distortions, including but not limited to catastrophizing, black-and-white thinking, and others. This classification result is concatenated to the original raw entry string to form a single string ready to be stored. More details on the construction of the model can be found in section 3.5.1.

5 RESULTS & DISCUSSION

In this section, we provide evaluations of the Cognify app based on user reviews and model accuracy. We conduct a pilot study in the first section to evaluate

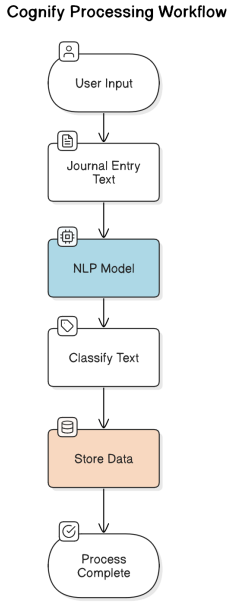


Figure 3: Cognify Processing Flowchart.

usability and trust of users for the app, then we discuss the result accuracy of the model compared to previous similar models.

5.1 Pilot Study of Cognify App

A pilot study was conducted with five participants to evaluate the app's usability, effectiveness, and user engagement, and edits were applied accordingly to ensure the app aligns with user expectations and addresses common concerns in mental health apps.

The participants for the pilot study were recruited through convenience sampling from a university setting, friends, and family. They were between the ages of 20 and 30 and represented diverse educational majors and backgrounds, including engineering, social sciences, and health-related fields. This variety ensured a basic level of heterogeneity in user perspectives, aiding the evaluation of the usability of the app across non-specialist populations.

The two-hour session includes a brief introduction to the mobile app and the concept of cognitive distortions, app download and exploration, followed by a feedback session featuring the System Usability Scale (SUS) to assess usability and user experience. Participants finally engaged in a discussion to share their comfort levels, impressions of the app, and alignment with their understanding of cognitive distortions. We discuss the analysis and results below according to usability, willingness to use as a concept, and trust in the app.

The usability analysis includes evaluating the responses from the SUS questions and calculating an overall score to determine usability, along with identifying concerns from the the open-ended responses. Calculating the average SUS value for the app yielded a score of 89.38 out of 100, which is remarkable; a SUS score above 80 indicates high usability, meaning the users found the app intuitive and easy to use. Overall, the participants had a smooth experience with minimal usability issues.

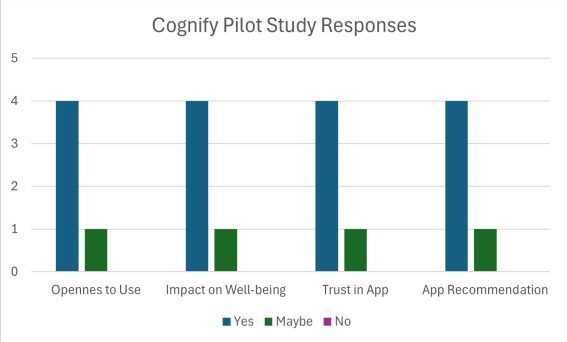


Figure 4: Cognify User Feedback Summary.

Regarding the willingness to use an app that detects and classifies cognitive distortions, we analyzed responses about whether users found the app useful for managing negative thinking, assessed their readiness to continue using the app and recommend it to their friends, and obtained feedback on their overall comprehension. 4 out of 5 participants (80%) said they would be open to using the app, with one participant saying maybe. Similarly, 4 of 5 believed the application could improve mental health, and would recommend the app to friends and family. The fifth participant voiced the need for improvements in clarifying the topic further and increasing engagement in the app for higher effectiveness.

Finally, few questions were asked on users' trust in using the app to emphasize the need for a privacy integration in any mental health application. This included examining responses regarding trust, particularly data privacy and AI analysis to determine whether users feel comfortable sharing their journal entries and having an AI model detect their cognitive distortions if any. 4 of 5 participants stated that they would trust the app if it applies a strong privacy mechanism and if there is a clear readable privacy policy. All participants expressed their concerns about AI analyzing their journal entry data, indicating a need for transparency in the way data is handled and a need for more control and clarity on the whereabouts of the datasets collected. Figure 4 highlights results of the most prominent questions.

In conclusion, the pilot study shows a promising future for Cognify as a cognitive distortion journaling

detection app. User feedback informed modifications will include higher transparency on how AI handles journal data. The clear usability of the app, along with the public interest in mental health recognition, paves way for further enhancements and extensions to fulfill further objectives such as therapist integration and privacy modules.

5.2 Cognitive Distortions Detection Model

To evaluate the model's performance, we use a dataset comprising 74,055 training samples—24,685 from the original dataset, 24,685 from SR augmentation, and 24,685 from RI augmentation. The validation and test sets contain 3,526 and 7,054 samples, respectively. We assess the model using standard text classification metrics, including precision, recall, F1-score, and accuracy. The fine-tuned RoBERTa model achieves scores of 63.05% for precision, 68.32% for recall, 64.27% for F1-score, and 68.32% for accuracy.

When evaluating such an automatic model for classifying cognitive distortions, it is crucial to consider the inherent risks of misclassification due to overlapping characteristics among various categories. For instance, the sentence "I failed this interview, I'll probably fail all interviews I get" simultaneously represents overgeneralization, magnification, and catastrophizing (Mostafa et al., 2021). Such overlaps present significant challenges for models typically trained to select a single definitive category, potentially overlooking other relevant distortions.

The complexity and inherent ambiguity of natural language further exacerbate this issue, contributing directly to low inter-annotator agreement rates. Annotation of cognitive distortions, which the model relies on during training, is inherently subjective, as annotators often struggle to choose a single dominant category, frequently inadvertently prioritizing secondary distortions (Pico et al., 2025). This subjectivity significantly complicates the creation of consistent and reliable training datasets.

Although experienced psychologists may detect a broader range of distortions in their patients, some types may still go undetected; automated cognitive distortion detection remains an invaluable tool within Cognitive Behavioral Therapy (CBT). The primary goal of CBT is to help patients self-identify their distorted thought patterns, thereby promoting self-awareness and therapeutic progress. Moreover, these automated systems support therapists by highlighting cognitive distortions that may not be explicitly evident during sessions, extending the therapists' observational capabilities.

6 CONCLUSION & FUTURE WORK

Cognify is the first mobile application to automatically detect and classify cognitive distortions necessary for the diagnosis of many psychological disorders while maintaining user privacy. Combining real-time feedback with privatized journaling, we are able to bridge the gap between self-help tools and therapist guided Cognitive Behavioral Therapy (CBT). Our solution provides a scalable proactive tool to support mental health with its privacy ensuring design and focus on accurate cognitive detection, making it an important contribution in the personalized digital mental health care sector in research.

Extensions on this research can include a chatbot system to converse with the user instead of having them write an entry directly; this can further eliminate user bias and produce more concrete data. Given the inherent ambiguity and subjectivity in cognitive distortion annotations, future research could focus on improving dataset quality and investigating the efficacy of multi-label classification frameworks to better capture overlapping distortions. Moreover, the privacy preserving framework can be further enhanced to be later applied to any mobile application holding sensitive data. Further user feedback can be obtained by expanding the pool size of participants over a longer period of time.

REFERENCES

Beck, J. S. (2022). *Cognitive behavior therapy: Basics and beyond*.

Bhave, U., Narendra, M. M., Bhadresh, D. J., Suhas, J. A., and Ajit, R. R. (2024). Mindwell solace: Your mental health companion. In *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–4.

Diano, F., Ponticorvo, M., and Sica, L. S. (2022). Mental health mobile apps to empower psychotherapy: A narrative review. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 306–311.

Elsharawi, N. and El Bolock, A. (2024). C-journal: a journaling application for detecting and classifying cognitive distortions using deep-learning based on a crowd-sourced dataset. In *Proceedings of the 2024 Joint International Conference on LREC-COLING*, pages 3224–3234.

Gaikwad, A., Nimbolkar, G., Keswani, R., Kolhe, V., and Navale, G. (2024). Mindful meadows: A mental health app. In *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–7.

Inakollu, A., Kranthi, S., and A, J. (2024). A novel approach to data security in cloud storage using erasure coding and re-encryption. In *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 972–976.

Krisnanik, E., Isnainiyah, I. N., and Resdiansyah, A. Z. A. (2020). The development of mobile-based application for mental health counseling during the covid-19 pandemic. In *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pages 324–328.

Latif, S., Hao, Y., Zhang, H., Bassily, R., and Rountev, A. (2020). Introducing differential privacy mechanisms for mobile app analytics of dynamic content. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 267–277.

Lim, S., Kim, Y., Choi, C.-H., Sohn, J.-y., and Kim, B.-H. (2024). ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mostafa, M., El Bolock, A., and Abdennadher, S. (2021). Automatic detection and classification of cognitive distortions in journaling text. In *WEBIST*, pages 444–452.

Na, H. (2024). CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

O'Loughlin, K., Neary, M., Adkins, E. C., and Schueller, S. M. (2019). Reviewing the data security and privacy policies of mobile apps for depression. *Internet Interventions*, 15:110–115.

Parker, L., Halter, V., Karliychuk, T., and Grundy, Q. (2019). How private is your mental health app data? an empirical study of mental health app privacy policies and practices. *International Journal of Law and Psychiatry*, 64:198–204.

Pico, A., Taverner, J., Vivancos, E., and Garcia-Fornes, A. (2025). Comparative analysis of the efficacy in the classification of cognitive distortions using llms. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 1: EAA*, pages 957–965. INSTICC, SciTePress.

Qi, H., Zhao, Q., Li, J., Song, C., Zhai, W., Luo, D., Liu, S., Yu, Y. J., Wang, F., Zou, H., et al. (2023). Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media. *arXiv preprint arXiv:2309.03564*.

Rasmy, M., Sabty, C., Sakr, N., and El Bolock, A. (2024). Enhanced cognitive distortions detection and classification through data augmentation techniques. In *Pacific Rim International Conference on Artificial Intelligence*, pages 134–145. Springer.

Samuel, M. and Shirley, C. (2023). Mindset, an android-based mental wellbeing support mobile application. In *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 989–996.

Shickel, B., Siegel, S., Heesacker, M., Benton, S., and Rashidi, P. (2020a). 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.

Shickel, B., Siegel, S., Heesacker, M., Benton, S., and Rashidi, P. (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. IEEE.

Shreevastava, S. and Foltz, P. (2021). Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158.

Suguna, M. and Shalinie, S. M. (2017). Privacy preserving data auditing protocol for secure storage in mobile cloud computing. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2725–2729.

Tasnim, R. A. and Eishita, F. Z. (2022). Arcod: A serious gaming approach to measure cognitive distortions. In *2022 IEEE 10th International Conference on Serious Games and Applications for Health(SeGAH)*, pages 1–8.

Vichare, A., Jose, T., Tiwari, J., and Yadav, U. (2017). Data security using authenticated encryption and decryption algorithm for android phones. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 789–794.

Vijay Sai, R., Geetha B, G., Yogeshwaran, S., Vignesh, A., and Santhosh, D. (2024). Implementation modular encryption to safeguard health data in mobile cloud environments. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1352–1357.

Wang, B., Deng, P., Zhao, Y., and Qin, B. (2023). C2d2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160.

Whaiduzzaman, M., Hossain, M. R., Shovon, A. R., Roy, S., Laszka, A., Buyya, R., and Barros, A. (2020). A privacy-preserving mobile and fog computing framework to trace and prevent covid-19 community transmission. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3564–3575.

Zhou, T., Cai, Z., Xiao, B., Wang, L., Xu, M., and Chen, Y. (2018). Location privacy-preserving data recovery for mobile crowdsensing. volume 2, pages 1–23.