Smart Health 36 (2025) 100565

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Continuous prediction of user dropout in a mobile mental health intervention program: An exploratory machine learning approach

Pinxiang Wang [a], Hanqi Chen [a], Zhouyu Li [a], Wenyao Xu [b], Yu-Ping Chang [b], Huining Li [a],*

[a] Department of Computer Science, North Carolina State University, United States
[b] Department of Computer Science and Engineering, University at Buffalo, United States

ARTICLE INFO

Keywords:
Mobile health
Mental health intervention
Dropout prediction
Machine learning

ABSTRACT

Mental health intervention can help to release individuals' mental symptoms like anxiety and depression. A typical mental health intervention program can last for several months, people may lose interests along with the time and cannot insist till the end. Accurately predicting user dropout is crucial for delivering timely measures to address user disengagement and reduce its adverse effects on treatment. We develop a temporal deep learning approach to accurately predict dropout, leveraging advanced data augmentation and feature engineering techniques. By integrating interaction metrics from user behavior logs and semantic features from user self-reflections over a nine-week intervention program, our approach effectively characterizes user's mental health intervention behavior patterns. The results validate the efficacy of temporal models for continuous dropout prediction.

1. Introduction

Mental health intervention is a therapeutic approach to improve psychological well-being and address various mental health disorders (Ybarra & Eaton, 2005). Mental health intervention program typically lasts for months or years, which begins with foundational modules and incrementally advances to more complex ones over time (Gimba et al., 2020). According to a report, user dropout rate of the intervention programs can range from 25% to 56% (Ybarra & Eaton, 2005), which largely reduce the overall effectiveness. While session-based incentives and email reminders are commonly used to improve retention, these approaches lack the ability to promptly detect early signs of disengagement. Early identification of user disengagement is essential to implement timely measures to address dropout, which can reduce its adverse effects on treatment outcomes. Building on this, automatic and efficient dropout prediction strategies are crucial for success of mental health intervention.

At the beginning, dropout prediction is used in online education platforms (Dalipi, Imran, & Kastrati, 2018; Jeon & Park, 2020) to identify disengaged students and improve course completion rates. Bremer, Chow, Funk, Thorndike, and Ritterband (2020) extended dropout prediction to digital health Intervention programs. They proposed a machine learning framework that uses a fixed-size feature matrix to model user behavior over time. However, this approach relies on fixed-length features, which often overlooks the temporal dynamics of user behavior. By treating time-series data as static inputs, it becomes less effective at capturing how user engagement develops and changes. Zantvoort, Scharfenberger, Boß, Lehr, and Funk (2023) studied different machine learning models, e.g., support vector machine (SVM), long short-term memory (LSTM), BERT, and text representations such as term

* Corresponding author.
E-mail addresses: pwang25@ncsu.edu (P. Wang), hchen54@ncsu.edu (H. Chen), zli85@ncsu.edu (Z. Li), wenyaoxu@buffalo.edu (W. Xu), yc73@buffalo.edu (Y.-P. Chang), hli83@ncsu.edu (H. Li).

https://doi.org/10.1016/j.smhl.2025.100565
Received 10 March 2025; Accepted 15 March 2025
Available online 25 March 2025
2352-6483/Published by Elsevier Inc.

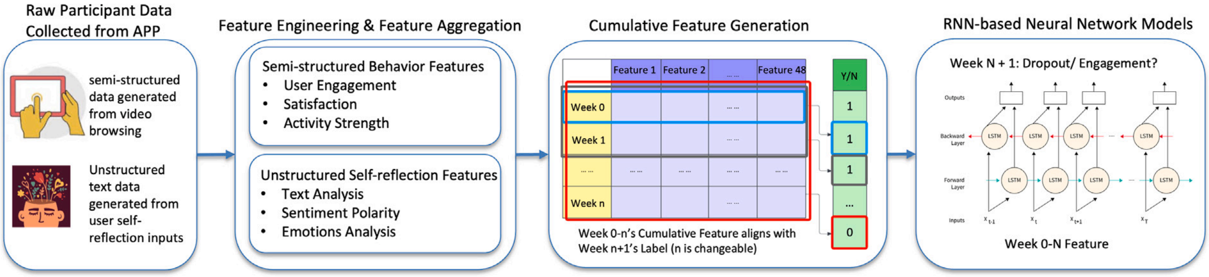
Raw Participant Data Collected from APP
semi-structured data generated from video browsing
Unstructured text data generated from user self-reflection inputs
Feature Engineering & Feature Aggregation
Semi-structured Behavior Features
User Engagement
Satisfaction
Activity Strength
Unstructured Self-reflection Features
Text Analysis
Sentiment Polarity
Emotions Analysis
Cumulative Feature Generation
Feature 1
Feature 2
Feature 48
Y/N
Week 0
Week 1
Week n
1
1
1
...
0
Week 0-n's Cumulative Feature aligns with Week n+1's Label (n is changeable)
RNN-based Neural Network Models
Week N + 1: Dropout/ Engagement?
Outputs
Backward Layer
Forward Layer
Inputs
LSTM
x_{t-1}
x_t
x_{t+1}
x_T
Week 0-N Feature

Fig. 1. System overview.

frequency-inverse document frequency (TF-IDF) and word embeddings for predicting dropout in digital mental health programs. However, their results were limited by a small sample size, and advanced models did not perform better than simpler ones, even with data augmentation. Although Large Language Models (LLMs) have excellent performance in data analysis in many areas, due to concerns about personal information privacy, the critical factor in healthcare, they may not be the optimal choice. LLMs are more vulnerable to data leakage through black-box extraction, inference, and reconstruction attacks (Lukas et al., 2023), and based on existing research, the larger the language model, the more vulnerable it becomes to data extraction attacks (Carlini et al., 2021). In contrast, traditional machine learning models provide better control and transparency regarding data handling, significantly reducing privacy risks, and their smaller scale makes them far less vulnerable than LLMs.

We learn the lessons from the previous works and develop a continuous dropout prediction framework, specifically tailored for our Mellowing Mind study—a mobile mental health intervention program designed for the African American community. This 9-week program is delivered through a smartphone app, where users complete different modules each week. For example, users need to watch different kinds of mindfulness videos and practice in daily life. Users are encouraged to perform self-reflection (Chen, Chang, & Stuart, 2020) by recording their happy events and unhappy events in our app.

We collect user interaction data with our app in the back end. First, we extract features from the heterogeneous interaction data, including user engagement metrics, sentiment trends, and emotional dynamics, which capture both static and evolving behavioral insights. Next, we adopt time-series modeling to effectively represent the temporal dependencies and cumulative trends in user behavior. Leveraging temporal models such as LSTM, BiLSTM, and GRU, we achieve high predictive performance, with AUC exceeding 0.95 and balanced accuracy surpassing 0.92 over a 5-fold cross validation. This approach not only ensures accurate dropout predictions but also supports timely interventions to enhance user retention and the overall success of mental health intervention programs (see Fig. 1).

2. Mellowing mind study

Our study has received approval from the University's Institutional Review Board. Our clinical team designs standard inclusion and exclusion criterias to recruit subjects from the African American community. 49 subjects are enrolled in our study. Among them, 71.5% are black or African American, 22.4% are White, and 6.1% are Asian. They are aging from 20 to 84 years old.

Subjects are asked to attend our mobile mental health intervention program, which is delivered through a smartphone app available on Google Play. Before the program, we host online training sessions to provide step-by-step guidance on how to use the app. Our intervention program lasts for 9 weeks, consisting of eight weekly sessions and an introductory session. Each weekly session features instructional videos and audio content, including guided meditations and practical, interactive suggestions for incorporating mindfulness into daily life. Participants are also encouraged to monitor their daily mood and record self-reflections on positive or negative events, as well as their personal insights and progress in mindfulness practice. All their interaction data with the App are recorded anonymously in the back end.

3. Feature engineering

It is crucial to extract effective features from heterogeneous and dynamic user interaction data with our app. In this section, we detail the feature engineering process.

3.1. Semi-structured behavior feature

In the mental health intervention program, users need to watch different kinds of mindfulness videos and practice in daily life. To quantify users' video-watching behaviors, we extract weekly features based on their interaction with mindfulness practice videos and audio content, as recorded through semi-structured screen-touch data. Specifically, the calculation features include activity completeness, pause duration, and a satisfaction index (ranging from "very calm" to "very stressed") for each mindfulness training video. These features are further analyzed to determine their weekly averages, standard deviations(STD), and rates of change(slope). Furthermore, we characterize behavioral patterns by calculating the most frequent starting times, engagement frequency, and time variability in users' interactions with these videos throughout the week. In total, we obtain 20 behavior features each week.

3.2. Unstructured self-reflection feature

Self-reflection on pleasant and unpleasant events is encouraged in the mental health intervention program. User self-reflections are recorded as unstructured text. From these data, we extract weekly features to capture sentiment trends and emotional dynamics. Specifically, we extract three types of features presented as follows.

3.2.1. Basic lexical and semantic features

Basic lexical and semantic feature analysis can provide an overview of the fundamental characteristics of the text (Pustejovsky, Bergler, & Anick, 1993). Therefore, we extract lexical and semantic features, such as word counts and part-of-speech frequencies (e.g., nouns, verbs, adjectives) from each instance of self-reflection data. These features reflect the general structure and complexity of the language, and metrics such as word frequency offer insight into the richness and level of engagement of the content.

3.2.2. Emotional features

Emotional features offer critical insights into the affective states expressed in user reflections, helping to evaluate mental engagement and well-being (Uban, Chulvi, & Rosso, 2021). Accordingly, we compute the following features.

Sentiment Polarity: Sentiment polarity scores (Loria et al., 2018) range from -1 (negative sentiment) to $+1$ (positive sentiment). The polarity score P_{polarity} is calculated as:

$$P_{\text{polarity}} = \frac{\sum_{i=1}^{n} S(w_i)}{n}. \tag{1}$$

where $S(w_i)$ represents the sentiment score of the ith word, and n is the total number of words in the text. **Emotion Distribution:** Emotion probabilities are derived using the pre-trained transformer model `Distilbert base uncased go emotions student`(Demszky et al., 2020). The probability ranges from 0–1 for each selected emotion. For an input $\mathbf{x}$, the probability of emotion e_j is computed as:

$$P(e_j|\mathbf{x}) = \frac{\exp(\mathbf{W}_j\mathbf{h} + b_j)}{\sum_{k=1}^{K} \exp(\mathbf{W}_k\mathbf{h} + b_k)},$$

where $\mathbf{h}$ is the hidden state of the model and K represents the total number of categories of emotions. This analysis enables tracking emotional evolution over time.

3.2.3. Engagement features

Engagement features are useful for evaluating user consistency over time (Zhou & Bhat, 2021), highlighting regularity or variability in engagement that may indicate potential dropout risks. We calculate engagement features based on user's active time points on doing tasks in our app.

3.3. Feature aggregation and calculation standards

To summarize features and maintain temporal properties, we further calculate the mean, standard deviation, and slope of the extracted features over a week. These measures help quantify the central tendency, variability, and trends of user behaviors, providing a comprehensive view of the dynamics of engagement. To ensure temporal consistency among users, we standardized all features in a weekly aggregated format.

3.4. Feature matrix summary

We extract a total of 48 features, as outlined in Table 1, categorized into six domains. Mindfulness Video-watching behavior (14 features), After-Mindfulness Satisfaction (3 features), Engagement and Activity Strength (3 features), Lexical and Semantic analysis (7 features), Emotion Variances (14 features) and Emotion Trends (7 features). These features provide a comprehensive representation of user behavior, emotional states, and engagement consistency. By aggregating data weekly using mean, standard deviation, and slope calculations, the unified feature matrix effectively integrates both video-watching behavior features and self-reflection features. This structured approach captures patterns over time, offering a reliable foundation for accurate and meaningful dropout predictions.

4. Prediction model

The weekly aggregated feature matrices in our study show strong temporal characteristics. Capturing these dynamic features is essential to achieve our prediction goals. Fei and Yeung (2015) employed temporal models such as vanilla RNN and LSTM on a weekly aggregated dataset for online student dropout prediction, demonstrating that RNN-based temporal models have significant potential for continuous dropout prediction tasks. Based on this finding, we selected LSTM, GRU, and biLSTM (Dey & Salem, 2017; Graves & Graves, 2012; Zhou et al., 2016). We also incorporated a customized negative weight loss function to address the unbalanced data set to process the temporal dependencies within our weekly feature matrices.

Table 1
Summary of extracted features.

Category	Feature name	Description
Mindfulness video-watching behavior	Active percentage, study day std, completeness percentage, completeness mean, completeness std, completeness slope, wasted mean, wasted std, wasted slope, comment count, comment date std, ets mct, ets mct frequency, ets time diff	Features related to user activity levels, task completion, and time efficiency. They capture weekly trends, variability, and overall participation consistency.
After-mindfulness satisfaction	Satisfaction mean, satisfaction std, satisfaction slope	Weekly satisfaction levels, variability, and trends over time, reflecting user feedback on the intervention.
Engagement and activity strength	Activity strength mean, activity strength std, activity strength slope	Metrics capturing the mean, variability, and weekly trends of activity levels, reflecting user engagement strength.
Lexical and semantic analysis	Word count mean, noun count mean, verb count mean, adj count mean, spelling error rate mean, sentiment mean, sentiment std	Text-based features derived from weekly comments, including word usage, sentiment analysis, and spelling error rates, providing linguistic and sentiment insights.
Emotion variance	Anger mean, anger std, sadness mean, sadness std, joy mean, joy std, relief mean, relief std, disgust mean, disgust std, optimism mean, optimism std, neutral mean, neutral std	Aggregated emotional metrics capturing mean and variability for 7 selected emotions, representing a broad spectrum of typical emotional states during the intervention. These emotions were chosen for their relevance and interpretability in understanding user behavior.
Emotion trends	Anger slope, sadness slope, joy slope, relief slope, disgust slope, optimism slope, neutral slope	Weekly trends in 7 selected emotional states, providing insights into temporal changes in users' feelings over the intervention period.

Table 2
Model performance comparison (5-fold average).

Model	Accuracy	Balanced Acc.	Precision	Recall	F1 Score	AUC	TNR	FPR	NPV	F1-Negative
LSTM	0.9901	0.9184	0.9908	**0.9987**	**0.9948**	**0.9565**	0.8379	0.1621	**0.9807**	0.8926
BiLSTM	0.9896	**0.9211**	**0.9910**	0.9980	0.9945	0.9399	**0.8442**	**0.1557**	0.9734	**0.8942**
GRU	**0.9855**	0.8815	0.9868	0.9981	0.9924	0.9405	0.7649	0.2351	0.9741	0.8257

The customized loss function is a weighted binary cross-entropy loss designed to handle the class imbalance by assigning three times more weight to negative samples than to positive samples. The loss for each sample is defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[-w_{pos}\, y_i\, \log(\hat{y}_i) - w_{neg}\,(1-y_i)\,\log(1-\hat{y}_i)\right]. \tag{2}$$

where y_i is the true label ($y_i \in \{0,1\}$), $\hat{y}_i$ is the predicted probability ($0 < \hat{y}_i < 1$), w_{pos} is the weight for positive samples and $w_{neg} = 3 \cdot w_{pos}$ is the weight for negative samples in our experiment settings. This loss function ensures that the model pays more attention to negative samples to mitigate the effects of the imbalance of the data set.

Based on previous work (Fei & Yeung, 2015), we selected LSTM as our baseline due to its excellent performance on time-series data. Furthermore, we incorporated two variants of LSTM: Gated recurrent unit (GRU) and bidirectional long-short-term memory (biLSTM), given their complementary strengths to different prediction requirements. GRU has a simplified architecture with reduced computational complexity that offers efficiency advantages for smaller datasets while maintaining strong temporal modeling capabilities. BiLSTM, with its ability to capture bidirectional information, provides a more comprehensive understanding of sequential dependencies. By comparing these three models, we aim to identify the most suitable architecture for our dataset, balancing prediction accuracy and computational efficiency.

5. Evaluation

5.1. Data preparation

The data preparation process includes missing value handling, feature engineering, data augmentation, and temporal aggregation to construct a cohesive dataset for model evaluation. Missing values are dealt with zero-padding techniques. To enrich the dataset, we applied well-designed data augmentation techniques: semi-structured log data were augmented using combination and permutation strategies (Wang et al., 2024), while self-reflection textual data used large language models and translation tools to generate semantically consistent variants (Ding et al., 2024).

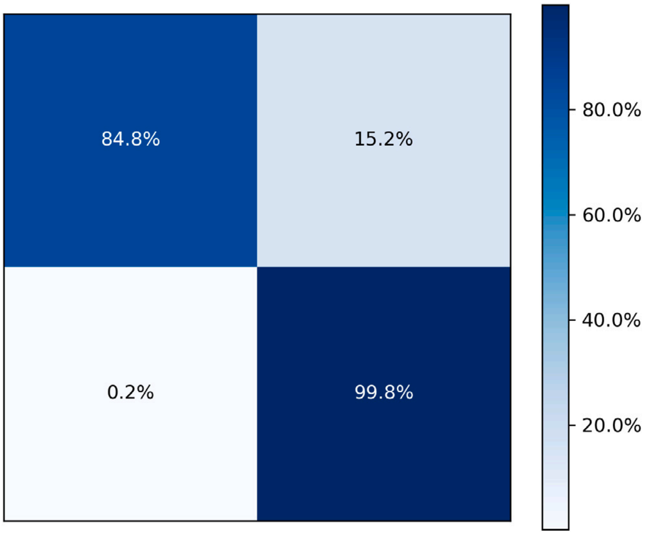

Fig. 2. Confusion matrix for BiLSTM.

The aggregated weekly characteristics were consolidated into uniform nine-week time step matrices to capture temporal patterns of user behavior. The dataset was structured such that each time step represented an accumulation of features from previous weeks, with labels assigned based on the dropout status of the subsequent week.

For instance, the first time step (week 0) was labeled according to dropout occurrence in the first week, while the second time step (weeks 0–1) was labeled based on dropout in the second week, and so forth. This approach resulted in cumulative datasets of dimensions 48 × 1, 48 × 2, …, $48 \times n$, where n varied from 1 to 9, reflecting the number of weeks aggregated. To standardize input for model training, any missing entries in datasets with dimensions smaller than 48 × 9 were filled with zeros, ensuring that all feature sets were consolidated into a uniform size of 48 × 9.

5.2. Implementation

We utilized RNN-based models (LSTM (Graves & Graves, 2012), BiLSTM (Zhou et al., 2016), GRU (Dey & Salem, 2017)) with a 64-size hidden layer and 9-time-step sliding windows, Each entry contains 48 features. Derived from the cumulative week segmentation strategy, our dataset has 404,000 positive samples and 25,000 negative samples. For each fold, around 308,000 positive samples and 21,000 negative samples were used for training, while 96,000 positive samples and 4000 negative samples were designated for testing. The models were trained for 20 epochs using the Adam optimizer (learning rate: 0.001) and a batch size of 32.

5.3. Evaluation metrics

We assessed the model using 5-fold cross-validation with a comprehensive set of metrics: accuracy, balanced accuracy, F1 score, and area under the curve (AUC). To explore the model's performance on negative samples, we emphasized metrics such as, and true negative rate (TNR), false negative rate (FPR), Negative Predictive Value (NPV), F1-Negative ensuring a thorough evaluation of the model's ability to correctly identify negative instances. To further explore the robustness and generalization of our models, we also analyzed the accuracy in different demographic groups such as gender, marital status, and education level.

5.4. Results and analysis

Based on the results shown in Table 2, BiLSTM demonstrated the best overall performance over the three models, achieving the highest balanced accuracy (0.9211) and TNR (0.8442), proving it to be particularly suitable for tasks that prioritize high accuracy, especially when working with unbalanced datasets. Fig. 2 futher showns the confusion matrix for BiLSTM.

Figs. 3, 4, and 5, indicate that gender and level of education have a minimal impact on accuracy. The model has consistent performance across these demographic groups. However, marital status exhibited significant variability in accuracy, with groups such as "Separated" and "Divorced" showing relatively lower performance. These categories may involve more complex or unpredictable emotional dynamics and features. Future work could focus more on designing refined strategies that target marital groups to implement more robust predictive models.

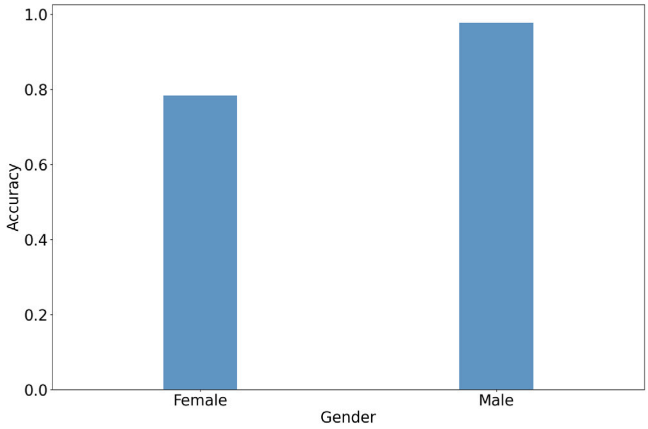

Fig. 3. The impact of gender on dropout prediction.

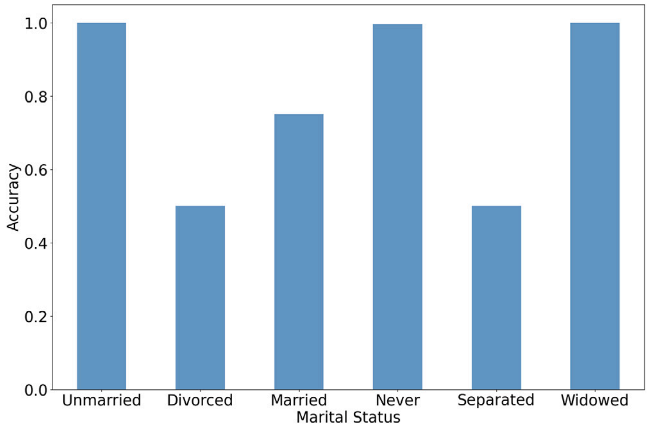

Fig. 4. The impact of marital status on dropout prediction.

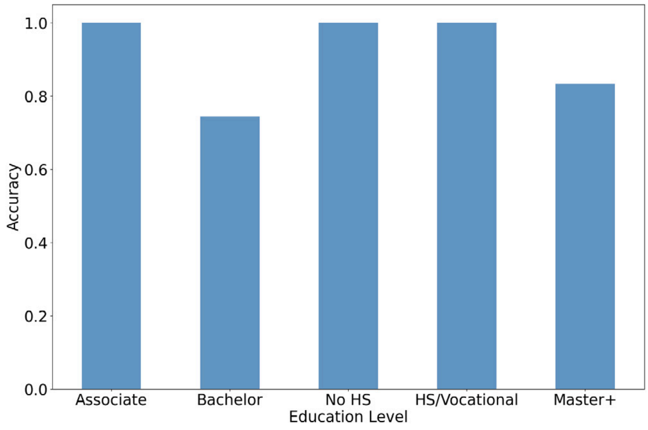

Fig. 5. The impact of education level on dropout prediction.

CRediT authorship contribution statement

Pinxiang Wang: Methodology, Writing – original draft. **Hanqi Chen:** Validation, Writing – review & editing. **Zhouyu Li:** Investigation, Writing – review & editing. **Wenyao Xu:** Software, Funding acquisition. **Yu-Ping Chang:** Data curation, Funding acquisition. **Huining Li:** Project administration, Supervision, Writing – review & editing.

Acknowledgments

Funding/support

Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (COVID-2020C2-11158).

Disclaimer

The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yu-Ping Chang reports financial support was provided by Patent-Centered Outcomes Research Institute. If there are other authors they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Bremer, V., Chow, P. I., Funk, B., Thorndike, F. P., & Ritterband, L. M. (2020). Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: machine learning approach. *Journal of Medical Internet Research*, *22*(10), Article e17738.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2021). Extracting training data from large language models. In *30th USENIX security symposium* (pp. 2633–2650).

Chen, S.-P., Chang, W.-P., & Stuart, H. (2020). Self-reflection and screening mental health on Canadian campuses: validation of the mental health continuum model. *BMC Psychology*, *8*, 1–8.

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE global engineering education conference* (pp. 1007–1014). http://dx.doi.org/10.1109/EDUCON.2018.8363340.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547.

Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems* (pp. 1597–1600). IEEE.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., et al. (2024). Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the association for computational linguistics ACL 2024* (pp. 1679–1705).

Fei, M., & Yeung, D.-Y. (2015). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE international conference on data mining workshop* (pp. 256–263). IEEE.

Gimba, S. M., Harris, P., Saito, A., Udah, H., Martin, A., & Wheeler, A. J. (2020). The modules of mental health programs implemented in schools in low-and middle-income countries: findings from a systematic literature review. *BMC Public Health*, *20*, 1–10.

Graves, A., & Graves, A. (2012). Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks* (pp. 37–45). Springer.

Jeon, B., & Park, N. (2020). Dropout prediction over weeks in MOOCs by learning representations of clicks and videos. arXiv:2002.01955. URL https://arxiv.org/abs/2002.01955.

Loria, S., et al. (2018). Textblob documentation. Release 0.15 2 (8) 269.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models. In *2023 IEEE symposium on security and privacy* (pp. 346–363). IEEE.

Pustejovsky, J., Bergler, S., & Anick, P. (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics*, *19*(2), 331–358.

Uban, A.-S., Chulvi, B., & Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, *124*, 480–494.

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., et al. (2024). A comprehensive survey on data augmentation. arXiv preprint arXiv:2405.09591.

Ybarra, M. L., & Eaton, W. W. (2005). Internet-based mental health interventions. *Mental Health Services Research*, *7*, 75–87.

Zantvoort, K., Scharfenberger, J., Boß, L., Lehr, D., & Funk, B. (2023). Finding the best match—a case study on the (text-) feature and model choice in digital mental health interventions. *Journal of Healthcare Informatics Research*, *7*(4), 447–479.

Zhou, J., & Bhat, S. (2021). Modeling consistency using engagement patterns in online courses. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 226–236).

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., et al. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 207–212).