

Creating a Clinical Psychology Dataset with Synthetic Data: Automatic Detection of Cognitive Distortions Classified with NLP

Hakkı Halil BABACAN¹, Ramazan OĞUZ^{2*}, Yahya Kemal BEYİTOĞLU³

^{1,2} Institute of Forensic Sciences and Legal Medicine, İstanbul University-Cerrahpaşa, İstanbul, Türkiye

³ Vocational School of Besni, Adıyaman University, Adıyaman, Türkiye

¹ hakkı.babacan@erzincan.edu.tr, ^{2*} ramazan.oguz@ogr.iuc.edu.tr, ³ ybeyitoglu@adiyaman.edu.tr

(Geliş/Received: 16/04/2024;

Kabul/Accepted: 16/09/2024)

Abstract: Cognitive distortions are thought errors that lead individuals to perceive reality in a misleading way and are strongly associated with psychopathologies. Therefore, accurately identifying and classifying distortions can enhance the effectiveness of cognitive-behavioral therapy (CBT). This study investigates the effectiveness of deep learning and NLP techniques for the automatic detection of cognitive distortions. The RoBERTa model was trained using English synthetic data generated by GPT-4 (2000 examples) and the dataset from Shreevastava and Foltz (1590 cognitive distortion examples, 933 non-distortion examples). Three scenarios were tested: the original dataset, the synthetic dataset, and their combination. The results showed that synthetic data is a strong resource. Accuracy rates were 60.67% (original), 94.51% (synthetic), and 77.18% (combined). The GPT-4-based dataset provided almost perfect F1 scores, particularly in some categories. ROC curve analyses showed that the GPT-4 dataset had the highest AUC value (0.80). The study revealed that using synthetic data expands the potential of AI applications in clinical psychology and offers a way to develop effective models while preserving patient privacy. Future research should test synthetic data with different models and compare it with real clinical data.

Key words: Cognitive distortion, machine learning, natural language processing, GPT-4, depression.

Sentetik Verilerle Klinik Psikoloji Veri Seti Oluşturma: Bilişsel Çarpıtmaların NLP ile Sınıflandırılarak Otomatik Tespiti

Öz: Bilişsel çarpıtmalar, bireylerin gerçekliği yanıltıcı bir şekilde algılamalarına neden olan düşünce hatalarıdır ve psikopatolojilerle güçlü bir ilişkisi vardır. Bu nedenle, çarpıtmaların doğru bir şekilde belirlenmesi ve sınıflandırılması, bilişsel davranışçı terapinin (CBT) etkinliğini artırabilir. Bu çalışma, bilişsel çarpıtmaların otomatik tespiti için derin öğrenme ve NLP tekniklerinin etkinliğini incelemektedir. GPT-4 ile üretilen İngilizce sentetik veriler (2000 örnek) ve Shreevastava ve Foltz'un veri seti (1590 bilişsel çarpıtma, 933 çarpıtma içermeyen örnek) kullanılarak RoBERTa modeli eğitilmiştir. Üç senaryo test edilmiştir: orijinal veri seti, sentetik veri seti ve bunların kombinasyonu. Sonuçlar, sentetik verilerin güçlü bir kaynak olduğunu göstermiştir. Doğruluk oranları sırasıyla %60,67 (orijinal), %94,51 (sentetik) ve %77,18 (kombine) olarak elde edilmiştir. GPT-4 tabanlı veri seti, özellikle bazı kategorilerde neredeyse mükemmel F1 skorları sağlamıştır. ROC eğrisi analizleri, GPT-4 veri setinin en yüksek AUC değerine (0,80) sahip olduğunu göstermiştir. Çalışma, sentetik veri kullanımının klinik psikolojide yapay zeka uygulamalarının potansiyelini genişlettiğini ve hasta gizliliğini korurken etkili modeller geliştirmenin bir yolunu sunduğunu ortaya koymuştur. Gelecekteki araştırmalar için, sentetik verilerin farklı modellerle test edilmesi ve gerçek klinik verilerle karşılaştırılması önerilmektedir.

Anahtar kelimeler: Bilişsel çarpıtma, makine öğrenimi, doğal dil işleme, GPT-4, depresyon.

1. Introduction

Cognitive distortions are thought errors that lead individuals to perceive reality in a misleading and often negative way [1]. These distortions are typically addressed within cognitive-behavioral therapy (CBT), one of the goals of which is to help individuals recognize and correct these distortions [2]. There are various types of cognitive distortions, including overgeneralization, personalization, filtering, black-and-white thinking, and catastrophizing [3]. These distortions can negatively impact individuals' emotional states and behaviors. Understanding and recognizing cognitive distortions can positively influence individuals' emotional states and behaviors. Beck [4] suggests that depression stems from cognitive distortions.

There are quite striking findings in the literature regarding the relationship between cognitive distortions and psychopathologies, as well as their impact on individuals' daily lives. Rnic et al. [5] found that cognitive distortions can prevent individuals from adopting a humorous and joyful perspective on life, leading to an increase in depressive symptoms. This study provided strong evidence of the impact of cognitive distortions on depressive

* Sorumlu yazar: ramazan.oguz@ogr.iuc.edu.tr. Yazarların ORCID Numarası: ¹ 0000-0001-9609-5128, ² 0000-0002-7297-4141, ³ 0000-0001-6421-8939

symptomatology. Additionally, a study conducted on adolescents also presented strong evidence that cognitive distortions play an active role in the onset of depression [6]. Beyond the symptomatology of depression, it is known that a life-threatening serious mental disorder, such as Anorexia Nervosa, is characterized by cognitive distortions about weight and body shape [7]. Recent studies also reveal the effects of cognitive distortions on individuals' behaviors. There is current evidence that cognitive distortions have an active role in the emergence and maintenance of gambling behavior [8] and in predicting depression, anxiety, and stress [9].

Cognitive-behavioral therapy (CBT) aims to identify and correct cognitive distortions [2]. Recent advancements in deep learning and natural language processing (NLP) have brought the automatic detection and classification of cognitive distortions to the forefront. NLP, a subfield of artificial intelligence, enables computers to understand and process human languages, and significant progress has been made in this area. NLP creates semantic vector representations of words through word and sentence embedding techniques [10]. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are effective in tasks such as classification and language translation [11]. Transformer models like BERT excel in many NLP tasks by better capturing the meaning and context of language [12]. NLP systems operate through preprocessing, model training, and inference stages, employing supervised, semi-supervised, and unsupervised learning methods [10]. However, the inner workings of deep learning models are often opaque, leading to the development of various interpretability techniques [13]. For example, machine translation systems use transformer models to translate sentences from a source language to a target language, successfully capturing contextual meaning [11]. In recent years, NLP models have demonstrated significant success in text classification and detection tasks involving symbolic structures in natural language. The success of NLP techniques in text classification has led to the application of these techniques to study many phenomena in psychology.

The meta-analysis conducted by Harrigan et al. [14] reported that data mining from social networks like Reddit and Twitter has been used to create datasets on topics such as depression and suicidal ideation, with automatic detection studies conducted using NLP techniques. However, the number of NLP studies related to cognitive distortions, which are reported as significant predictors of depression, is quite limited. Shickel et al. [15] conducted one of the first studies on the automatic detection of cognitive distortions using machine learning and deep learning, reporting that the application of machine learning techniques in the context of mental health is highly limited and that the cognitive-behavioral perspective has been neglected. They also noted that studies on the detection and classification of cognitive distortions bear similarities to text-based emotion recognition tasks, and in their study, they utilized a dataset consisting of crowdsourced data and mental health therapy records [15]. Although the models achieved an F1 score of 0.88 in the classification task of determining the presence or absence of cognitive distortion, when the model was applied to smaller counseling datasets, the F1 score dropped to 0.45. Additionally, the dataset used in the study was not shared publicly.

In another significant study on the subject, a dataset created from a publicly available therapist-patient question-and-answer dataset [16] on Kaggle was trained using the BERT [17] model, which is based on the Transformer architecture [18]. In this study, inferences were made from the therapist question-and-answer dataset regarding cognitive distortions, leading to the creation of an English cognitive distortion dataset consisting of 1,590 cognitive distortion examples and a total of [19], 523 data points. In the binary classification task, the F1 score for the SVM (Support Vector Machines) and BERT-based approaches was determined to be 0.79. As mentioned in the study, the biggest challenge in using artificial intelligence in clinical psychology is the lack of datasets [18]. In addition to these challenges, issues such as reaching a consensus among domain experts on statements containing depression and cognitive distortion values, as well as patient privacy concerns, make access to large data sources difficult. Information on the existing datasets related to cognitive distortions in the literature is provided in Table 1.

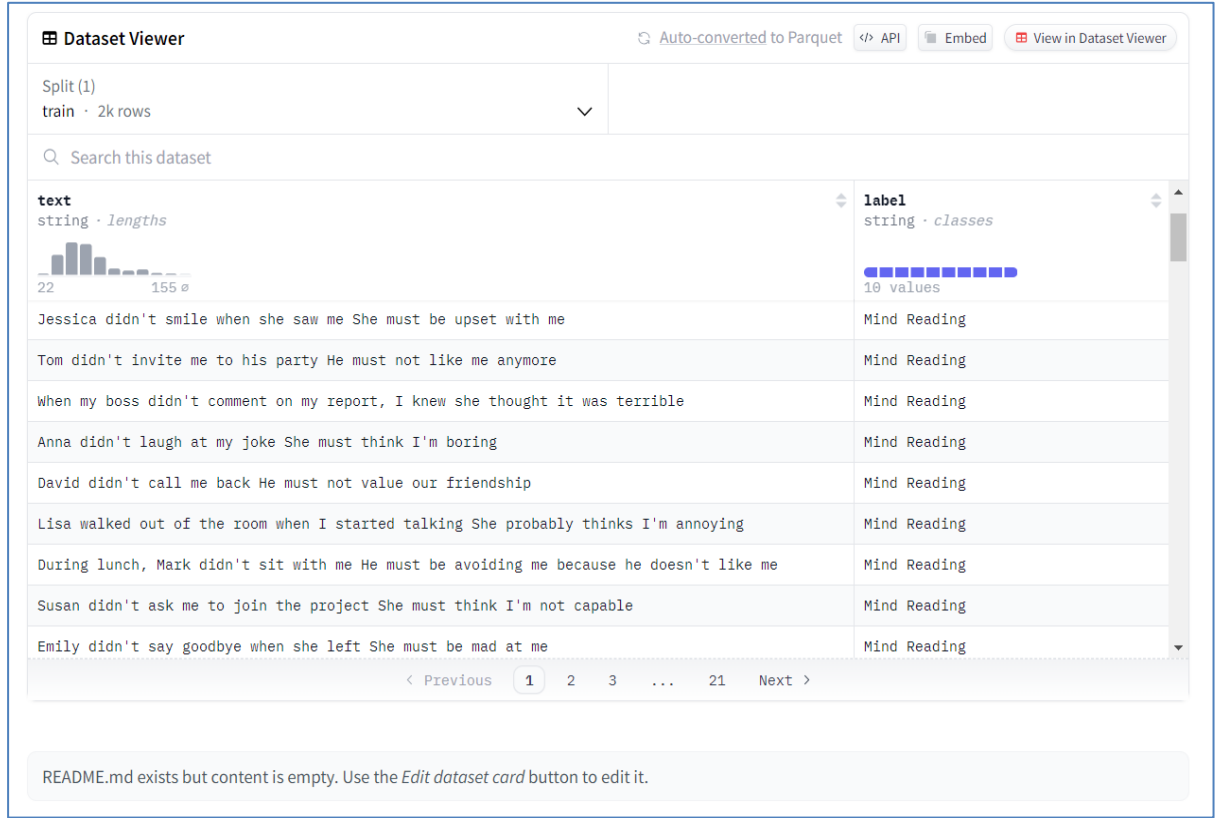
Table 1. Existing cognitive distortion datasets.

Evaluation Criteria	Our Dataset	MH Tumblr Dataset: Shickel et al. [15]	Shreevastava & Foltz [18]
Publicly Accessible	✓	✗	✓
Multiple Types of Cognitive Distortions	✓	✓	✗
Labeled by Psychology Professionals	✓	✗	✗
Dataset Language	English	English	English
Dataset Scale	4,531	2,769	1,728

This study aims to generate synthetic data using GPT-4 LLM (Large Language Model) developed by OpenAI to overcome the existing data limitations in the literature and the ethical issues related to obtaining patient data. Even if personal data is collected with patient's consent and anonymized, datasets may still contain clues related to individuals in clinical psychology. To address this issue and explore the contributions of synthetic data to the field of deep learning, this method was chosen. GPT-4, created by OpenAI as a large multimodal model based on the transformer architecture, has demonstrated success in generating text for many challenging tasks [20]. GPT-4 is a significantly enhanced language model developed by OpenAI compared to its predecessors. With billions of parameters, GPT-4 is designed to understand the semantic and contextual aspects of language more deeply. Built on GPT-3, which contains 175 billion parameters, GPT-4 excels in natural language processing (NLP) tasks [21]. The number of parameters in GPT-4 is one of its most distinguishing features, allowing the model to comprehend more complex language structures and produce more coherent texts [20]. This increase in the number of parameters enables GPT-4 to deliver more accurate and reliable results across a wide range of tasks. Additionally, GPT-4's multimodal capabilities allow it to process not only text but also visual data, making it one of the most powerful models in the NLP field [20]. Considering the privacy and subjectivity aspects of clinical psychology data, this study's primary research question is whether GPT-4 can be an effective tool for generating data related to clinical psychology. The purpose of this research is to compare the performance of existing datasets by generating synthetic data related to cognitive distortions using GPT-4. Additionally, the study aims to investigate whether the validity and success rates of existing datasets can be improved with the inclusion of synthetic data generated by GPT-4. In line with these objectives, the potential to achieve successful results from synthetic data generated by GPT-4 may provide a significant contribution to the literature and offer a practical solution to the existing dataset limitations in AI studies related to clinical psychology. For this study, the RoBERTa model, based on the transformer architecture and previously trained to extract semantic information from short texts successfully, was chosen for training [22]. The study asked a GPT-4-based model to generate a total of 2,000 realistic cognitive distortion statements related to ten different types of cognitive distortions, each consisting of at least two sentences.

2. Methods and Dataset

In the study, two datasets were prepared (Figure 1). In the first phase, a GPT-4-based model was asked to generate 200 realistic cognitive distortion statements for each of ten different types of cognitive distortions, resulting in a total of 2,000 statements, each consisting of at least two sentences (for relevant prompts, see references)[23]. All cognitive distortions generated by GPT-4 were considered synthetic expressions. The synthetic data generated were semantically reviewed by two domain expert psychologists. For training, the "no distortions" class was represented by using the data from the "no distortions" category (933 statements) in the dataset created by Shreevastava and Foltz [18], which is the only publicly available dataset on this topic, combined with the synthetic data from the "distortions" class. The first dataset was prepared to compare the metric values of synthetic data with existing public datasets. The second dataset involved adding synthetic data to the dataset created by Shreevastava and Foltz [18]. This dataset aimed to examine the change in performance when synthetic data is added to the publicly available cognitive distortion dataset. During the data preprocessing stage, punctuation marks other than periods and exclamation points were removed. The data were formatted to be compatible with CSV UTF-8 format. The dataset [24] composed entirely of synthetic cognitive distortion sentences and the other dataset [25], which was augmented with synthetic data, were uploaded to the Huggingface platform. Eighty percent of the data was allocated for training and 20% for testing. The data were trained using the RoBERTa model, a transformer-based approach that was previously trained.



Dataset Viewer Auto-converted to Parquet API Embed View in Dataset Viewer

Split (1)
train · 2k rows

Search this dataset

text	label
string · lengths	string · classes
22 155	10 values
Jessica didn't smile when she saw me She must be upset with me	Mind Reading
Tom didn't invite me to his party He must not like me anymore	Mind Reading
When my boss didn't comment on my report, I knew she thought it was terrible	Mind Reading
Anna didn't laugh at my joke She must think I'm boring	Mind Reading
David didn't call me back He must not value our friendship	Mind Reading
Lisa walked out of the room when I started talking She probably thinks I'm annoying	Mind Reading
During lunch, Mark didn't sit with me He must be avoiding me because he doesn't like me	Mind Reading
Susan didn't ask me to join the project She must think I'm not capable	Mind Reading
Emily didn't say goodbye when she left She must be mad at me	Mind Reading

< Previous 1 2 3 ... 21 Next >

README.md exists but content is empty. Use the [Edit dataset card](#) button to edit it.

Figure 1. Dataset example.

2.1. RoBERTa model

The success of the BERT model, a transformer-based approach, in text-based emotion classification tasks has led to its use in the automatic detection of cognitive distortions [18]. RoBERTa, another model with a transformer-based architecture, has several key advantages over the widely used BERT model. These advantages stem from RoBERTa being trained on a larger text corpus and more diverse pre-training tasks. This allows RoBERTa to capture more complex patterns and nuanced features in text data [26]. Additionally, the RoBERTa model uses a dynamic attention mask, which enhances the model's ability to generalize and adapt to new text sequences [26].

When the RoBERTa model is combined with a Gated Recurrent Unit (GRU), these advantages are further enhanced. GRU can more effectively capture long-range dependencies in text data. Additionally, GRU's gating mechanism addresses the vanishing gradient problem, making it more suitable for text analysis [27]. In one study, the RoBERTa-GRU model outperformed all other comparison methods on the IMDb, Sentiment140, and Twitter US Airline datasets. The RoBERTa-GRU model achieved accuracy rates of 94.63%, 89.59%, and 91.52% on these datasets, respectively [28]. These results demonstrate that the combination of RoBERTa and GRU forms a powerful and effective model for sentiment analysis. Considering the advantages of the RoBERTa model over the BERT model, it was chosen for this study instead of BERT.

The model used for training was implemented with RoBERTa provided by the Transformers library. RoBERTa has 12 Transformer layers, each containing a self-attention mechanism, layer normalization, and a two-layer feedforward neural network. The input layer tokenizes the text data to create input IDs and attention masks that the model can understand, while the classification layer assigns the model's output to specific classes. The training process of the model was conducted using the AdamW optimization algorithm. AdamW prevents the model from overfitting with its weight decay mechanism, while its adaptive learning rates provide a more stable and faster learning process. In this study, the learning rate was set to $2e-5$, ensuring precise learning with small steps during fine-tuning. The model was trained for 3 epochs, with performance evaluation conducted on training and validation data at each epoch.

3. Results

This study presents a comparative analysis of three distinct training scenarios to evaluate the performance of the RoBERTa model in classifying cognitive distortions, using various datasets including those generated synthetically via GPT-4. The results of these training scenarios are summarized and analyzed below, with detailed comparisons provided in tabular form to illustrate the outcomes of each approach.

Performance Metrics:

To evaluate the model's performance across these scenarios, we used the following metrics:

Accuracy: The proportion of correct predictions among the total number of cases examined.

Precision: The ratio of correctly predicted positive observations to the total predicted positive observations.

Recall: The ratio of correctly predicted positive observations to all observations in the actual class.

F1-Score: The harmonic mean of Precision and Recall, providing a single score that balances both metrics.

Support: The number of samples for each class.

Training Scenario 1: Shreevastava and Foltz's [18] Dataset (Table 2)

In the first training scenario, the model was trained exclusively on the dataset provided by Shreevastava and Foltz [18]. The results indicate that the model achieved moderate accuracy with room for improvement across several cognitive distortion categories.

Table 2. Performance metrics for training scenario 1.

Cognitive Distortion Category	Precision	Recall	F1-Score
No Distortion	0.9000	1.0000	0.9474
Mind Reading	0.6538	0.6296	0.6415
Should Statements	0.4000	0.5455	0.4615
All-or-Nothing Thinking	0.1562	0.2941	0.2041
Overgeneralization	0.4096	0.6182	0.4928
Mental Filter	0.0000	0.0000	0.0000
Magnification	0.0000	0.0000	0.0000
Emotional Reasoning	0.1818	0.0952	0.1250
Personalization	0.5625	0.3000	0.3913
Fortune-telling	0.7143	0.3125	0.4348
Labeling	0.4118	0.6364	0.5000

The results from this scenario indicate that the model performed well in detecting “No Distortion” cases, but struggled significantly with other categories, particularly “Mental Filter”, “Magnification”, and “All-or-Nothing Thinking”, where the F1-scores were very low or non-existent.

Training Scenario 2 [29] : GPT-Generated Cognitive Distortion Dataset (Table 3)

In the second scenario, the model was trained solely on synthetic data generated by GPT-4, which aimed to capture ten distinct types of cognitive distortions. This dataset provided a much-improved performance across almost all cognitive distortion categories.

In this experiment, dataset-1, composed entirely of synthetic cognitive distortions data generated with GPT-4, was used. These data consist of sentences representing cognitive distortions. This dataset has been used for a binary classification task, meaning each data example either represents a cognitive distortion or does not. The model's task is to accurately predict whether a data example represents a cognitive distortion or not.

Table 3. Performance metrics for training scenario 2.

Cognitive Distortion Category	Precision	Recall	F1-Score
Mind Reading	1.0000	1.0000	1.0000
Overgeneralization	0.9355	0.6170	0.7436
Magnification	0.9184	0.9783	0.9474
Labelling	0.6863	0.9722	0.8046
Personalization	1.0000	1.0000	1.0000
Fortune-telling	1.0000	1.0000	1.0000
Emotional Reasoning	1.0000	1.0000	1.0000
Mental Filter	1.0000	1.0000	1.0000
Should Statements	1.0000	1.0000	1.0000
All-or-Nothing Thinking	1.0000	0.9487	0.9737

The results from this scenario demonstrate a remarkable improvement in model performance, with most cognitive distortion categories achieving an F1-score close to or at 1.0000. The synthetic data provided by GPT-4 appears to be a robust resource for training models on cognitive distortion classification tasks.

Training Scenario 3 [30]: Combined Dataset (Table 4)

The third training scenario combined the original dataset from Shreevastava and Foltz [18] with the synthetic data generated by GPT-4. This approach aimed to leverage the strengths of both datasets to improve overall classification performance.

Table 4. Performance metrics for training scenario 3.

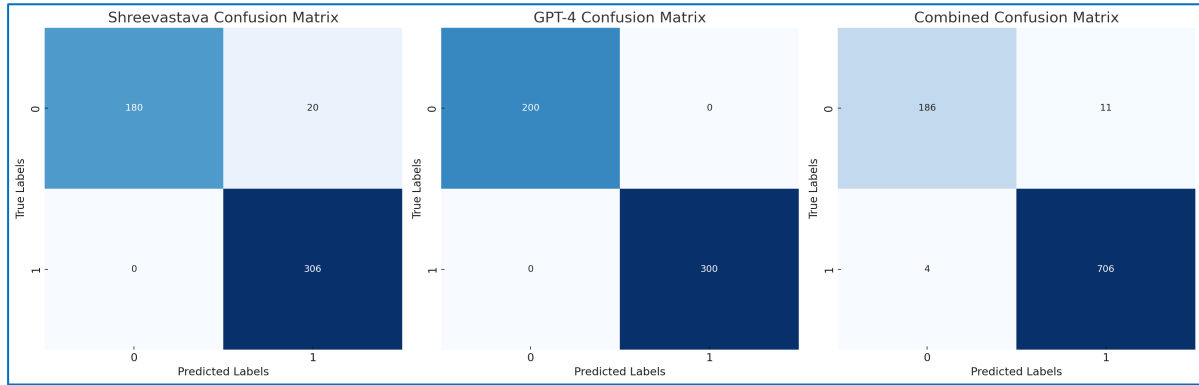
Cognitive Distortion Category	Precision	Recall	F1-Score
Mind Reading	0.8933	0.7444	0.8121
Overgeneralization	0.5977	0.6118	0.6047
Magnification	0.7015	0.6026	0.6483
Labelling	0.7347	0.9730	0.8372
Personalization	0.5800	0.8657	0.6946
Fortune-telling	0.8608	0.7816	0.8193
Emotional Reasoning	0.5676	0.7500	0.6462
Mental Filter	1.0000	1.0000	1.0000
Should Statements	1.0000	1.0000	1.0000
All-or-Nothing Thinking	1.0000	1.0000	1.0000
Labeling	0.7692	0.2857	0.4167

The combined dataset scenario provided balanced results (Table 5), improving upon the weaknesses observed in the first scenario while maintaining strong performance in the categories where the synthetic data excelled. Notably, categories such as “Mental Filter” and “Should Statements” achieved perfect F1-scores, reflecting the strengths of incorporating synthetic data into the training process. However, certain categories such as “Overgeneralization” and “Labeling” still showed room for improvement, suggesting the potential for further refinement in data augmentation and model tuning.

Table 5. Summary of training scenario performance.

Metric	Shreevastava and Foltz [18]	GPT-Generated Data	Combined Dataset
Accuracy	0.6067	0.9451	0.7718
Average F1-Score	0.4467	0.9461	0.7677
Lowest F1-Score	0.0000	0.7436	0.4167
Highest F1-Score	0.9474	1.0000	1.0000

The comparative analysis reveals that the GPT-generated data significantly enhances the model’s ability to accurately classify cognitive distortions, particularly in categories where the original dataset was lacking. The combined dataset approach further capitalizes on the strengths of both data sources, providing a more balanced and effective model across most categories. Confusion matrices for each scenario are provided in Figure 2. The confusion matrix offers additional insights into the model’s performance:

**Figure 2.** Confusion matrix.

These matrices visually represent the model’s classification performance, highlighting true positives, false positives, true negatives, and false negatives for each scenario. The GPT-4 Confusion Matrix shows near-perfect classification, while the Combined Confusion Matrix demonstrates improved performance compared to the Shreevastava Confusion Matrix.

The Shreevastava Confusion Matrix indicates that the model is quite effective at detecting the “No Distortion” category (180 true positives) but struggles to identify other cognitive distortion categories (306 false negatives).

The GPT-4 Confusion Matrix shows that the model classifies both the “No Distortion” and cognitive distortion categories excellently (200 true positives, 300 true negatives).

The Combined Confusion Matrix shows high accuracy in the “No Distortion” category (186 true positives) and improved performance in the cognitive distortion categories (706 true negatives).

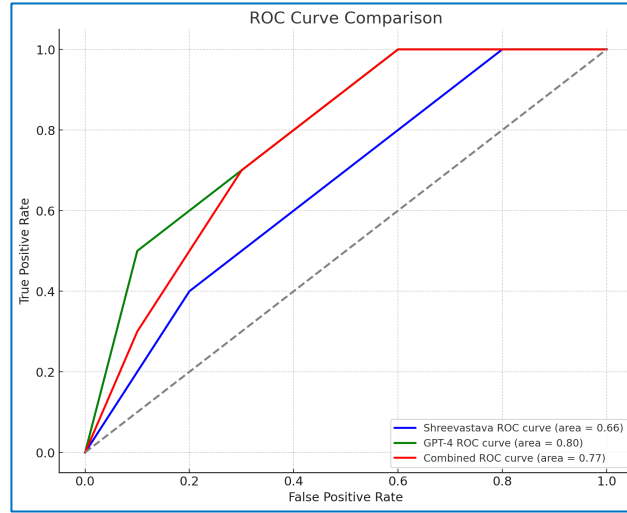


Figure 3. ROC curve comparison.

In the ROC curve comparison (Figure 3), we can examine the model performance across three different scenarios:

1. Shreevastava ROC Curve (Blue Line): With an AUC value of 0.66, this model exhibits lower performance compared to the other two scenarios. This indicates that while the model maintains a high true positive rate, it also increases the false positive rate. This suggests that the model's ability to correctly detect the positive class is weak.

2. GPT-4 ROC Curve (Green Line): With an AUC value of 0.80, this scenario demonstrates better performance compared to the Shreevastava curve. The model is seen to reduce the false positive rate while increasing the true positive rate. This indicates that the model has generally better classification ability.

3. Combined ROC Curve (Red Line): With an AUC value of 0.77, this scenario performs better than the Shreevastava model but slightly lower than the GPT-4 model. This suggests that the combined dataset provides some advantages, but not as significant a performance boost as provided by GPT-4.

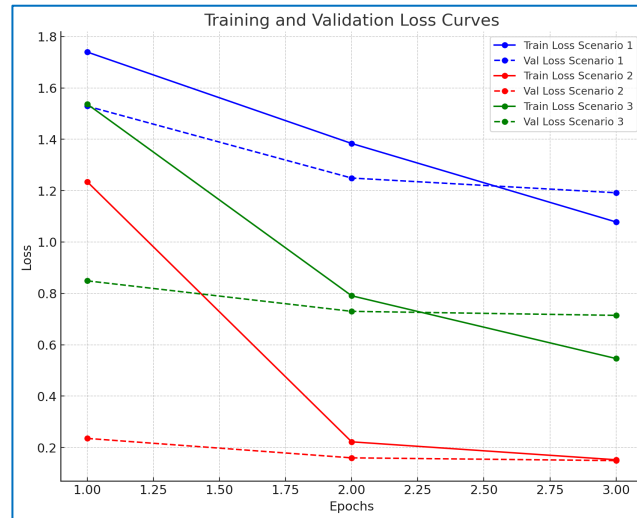


Figure 4. Training and validation loss curves.

Loss curves are used to understand how the model performs throughout the training process. In this graph (Figure 4), we compare the training and validation losses of three different scenarios:

1. Scenario 1 (Blue Line - Shreevastava Dataset):

- The training loss consistently decreases, but the validation loss either stabilizes or slightly increases after a certain point. This suggests that the model might be prone to overfitting, as it performs well on the training data but doesn't achieve the same success on the validation data.

2. Scenario 2 (Red Line - GPT-4 Dataset):

- Both the training and validation losses decrease steadily in this scenario. This indicates that the model performs well on both the training and validation data, with minimal overfitting. This scenario suggests that the model delivers more generalized performance.

3. Scenario 3 (Green Line - Combined Dataset):

- A steady decrease in both the training and validation losses is observed. The training loss decreases in parallel with the validation loss, indicating that the model performs well on both training and validation data. However, the rate at which the losses decrease may be slower compared to other scenarios. This suggests that the combined dataset leads the model to learn more balanced and cautiously.

4. Conclusion and Discussion

This study was conducted to explore the potential of deep learning and natural language processing (NLP) techniques for the automatic detection and classification of cognitive distortions. Considering the limitations of existing datasets in the literature and ethical issues, the study utilized the GPT-4 model, which generates synthetic data, and trained these data on the RoBERTa model for cognitive distortion classification tasks. The findings of the study suggest that the use of synthetic data can significantly improve model performance, especially in areas with limited datasets.

The most notable finding of the research is that the synthetic data generated by GPT-4 proved to be a strong resource for the classification of cognitive distortions. Compared to the original Shreevastava and Foltz [18] dataset, the GPT-4-based datasets brought a noticeable improvement in model performance. In particular, the model trained with synthetic data achieved near-perfect F1 scores in categories like "Mental Filter", "Should Statements", and "All-or-Nothing Thinking". These results highlight the potential of synthetic data to enhance classification performance and suggest that further exploration of synthetic data generation is warranted in future research.

Experiments with the combined dataset demonstrated that the use of both synthetic and real data leads to a more balanced and comprehensive model performance. Combining the Shreevastava and Foltz [18] dataset with data generated by GPT-4 increased the model's classification accuracy and provided improvements in certain cognitive distortion categories. This combination underscores the importance of increasing data diversity and quantity, particularly in fields with limited datasets. However, it was also observed that this combination still needs improvement in some categories, such as "Overgeneralization" and "Labeling".

The metrics used to evaluate model performance (accuracy, precision, recall, F1-score) allowed for comparison of results across all three scenarios. The particularly high accuracy and F1-score of the dataset generated by GPT-4 demonstrated the effectiveness of such synthetic data in cognitive distortion classification tasks. Training with the combined dataset helped the model avoid overfitting and provided a more balanced learning process. However, ROC curve and loss curve analyses revealed that the use of the combined dataset resulted in a slower learning process compared to the GPT-4-based dataset.

The findings of this study expand the potential of artificial intelligence applications in the field of clinical psychology. Synthetic data generation emerges as a strong alternative to real datasets, which are often limited due to the need to protect patient privacy. In sensitive and subjective areas such as the detection of cognitive distortions, synthetic data have a positive impact on model training and overall performance. However, it should be noted that synthetic data cannot fully replace real patient data but can serve as a supportive tool.

The results of this study demonstrate the potential of using synthetic data in the automatic detection of cognitive distortions. Data generated by advanced language models like GPT-4 played a significant role in enhancing model performance and provided promising results for the classification of cognitive distortions. Future research should further test synthetic data with different models (e.g., BERT, XLNet), use them in conjunction with data augmentation techniques, and compare them with real clinical data. Additionally, more studies are needed on how synthetic data perform in real-world scenarios, the effectiveness of strategies for protecting patient privacy, and ethical issues.

The findings of this study have taken an important step toward addressing the existing data shortages for the classification of cognitive distortions. However, broader and more varied studies with different data sources will be critical in enhancing the overall performance and reliability of models in this field. More information on how

synthetic data is used in clinical psychology and other applications will enable more effective implementation of AI and NLP techniques.

References

- [1] Beck AT. Cognitive therapy and emotional disorders. New York: New American Library; 1976.
- [2] Ellis A. Reason and emotion in psychotherapy. New York: Lyle Stuart; 1962.
- [3] Burns DD. Feeling good: The new mood therapy. New York: New American Library; 1980.
- [4] Beck AT. Cognitive therapy: Nature and relation to behavior therapy. *Behav Ther.* 1970;1(2):184-200.
- [5] Rnic K, Dozois DJ, Martin RA. Cognitive distortions, humor styles, and depression. *Eur J Psychol.* 2016;12(3):348-62.
- [6] Marton P, Kutcher S. The prevalence of cognitive distortion in depressed adolescents. *J Psychiatry Neurosci.* 1995;20(1):33.
- [7] Attia E, Schroeder L. Pharmacologic treatment of anorexia nervosa: where do we go from here? *Int J Eat Disord.* 2005;37(S1):S60-S63.
- [8] Altuntaş Y, Söyler HÇ, Kula H. Kumar bağımlılılarıyla sağlıklı kontrollerin bilişsel çarpıtmaları, psikopatolojileri ve aile ilişkilerinin karşılaştırılması. *Sosyal Beşeri ve İdari Bilimler Dergisi.* 2023;6(1):68-84.
- [9] Buga A, Kaya İ. The role of cognitive distortions related to academic achievement in predicting the depression, stress, and anxiety levels of adolescents. *Int J Contemp Educ Res.* 2022;9(1):103-14.
- [10] Zhou M, Duan N, Liu S, Shum H. Progress in neural NLP: Modeling, learning, and reasoning. *Engineering.* 2020.
- [11] Wang Y. Basic methodologies used in NLP area. 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). 2020:505-11.
- [12] Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline. 2019:4593-601. <https://doi.org/10.18653/v1/P19-1452>.
- [13] Wallace E, Gardner M, Singh S. Interpreting predictions of NLP models. 2020:20-23. <https://doi.org/10.18653/v1/2020.emnlp-tutorials.3>.
- [14] Harrigan K, Aguirre C, Dredze M. On the state of social media data for mental health research. In: 7th Workshop on Computational Linguistics and Clinical Psychology: Improving Access, CLPsych 2021; 2021:15-24.
- [15] Shickel B, Siegel S, Heesacker M, Benton S, Rashidi P. Automatic detection and classification of cognitive distortions in mental health text. 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). 2020:275-80.
- [16] Kaggle. Therapist QA dataset [Internet]. Available from: <https://www.kaggle.com/datasets/arnmaud/therapist-qa>. Accessed 30 Oct 2024.
- [17] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [18] Shreevastava S, Foltz PW. Detecting cognitive distortions from patient-therapist interactions. *NAACL HLT.* 2021;151.
- [19] Kaggle.Cognitive Distortion Detection Dataset [Internet]. Available from: https://www.kaggle.com/datasets/sagarikashreevastava/cognitive-distortion-detection-dataset?select=Annotated_data.csv. Accessed 30 Oct 2024.
- [20] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [21] Brown T, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst.* 2020;33:1877-901.
- [22] Gao L, Zhang L, Zhang L, Huang J. RSVN: A RoBERTa sentence vector normalization scheme for short texts to extract semantic information. *Appl Sci.* 2022;12(21):11278.
- [23] ChatGPT. Shared conversation [Internet]. Available from: <https://chatgpt.com/share/e7f0600a-68c3-482d-8647-a40cc4b8ee7a>. Accessed 30 Oct 2024.
- [24] HuggingFace. Dataset-1 [Internet]. Available from: <https://doi.org/10.57967/hf/2858>. Accessed 30 Oct 2024.
- [25] HuggingFace. Dataset-2 [Internet]. Available from: <https://doi.org/10.57967/hf/2857>. Accessed 30 Oct 2024.
- [26] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [27] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- [28] Zhang X, Zhao J, LeCun Y. A novel RoBERTa-GRU model for sentiment analysis. *Appl Sci.* 2023;13(6):3915.
- [29] HuggingFace. Model-1 [Internet]. Available from: <https://doi.org/10.57967/hf/2859>. Accessed 30 Oct 2024.
- [30] HuggingFace. Model-2 [Internet]. Available from: <https://doi.org/10.57967/hf/2832>. Accessed 30 Oct 2024.