Exploring Generative A.I. for Addressing Class Imbalance in Cognitive Distortion Predictive Models

Contact Information:
Dept. of Mathematics and Computer Science
Science Building S121D

Connor Mulholland, BSCS Candidate; Paula Lauren, PhD, Professor, CoAS

Abstract

Modern tools for natural language generation may enable Artificial Intelligence (AI) models to be better trained on imbalanced data by generating records for the minority classes. Specifically, the psychoanalytic cognitive distortion data is based on the irrational or biased ways of thinking which can contribute to negative emotions and behaviors, which are crucial in many types of therapy. We compare various types of Generative AI models for generating new records and the current limitations of these new models. We find that pretrained Sentence-Bidirectional Encoder Representations from Transformers (Sentence-BERT) embeddings (i.e., multilingual-e5-large-instruct) used to train a Support Vector Machine (SVM) classifier model yields the best binary results with an F1-score of 0.756. The addition of generated data using the Mistral-7B-Instruct-v0.2 model with recursive data generation on the binary classification task resulted in a marginal boost in performance with an F1-score of 0.765 using the same Sentence-BERT embeddings with SVM classification model.

Introduction

Cognitive distortions are irrational or biased ways of thinking that can contribute to negative emotions and behaviors. Cognitive-behavioral therapy (CBT) is a common therapeutic approach that aims to help individuals identify and challenge these distortions. Detecting cognitive distortions from patient-therapist interactions is a challenging task that has been addressed in recent research, and could be crucial in many types of therapy to help therapstis and individuals identify and challenge these distortions. Prior research has attempted to use a dataset of human-labeled patient-therapist interactions to train a model to detect cognitive distortions[1]. However, the dataset is highly imbalanced [fig. 1], with the majority of the data belonging to the non-distorted class, which poses a challenge for predictive modeling. Generative AI models have the potential to address class imbalance by generating new data for the minority classes. In this study, we explore the use of generative AI models to address class imbalance in cognitive distortion predictive models by comparing the performance of different generative AI models, vectorizers, and classification algorithms on the task of detecting cognitive distortions from patient-therapist interactions. We consider both binary and multi-class classification tasks and evaluate the models using the weighted F1-score as the performance metric.

Initial Data Analysis

We use a dataset of patient-therapist interactions to train our models. The dataset contains 10 classes of cognitive distortions, with the majority of the data belonging to the non-distorted class. We first treat the problem as a binary classification task, where we combine all the distorted classes into a single class and compare the performance of different classification algorithms. We use the Term Frequency-Inverse Document Frequency (tf-idf) vectorizer to convert the text data into numerical features and train a Linear Support Vector Classifier (LinearSVC) on the data. The best initial F1-score for the binary classification task is 0.73 using tf-idf and LinearSVC. We then treat the problem as a multi-class classification task and compare the performance of different classification algorithms. The best initial F1-score for the multi-class classification task is 0.21 using tf-idf and LinearSVC. The addition of the non-distorted class to the multi-class classification task resulted in a marginal improvement in performance, with an F1-score of 0.32 using tf-idf and LinearSVC.

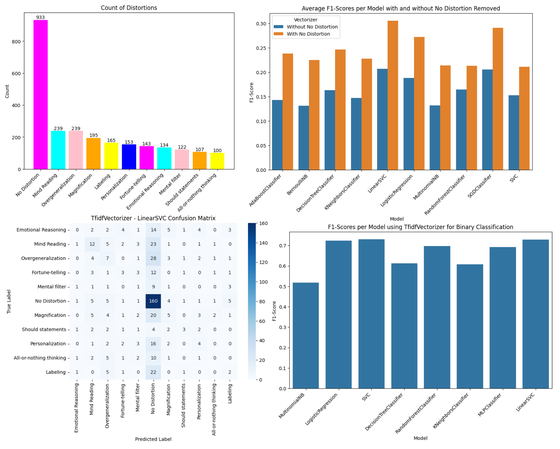

Figure 1: (Top-Left) Class Imbalance; (Top-Right) Average F1-Scores with and without No Distortion data; (Bottom-Left) Confusion Matrix of best model with No Distortion data; (Bottom-Right) F1 Scores for Binary Classification

Sentence-BERT Embeddings

Sentence-BERT (SBERT) is a variant of the BERT model that is specifically trained to generate sentence embeddings. We compare the performance of different SBERT models on the task of detecting cognitive distortions from patient-therapist interactions [fig. 2]. We use the Sentence Transformer library to generate SBERT embeddings for the text data and train different classification algorithms on the embeddings. The best F1-score for the multi-class classification task is 0.262 using the "intfloat/multilingual-e5-large-instruct" (Multilingual) SBERT model and LinearSVC. The best F1-score for the binary classification task is 0.756 using the same SBERT model with SVM.

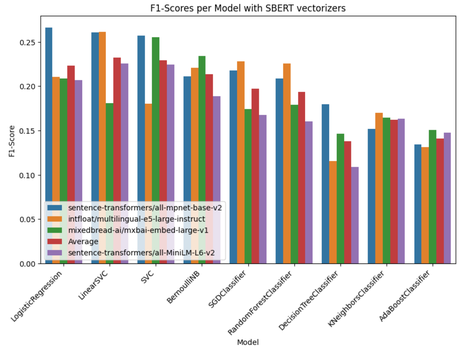

Figure 2: F1-Scores for SBERT Models

Generative AI Models

We explore the use of generative AI models to address class imbalance in cognitive distortion predictive models. For the first set of experiments, we use the "mistralai/Mistral-7B-Instruct-v0.2" (Mistral) model to generate new records for the minority classes in the dataset. We used a recursive technique to generate new records taking the generated data and appending the same question and feeding it back to the model, extracting and storing the generated data. We then used a sampling technique by pulling three random samples from the generated data, acting as if the model generated that data, and asking the model to generate a fourth sample similar to the three and including the distortion with it. We used this sampling technique on GPT-4 as it was the most inexpensive, the fastest, and the cleanest method to generate data, although later we found that it was not the most effective. Results show that the addition of generated data using the Mistral model with recursive data generation on the binary classification task resulted in the highest F1-score boost for both the binary and multi-class classification tasks, with the ladder results shown below [fig. 3].

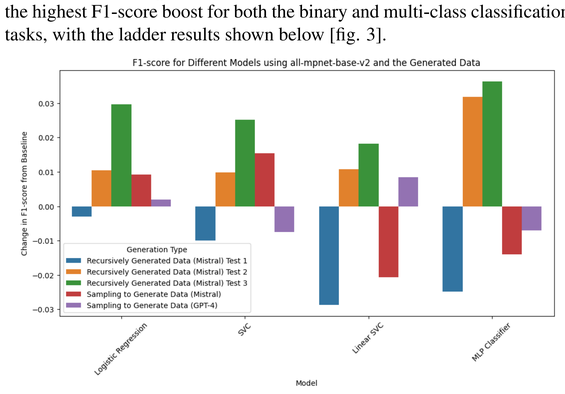

Figure 3: Average Difference in F1-Score from baseline tests (no added generated data) for Different Generation Techniques and Classification Tasks

Results and Evaluation

The best F1-score for the multi-class classification task is 0.298 using the Multilingual SBERT model and MLPClassifier with hyperparameter tuning and recursive generated Mistral data 2. The best F1-score for the binary classification task is 0.765 using the Multilingual SBERT model and SVM with hyperparameter tuning and recursive generated Mistral data 2.

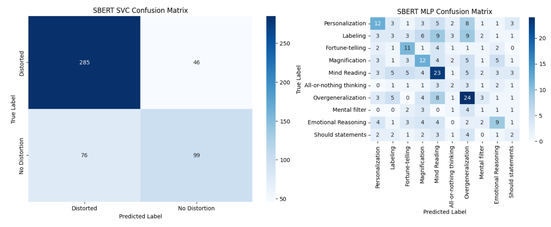

Figure 4: Confusion Matrix of best models' performance on the binary (left) and multi-class (right) classification task

Discussion

The inter-annotator agreement (IAA) score posed a challenge in this study, as the human annotators had difficulty agreeing on the labels for the data (Table 1). This may have affected the performance of the generative AI models. The best model we have found for detecting cognitive distortions from patient-therapist interactions is the Multilingual SBERT model with SVM classification and hyperparameter tuning, achieving an F1-score of 0.765. The best model we have found for classifying the data into the 10 classes is the Multilingual SBERT model with MLPClassifier and hyperparameter tuning, achieving an F1-score of 0.298.

Future work includes exploring different methods of text generation and categorization approaches incorporating other large language models, such as Claude and Gemini, as well as refining prompt engineering techniques to generate more data, potentially yielding higher F1-scores [fig. 3].

Cognitive Distortion	Percentage of Records containing a Secondary Distortion
All-or-Nothing Thinking	23%
Emotional Reasoning	27%
Fortune-Telling	29%
Labeling	38%
Magnification	26%
Mental Filter	34%
Mind Reading	17%
Overgeneralization	22%
Personalization	32%
Should statements	20%

Table 1: Analysis of Secondary Distortions in Cognitive Distortions

Conclusion

In conclusion, the addition of generated data using the Mistral-7B-Instruct-v0.2 model with recursive data generation on the binary classification task resulted in a boost of 0.01 with an F1-score of 0.765 using the same SBERT embeddings with SVM classification model. On the other hand, the same additional data on the multi-class classification task resulted in a boost of 0.036 with an F1-score of 0.298 using the same SBERT embeddings with MLPClassifier and hyperparameter tuning. The binary classification task is useful for a future system and would be a good first step for further analysis by a human, but there is lots more work to be done within this field and this project.

Acknowledgements

We would like to acknowledge the support of the QUEST Program and Dr. Lauren for their guidance and assistance throughout this research project.

References

[1] Shreevastava, S., & Foltz, P. (2021, June). Detecting cognitive distortions from patient-therapist interactions. ACL Anthology. https://aclanthology.org/2021.clpsych-1.17/