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. The paper who first utilized this data[1] uses a dataset of patient-therapist interactions to train a model to detect cognitive distortions. However, the dataset is highly imbalanced, with the majority of the data belonging to the non-distorted class. This class imbalance poses a challenge for training accurate predictive models. Generative AI models have the potential to address class imbalance by generating synthetic 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. We compare the performance of different generative AI models 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 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 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 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.

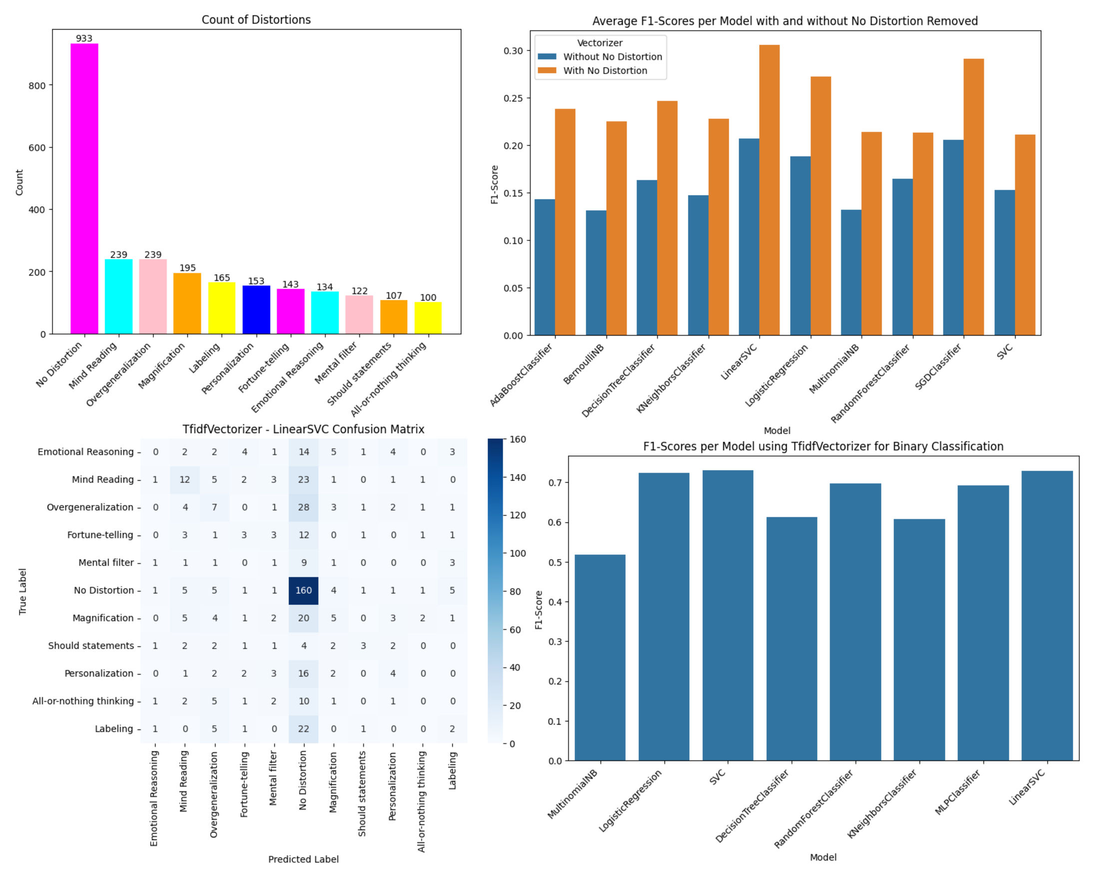

Figure 1: (Top-Left) Class Imbalance, (Top-Right) Average F1-Scores with and without No Distortion data, (Bottom-Left) Confusion Matrix of best model with No Distortion data, (Bottom-Right) F1 Scores for Binary Classification

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. 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" SBERT model and LinearSVC. The best F1-score for the binary classification task is 0.756 using the "intfloat/multilingual-e5-large-instruct" SBERT model with SVM.

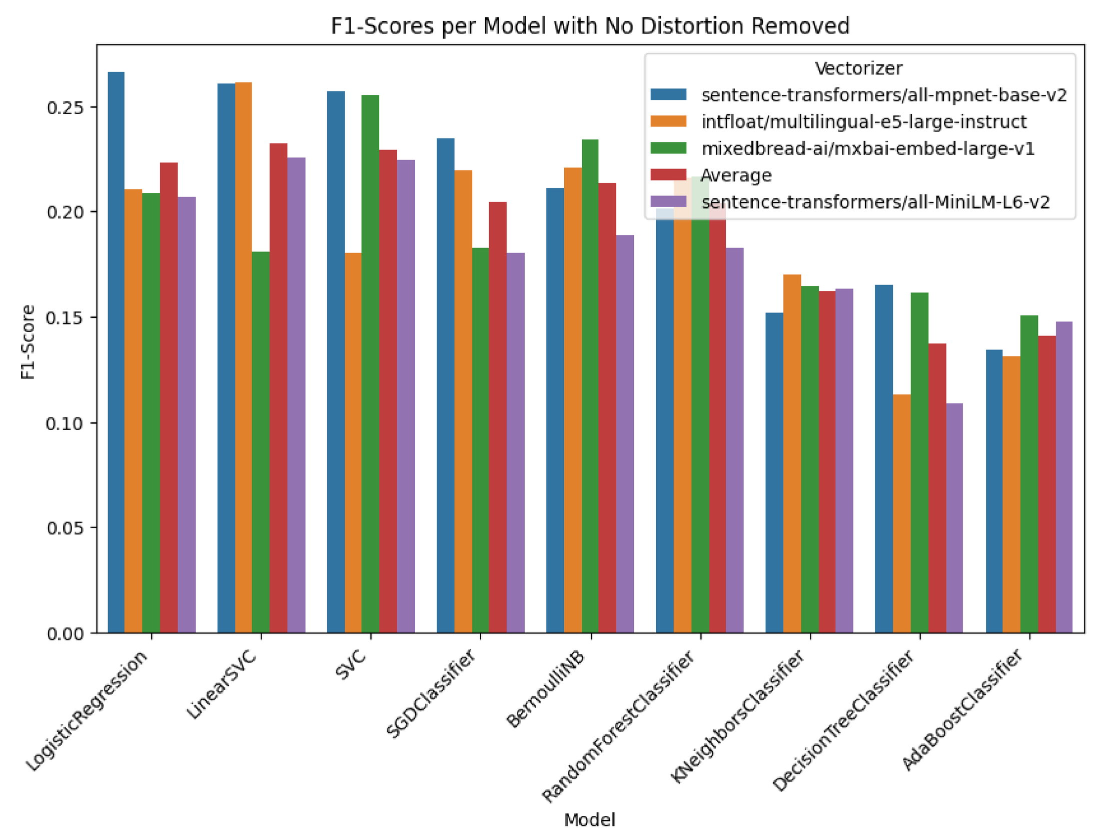

Figure 2: F1-Scores for SBERT Models

Generative AI Models

We explore the use of generative AI models to address class imbalance in cognitive distortion predictive models. We compare the performance of different generation techniques, including fine-tuning GPT-2 models and recursive data generation. We also compare the performance of different classification algorithms on the generated data. The best F1-score for the multi-class classification task is 0.285 using the "intfloat/multilingual-e5-large-instruct" SBERT model and MLPClassifier with hyperparameter tuning. The best F1-score for the binary classification task is 0.765 using the "intfloat/multilingual-e5-large-instruct" SBERT model and SVM with hyperparameter tuning. 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 SBERT embeddings with SVM classification model.

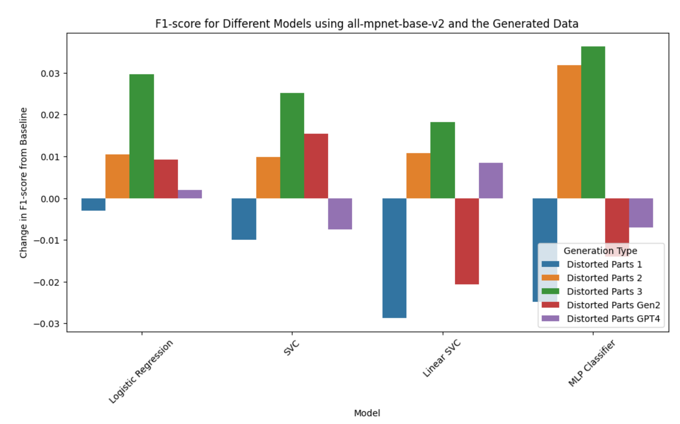

Figure 3: Fine Tuning Architecture

The Fine-Tuning Process

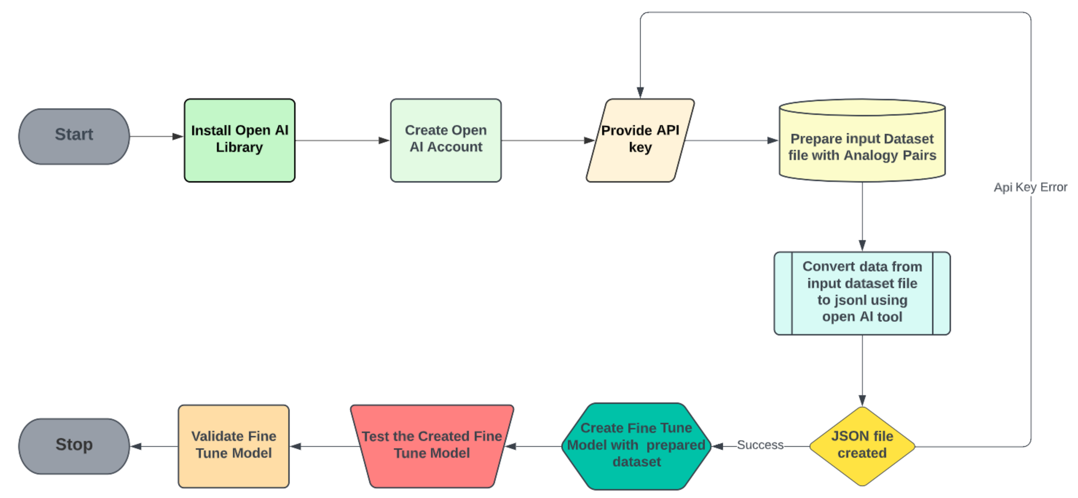

Figure 4: Fine-Tuning Process Step-by-Step

The below illustration depicts the outcome of the fine-tuned model for the example elaborated in Figure 3.

Bananas are more closely related to avocados than to limes because both belong to the same plant family, called the Musaceae family.

As illustrated in Figure 6, fine-tuning a pre-trained language model can enhance its performance on a given task or dataset. By adjusting the pre-trained model on a specific dataset, the model can learn the nuances and subtleties of the data and perform more accurately and meaningfully.

In addition to the example shown in Figure 6, there are several other examples where fine-tuning has demonstrated impressive results. These instances are shown in the table below.

Prompt	GPT	Fine-Tuned
Yellow is to banana and green is to	Lime	Avocado
Computer is to software and brain is to	Hardware	Cognition
Knife is to cooking as screwdriver is to	DIY projects	Repair
Shovel is to gardening as keyboard is to	Computing	Typing
Microphone is to sound as camera is to	Light	Images
Car is to engine as computer is to	Keyboard	Processor
Glacier is to iceberg as stream is to	Rock	River
Rust is to decay as charcoal is to	Burning	Ash
Forgiveness is to healing as medicine is to	Health	Recovery

Table 1: Analysis of Fine-Tuned Model

F1-Score Results Analysis

- The challenge at hand is that OpenAI, a leading AI research organization, will be offering a limited-time free subscription to their API service. This subscription will only be available for a duration of three months and will come with a limited amount of credits for using the API and fine-tuning it. Users may need to carefully plan and prioritize their use of the API within the allotted time and credit.
- During the process of fine-tuning a learning model, users may have to wait for an unpredictable amount of time for the model to be fully

trained. This wait time can vary depending on several unknown factors, and can range from minutes to hours or even days. This unpredictability can be a frustrating and time-consuming challenge for users who need to balance their fine-tuning activities with other tasks and responsibilities.

Conclusion

In conclusion to our research, fine-tuning is a crucial process that can significantly improve the performance of pre-trained NLP models on specific tasks or domains. Our findings show that the pre-trained GPT models can benefit from fine-tuning, as they are trained on massive datasets with broad contexts and may not generalize well on specific tasks or domains. Fine-tuning allows the models to adapt to the target task or domain by learning the relevant patterns and relationships from a smaller dataset, leading to higher accuracy and efficiency.

Our analysis of the fine-tuning results shows that the fine-tuned GPT models outperformed their pre-trained large language model and achieved better results in the analogy task. This indicates the effectiveness of the fine-tuning process in improving the models' performance on specific tasks or domains.

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] Detecting Cognitive Distortions from Patient-Therapist Interactions. (2021). *ACL Anthology*. Retrieved from https://aclanthology.org/2021.clpsych-1.17.pdf