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.

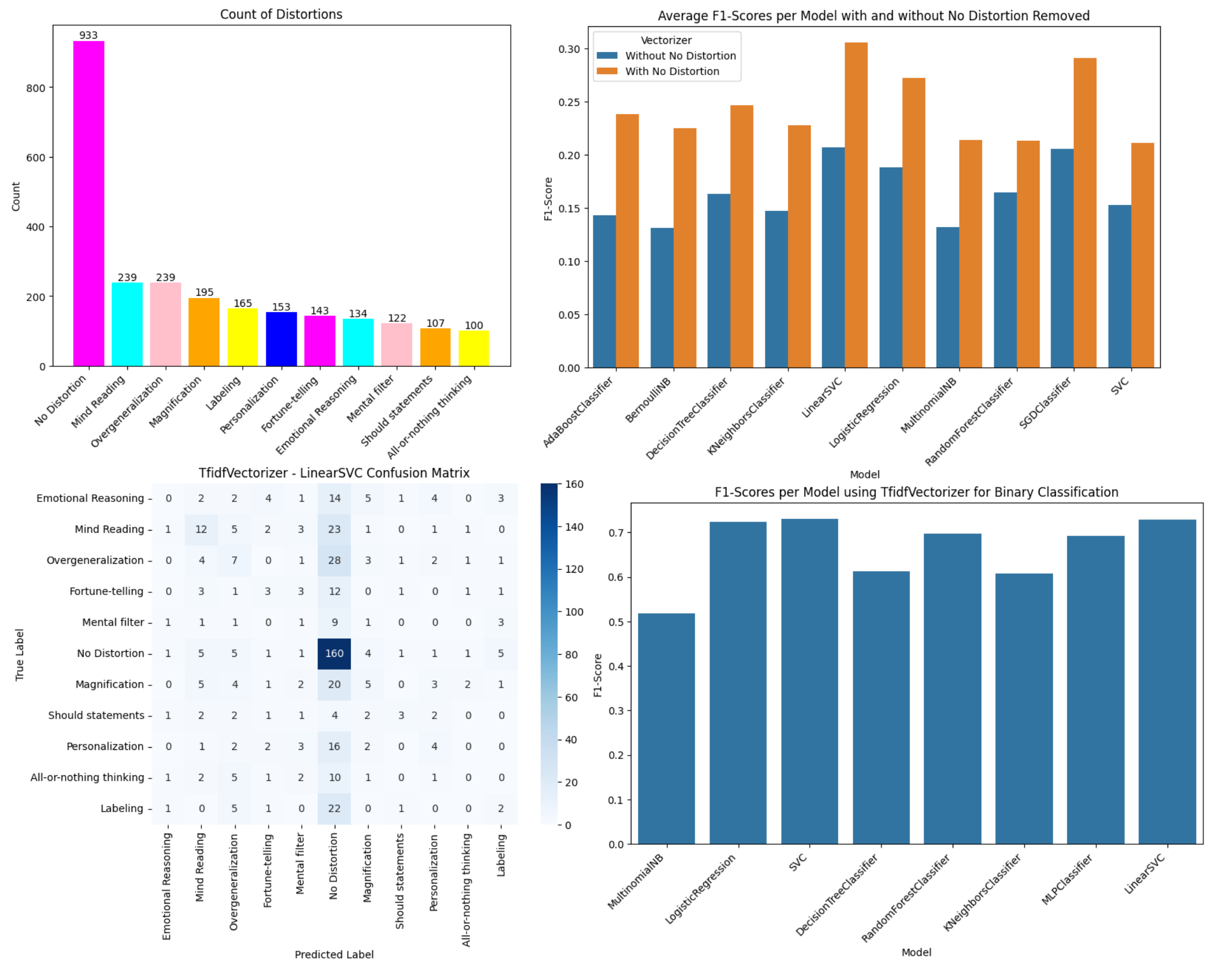

Figure 1: (Top-Left) Class Imbalance, (Top-Right) Average F1-Scores with and without No Distortion data, (Bottom-Left) Confusion Matrix of best model with No Distortion data, (Bottom-Right) F1 Scores for Binary Classification

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.

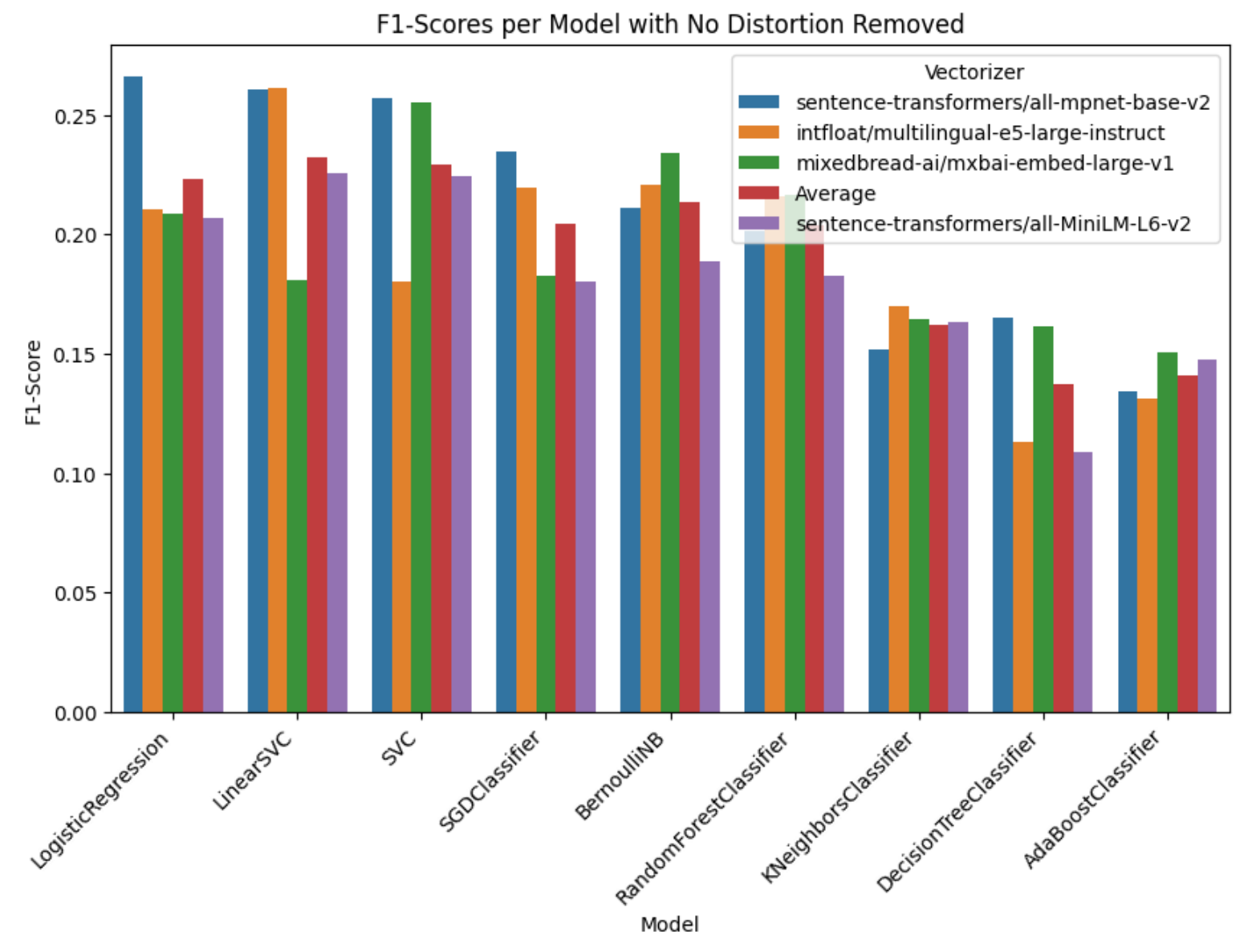

Figure 2: F1-Scores for SBERT Models

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.

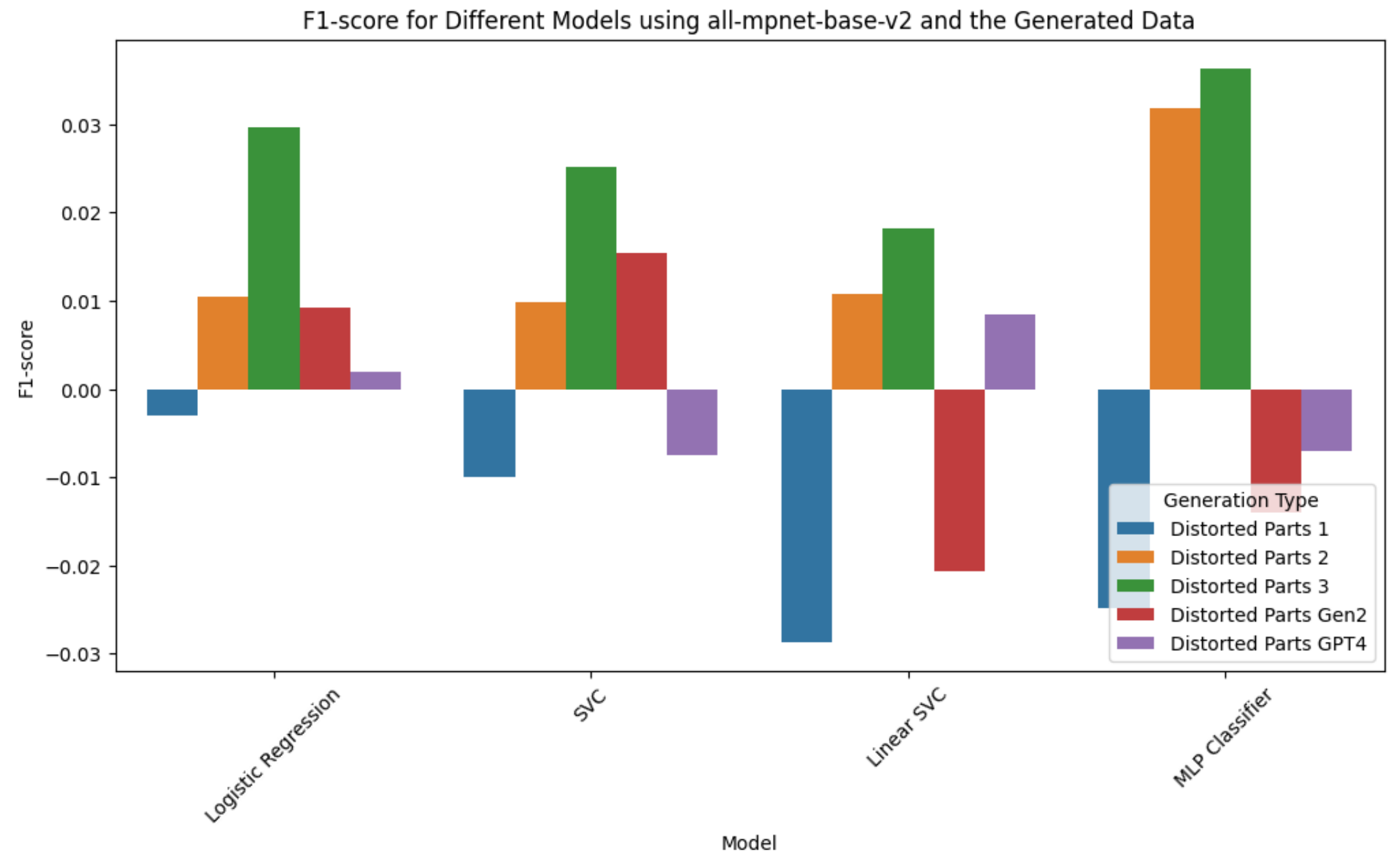

Figure 3: Fine Tuning Architecture

Results and Evaluation

The best F1-score for the multi-class classification task is 0.298 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.

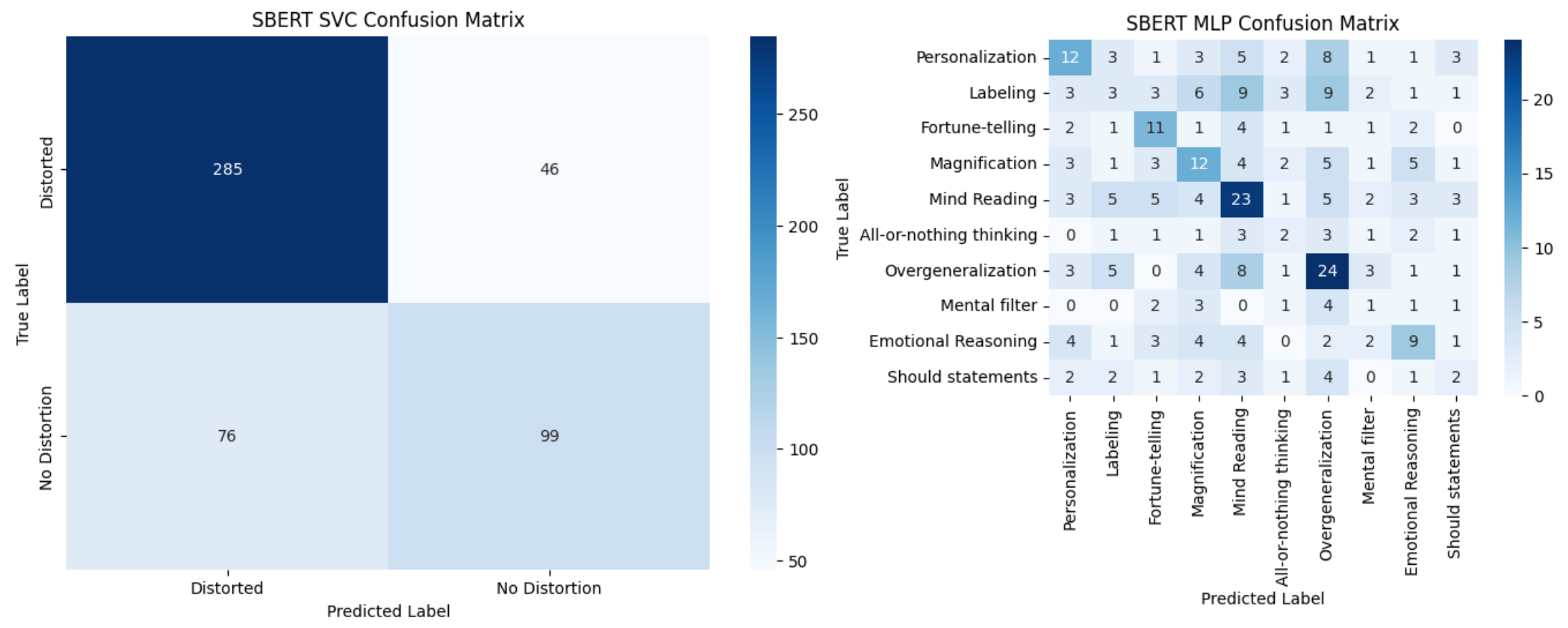

Figure 4: F1-Scores for Final Models

Discussion

The limitations of this study include the inter-annotator agreement (IAA) score, which posed a challenge for the generative AI models. If humans cannot agree on the labels for the data, it would pose a major challenge for classification AI. The original paper's method of separating the data into distorted and non-distorted classes was validated in this study. 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.

Cognitive Distortion	Percentage of Records containing a Secondary Distortion
All-or-Nothing Thinking	0.23
Emotional Reasoning	0.27
Fortune-Telling	0.29
Labeling	0.38
Magnification	0.26
Mental Filter	0.34
Mind Reading	0.17
Overgeneralization	0.22
Personalization	0.32
Should statements	0.20

Table 1: Analysis of Secondary Distortions in Cognitive Distortions

Conclusion

In conclusion, the best model for detecting cognitive distortions from patient-therapist interactions is the "intfloat/multilingual-e5-large-instruct" SBERT model with SVM classification and hyperparameter tuning, which achieved an F1-score of 0.765. 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. Future work includes exploring different methods of text generation and categorization, such as fine-tuning GPT-2 models and using GPT-4 to categorize data. The binary classification task is useful for a future system and would be a good first step for further analysis by a human.

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