RWTH AACHEN UNIVERSITY

Aryan Arora

University of Pennsylvania

Explainable AI and Text Analytics

RWTH Aachen University
Exaia Technologies
Elma Kerz, Daniel Wiechmann, Yu Qiao

UROP – Undergraduate Research Opportunities Program

Aachen
21 July 2025

Table of Contents

Abstract

We investigate the use of large language models (LLMs) to generate synthetic text data for **bipolar disorder** and control user personas, and apply explainable AI (XAI) techniques to analyze the results. The project's aim is to **augment mental health text datasets** with realistic synthetic posts and evaluate their fidelity and utility. We implemented a two-stage approach: (1) using zero-shot and few-shot prompting (with both predefined attribute personas and inferred personas) to create a diverse corpus of Reddit-style posts, and (2) extracting rich linguistic features (344 dimensions via CYMO analytics) to compare synthetic vs. real data distributions and inform classification models. Key findings indicate that **high-quality synthetic data can vaguely mirror real user content** in terms of linguistic features, especially when using carefully engineered prompts and persona designs. Furthermore, incorporating the synthetic posts into a bipolar-detection classifier **marginally improved** its performance on real-world data, demonstrating the **value of synthetic augmentation** for enhancing model accuracy and generalization. Overall, the project showcases how explainable text analytics and synthetic data generation together can address data sparsity and bias in mental health NLP, while providing transparency into model decision factors.

Introduction

Bipolar disorder (BPD) is a severe mood disorder characterized by cyclic episodes of mania and depression, with roughly 2.8% of adults meeting criteria for in a single year and 4.4% of people experiencing it at some point in their lives[1]. The condition causes severe role impairment to individuals in social media, work, and family domains, imposing an economic burden exceeding $190 billion per year[2]. Individuals diagnosed with BPD have a suicide rare 20 to 30 times higher than that of the general population.[3]

Traditional methods used as screening tools for the disorder include the the 13-item Mood Disorder Questionnaire (MDQ) and the Structured Clinical Interview for DSM (SCID), which relies on the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, 5th Edition), the standard classification system used by mental health professionals to diagnose and describe psychiatric conditions[4]. However, these screening tools have significant limitations. The MDQ has low sensitivity outside psychiatric settings, only identifying around 28% of true bipolar cases in the general population, and the SCID's time and resource demands contribute to frequent misdiagnoses, leading to an average diagnostic delay of ~10 years[5].

To find a solution, we look towards social media. As a prominent platform where people engage in discussions and share information about their lives, social media serves as an important source of information related to mental health conditions for users. Language from social media, such as first-person narratives from Reddit posts, can be used as digital biomarkers for mental health, helping to identify users in need of clinical support and broadening our understanding of the disorder as a whole.

Supervised machine learning models that have been applied to online user-generated text have shown promise in detecting mental health conditions. However, their effectiveness is limited by (i) demographic imbalances in available training data and (ii) limited training data for underrepresented groups, both of which can introduce bias in model prediction. For instance, a classifier trained predominantly on data from young, English-speaking users might fail to generalize to older users or those from different cultural backgrounds. Synthetic data generation offers a compelling solution to this problem– by simulating realistic posts across diverse personas, we can augment unbalanced data, correct biases in sampling, and improve model generalization.

Recent advances in prompt-based generation include zero-shot/few-shot strategies guided by structured persona definitions (Li et. al 2023)[6] and large-scale persona hubs (Ge et. al 2024)[7]. These approaches enable targeted synthesis of data without fine tuning. However, rigorous evaluation of the realism, psycholinguistic fidelity, and downstream utility of such synthetic text remains underexplored in the mental health domain. To address this gap, we implement a two stage pipeline. First, we generate synthetic Reddit-style posts for bipolar and control personas using zero-shot and few-shot prompting techniques. Second, we extracted high-resolution linguistic features using CYMO[8], a next-generation NLP tool for text analytics. We used these features to qualitatively and quantitatively compare the distributions of real and synthetic text, and to train classifiers both to distinguish synthetic from real data and to detect bipolar disorder.

This report details our methods and findings.

Project Description

Objectives and Scope

- O1: Implement zero-shot synthetic data generation
- O2: Extend to few-shot and inferred few-shot generation
- O3: Extract CYMO features to compare real and synthetic distributions
- O4: Evaluate real-synthetic distinguishability and downstream detection performance

Pipeline Overview

- Persona Design & Prompt Engineering (attribute-controlled vs. inferred/zero-shot vs. few-shot).
- Synthetic Data Generation (8 datasets with Mistral-7B/Meta-Llama-70B).
- Sociodemographic Detection (manual + automated regex-based extraction).
- Feature Extraction (CYMO sentence-level → text-level aggregation).
- Fidelity & Realism Analysis (UMAP, JSD, WD, density plots).
- Classifier Evaluation (real vs. synthetic; bipolar detection)

Synthetic Data Generation

Generating high-quality synthetic data requires outputs that resemble real-world behavior in both content and style. These qualities are especially important when the synthetic data is intended to reflect identity-linked language, such as Reddit comments authored by individuals with Bipolar Disorder. Our approach aims to maximize realism and diversity in generated text by exploring different combinations of **prompting techniques** and **persona construction methods.**

We focus on two primary components of the generation process: the **prompting method** (how we instruct the language model, e.g. zero-shot vs. few-shot) and the **persona design** (how we construct the identity of the synthetic "author," e.g. via predefined attributes or inferred persona descriptions). Combining these two dimensions yields four distinct generation settings: (1) attribute-controlled zero-shot, (2) attribute-controlled few-shot, (3) inferred zero-shot, and (4) inferred few-shot.

In all settings, the model is instructed to write a Reddit comment as if by a user with bipolar disorder, but **without explicitly mentioning** any mental health diagnoses in the text (to ensure authenticity of expression). Each setting produces a dataset of 5,400 comments, corresponding to 5,400 unique synthetic personas. Below, we detail each generation method and setting:

a) Zero-shot Synthetic Data Generation Method

Following Li et al. (2023), our zero-shot prompting method assumes no direct examples of real data are given to the model. A prompt consists of several components appended together:

- **Context Prompt** sets persona slots:
  - Ex: "Your age is {age}. You are a {gender} from {nationality}, occupation {occupation}, interests {interests}. You are diagnosed with Bipolar Disorder."
- **Generation Prompt** instructs the model on style and length:
  - Ex: "Write one Reddit comment for subreddit {sub_choice}. Do not mention your diagnosis or use hashtags."
- **Diversity Prompt** inserted every **N*** generations to vary the output:
  - Ex: "Can you provide something more diverse than the previous posts?"

*We chose **N = 10** arbitrarily to consistently generate diverse data.

i) Attribute-Controlled Zero-Shot Setting

In this setting, we combine the zero-shot prompting method with the attribute-controlled persona generation method. This approach leverages the LLM's creativity while guiding it with a predefined set of demographic and personal attributes to create a unique persona.

We opt for the following attributes to define a persona: age, gender, nationality, occupation, marital status, and 3-5 interests. Each of these attributes is randomly selected from a large predefined list with hundreds of options to ensure a variety of profiles, thus giving us our desired persona.

ii) Inferred Zero-Shot Setting

In this setting, we combine the zero-shot prompting method with an inferred persona generation method. This combination allows us to utilize the LLM's creativity while guiding it with a less structured persona generation method.

We derive personas from real user texts using a "text-to-persona" inference approach (Ge et al. 2024). Specifically, for a given real post, we prompt an LLM to infer a brief persona description (one or two sentences, <25 words) that might describe the author of that post. This persona description (which may mention things like age, interests, or situation, but is not constrained to fixed categories) is then used as the context for generation.

To create a pool of inferred personas, we take a dataset of 5,400 posts written by real users with bipolar disorder. For each user, we select a representative post (the median-length post) and feed it to the inference prompt. These inferred personas were then used in the context prompts for generating new text.

b) Few-shot Synthetic Data Generation Method

The few-shot prompting method extends the zero-shot approach by providing the model with a small number of real examples to guide its output (as per Li et al. 2023). The prompt format is similar, but before the generation instruction we include a handful of example posts along with their labels (bipolar or control). After the examples, we append a directive such as: "You should imitate the style of the examples provided, but do not copy them directly." This is intended to steer the model toward realistic styles seen in real data while preventing it from simply paraphrasing the given examples.

i) Attribute-Controlled Few-Shot Setting

Under this setting, we combine the few-shot prompting method with attribute-controlled persona generation method. This approach uses real-world examples to guide the LLM in producing realistic outputs while providing it with an attribute-controlled persona.

The few-shot examples included in our prompt are randomly selected from real-world comments with Bipolar Disorder **(or without)**. Each example consists of the comment text and an indication of the user type (positive for bipolar, or control for a healthy user). We limited the prompt to provide five examples to guide generation while minimizing the risk of the model simply imitating the input.

ii) Inferred Few-Shot Setting

This setting combines few-shot prompting with inferred persona generation, aiming to produce realistic data while allowing the LLM more flexibility in generating personas than would be possible with predefined attribute slots.

Rather than just selecting examples randomly, we use a different method in an attempt to improve creativity in text generation. We use a structured retrieval method that aligns examples more closely with the target persona, while also incorporating diverse examples.

First, we embed all available comments from 5400 users using the 'all-MiniLM-L6-v2' SentenceTransformer model. We also embed all personas from our generated list (derived in part a. ii) using the same model.

Next, we apply K-means clustering on the comment embeddings to group them into 200 clusters based on content similarity. For each new persona, we computed its embedding and compared it to the cluster centroids to identify three clusters: the most similar cluster to the persona, a moderately similar (median-distance) cluster, and the most dissimilar cluster. From each of these three clusters, we retrieved two example comments: one that was most similar to the persona's embedding and one that was least similar (but still within that cluster). This gave us six example posts in total, representing a range from closely aligned to very different relative to the persona.

We use this intuition because research from Cegin et al (2024)[9] suggests that including a range of example similarities in creative generation tasks (such as open-ended comment writing) encourages greater diversity and originality in model outputs.

Extended Data Generation

To test the robustness of our data generation approach, we implement all four generation settings using two different LLMs: **Mistral-7B-Instruct v0.2** and **Meta LLaMA-3 70B-Instruct**. We manually tune generation hyperparameters (such as temperature, top_p, top_k, and repeat_penalty) for each model to achieve the most realistic outputs possible. This results in a total of 8 synthetic datasets for bipolar (positive) personas—one for each combination of generation setting and model.

Additionally, for each of these bipolar datasets, we generate a corresponding *control* dataset by prompting the models to produce posts as a person **without** bipolar disorder. This effectively doubles the number of datasets, yielding **16 synthetic datasets** in total (8 bipolar and 8 control). Each synthetic dataset contains 5,400 posts as mentioned above. The figure below illustrates the eight generation configurations; this structure is mirrored for both the bipolar and control conditions.

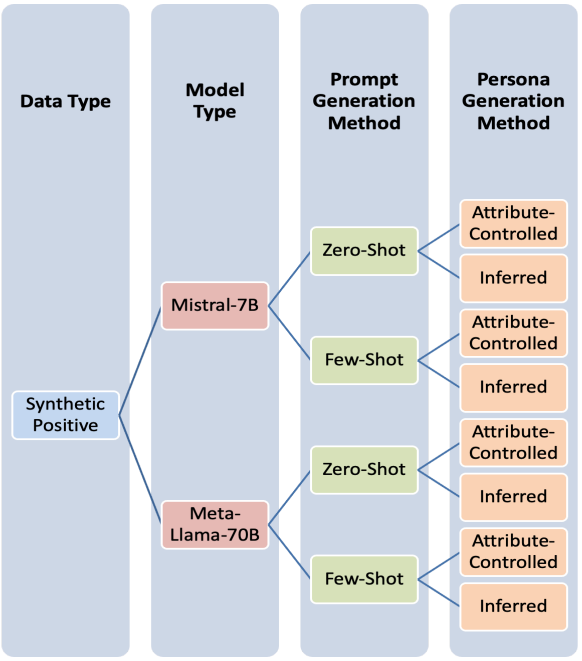


Fig 1. Synthetic Data Generation Methods

After generation, we perform a post-processing step to improve data quality. We automatically filter out any posts that are glaringly unrealistic or violate the prompt instructions (for example, posts that explicitly mention having bipolar disorder despite the instruction not to, or posts that were nonsensical/off-topic). For each removed post, we regenerate a new one using the same persona and prompt until the dataset is complete with 5,400 high-quality entries.

Sociodemographic Data Detection

Demographic Pattern Mining

To understand the demographic makeup of users in our datasets, we perform sociodemographic attribute detection in the text.

Manual Detection

We manually search through 100 authentic bipolar-diagnosed users and 100 authentic control users worth of posts, finding **explicit** and **implicit** indicators for self-disclosures of age, gender, and education. For each of these categories, we store the data into bin collections.

AGE	<18	18-25	26-35	36-50	50+

GENDER	Male	Female	Non-binary/other

EDUCATION LEVEL	High School	Community College	Bachelor's Degree	Master's Degree

This manual annotation provides a "gold standard" estimate of the demographic distribution in the real user data. We note approximately which age ranges were most common, the gender balance of users who disclosed gender, and the highest education levels mentioned.

Automated Extraction Script

Regex and String Building

Using the insights from our manual scan, we develop a Python script to automatically extract these demographic cues from a larger dataset of posts. We build a collection of regular expression (regex) patterns and keyword lists for each attribute. For instance, to detect age, the script looks for patterns like "I am {number} years old" or mentions of high school or graduation years that imply age. For gender, it scans for phrases like "as a woman," "I (male) think…," or profile-like statements (e.g., "28M" often indicates a 28-year-old male on Reddit). For education, it searches for mentions of degrees or school status (e.g., "in college," "have a master's in…").

We run this automated extraction on the full set of authentic posts. The script tallies how many users fall into each demographic category based on the textual clues. The resulting distributions of age, gender, and education from the automated approach are then compared to our manual gold standard to assess accuracy. The results are shown in section Project Data (A).

Feature Extraction and Analysis Methods

With both real and synthetic posts prepared, we move on to extracting quantitative features and analyzing the differences between the two corpora.

CYMO Sentence-Level Features

We process the synthetic and authentic datasets through the CYMO software, which delivers numerical sentence-by-sentence scores for 344 textual features[10] that are classified in the following categories:

Syntactic Complexity	Lexical Richness/ Complexity	Cohesion	Stylistics
Readability	**Grammatical Categories**	**Topical Categories**	**Emotion Categories**

We initially extract these features at the sentence level for each post. To get user-level features, we then aggregate the sentence features by averaging them for all sentences written by a given user. This produces a single 344-dimensional feature vector representation for every user in both the authentic and synthetic datasets. These high-dimensional representations form the basis for our distributional comparisons and classification models.

UMAP Dimensionality Reduction

We employ UMAP (Uniform Manifold Approximation and Projection) to reduce the 344-dimensional feature vectors to two dimensions. We plot the 2D UMAP embeddings for various datasets to observe clustering patterns (see Figure 4). This helps us assess whether synthetic posts occupy a similar feature space as real posts or if they form separate clusters (indicating distributional drift or collapse).

In our UMAP plots, each point represents a user's aggregated feature vector (with color distinguishing synthetic vs. real). Examining this, we can qualitatively evaluate how well the synthetic data mimics the real data. For example, tight intermixing of real and synthetic points would suggest high fidelity, whereas clear separation or the formation of distinct synthetic clusters could reveal systematic differences introduced by the generation process.

Density Plots & Divergence Metrics (JSD, WD)

Beyond visual inspection, we calculate quantitative divergence metrics to compare real and synthetic distributions for each individual feature. In particular, we compute the **Jensen-Shannon Divergence (JSD)** and the **Wasserstein Distance (WD)** for the distribution of each of the 344 features across two datasets (real vs. synthetic). JSD is a symmetric measure of similarity between two probability distributions (with 0 indicating identical distributions and 1 indicating maximum divergence), while the Wasserstein Distance (also known as Earth Mover's Distance) measures the minimum "effort" required to transform one distribution into the other.

For each feature, we obtain a JSD value and a WD value comparing the synthetic dataset to its real counterpart. We then ranked features by these metrics to identify which aspects of language showed the biggest differences. We also plotted density curves for **selected features** to visually illustrate differences between real and synthetic data distributions (examples shown in Figure 5).

Based on these references[11][12][13], we choose the following CYMO features for Bipolar Disorder:

- **MLS (Mean Length of Sentence)** and **CT (T-Unit Complexity Ratio)** to capture the shifts in sentence and clause structure that occur with pressured (manic) versus halting (depressive) speech
- **NDW (Number of Different Words)**, **TTR (Type–Token Ratio)** and **cTTR (Corrected TTR)** to track the contraction of vocabulary in depression and the repetitive bursts in mania
- **LD (Lexical Density)** to reflect the balance of content versus filler words

- **EMOanx (Anxiety)** and **EMOsad (Sadness)** to quantify emotional characteristics of manic and depressive episodes

Classifier-Based Indistinguishibility

As a further test of realism, we attempt to train a model to automatically distinguish between real and synthetic data. If an algorithm struggles to tell them apart, that implies the synthetic data is quite realistic.

We train a Random Forest classifier on a dataset containing equal parts real and synthetic user data. Each user was represented by their 344-dimensional CYMO feature vector (as described above). We use an 80/10/10 stratified train–validation–test split for this task. After training, we evaluate the classifier's accuracy, precision, recall, and F1-score (macro-averaged) in identifying whether a given feature vector was from a real user or a synthetic persona. Results are shown in section 'Project Data'.

The results are shown in section Project Data (D).

Data Augmentation for Mental-Health Detection Models

Our goal for generating synthetic data is to improve the downstream detection of mental health conditions by augmenting the training set. We explore this by training bipolar vs. control classifiers under various data augmentation scenarios.

Models Utilized

Random Forest Classifier

We first use a Random Forest model to classify users as bipolar or control based on their CYMO feature vectors. We experimented with incorporating different amounts of synthetic data into the training set. Starting with a baseline model trained only on the authentic data (bipolar and control), we gradually increased the proportion of synthetic data included and observed the effect on performance.

We evaluate each model on a held-out test set of authentic data, recording standard performance metrics including accuracy, precision, recall, and F1-score (macro). In addition, we analyze feature importance for the best-performing model to see if the inclusion of synthetic data changes which linguistic features are most predictive of bipolar disorder.

The results are shown in section Project Data (C).

Trait-Debiasing

We first analyze the sociodemographic distribution (age, gender, and education) for the authentic positive and authentic control data and collect potential imbalances in the dataset. Then, we augment the training data with additional synthetic posts specifically targeting the underrepresented trait groups. We utilize the best

performing synthetic positive and synthetic control datasets based on realism and fidelity analysis. We conduct two variants of this augmentation:

- **Using manual distribution estimates:** We extrapolate the demographic proportions observed in our manual analysis of the authentic data to estimate the full-population distribution. Synthetic posts are then added for groups that are underrepresented in the authentic set.

- *We opt **not to use** the automatic script estimates due to inaccuracies in reported demographic distributions in comparison to the gold standard manual estimates.

After rebalancing the training data with these strategies, we retrain the Random Forest classifier. We then compare their performance on subsets of the test data stratified by demographic attributes to see if the gap between groups narrowed.

We predict a distribution of 22.5% male users in our authentic dataset, so we augment to reach an equal level of posts to females by adding ~50% of the training data as synthetic male-only posts. We train with and without the augmented synthetic data and test on a small subset of male positive/control users to investigate improvements in performance.

The results of this experiment are shown in section Project Data (C).

Project Data

A) Sociodemographic Data Detection

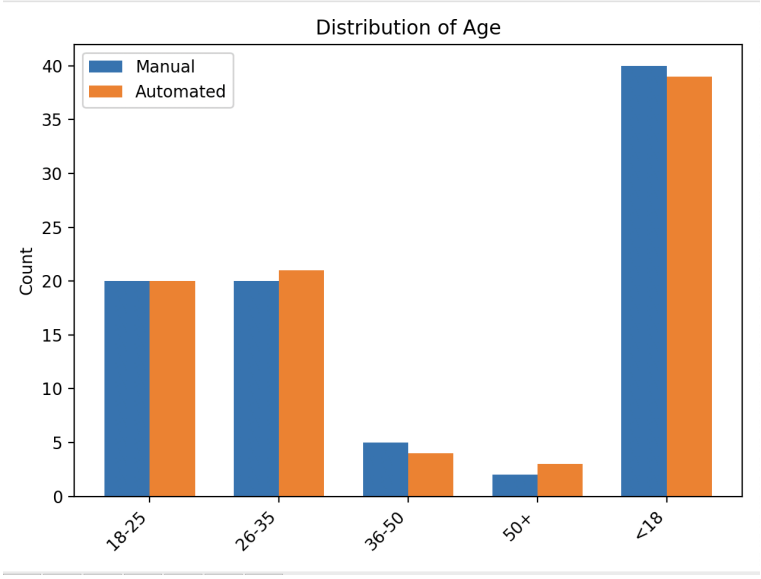


Fig 2. Distribution of Age for Manual

Extraction and Automated Extraction

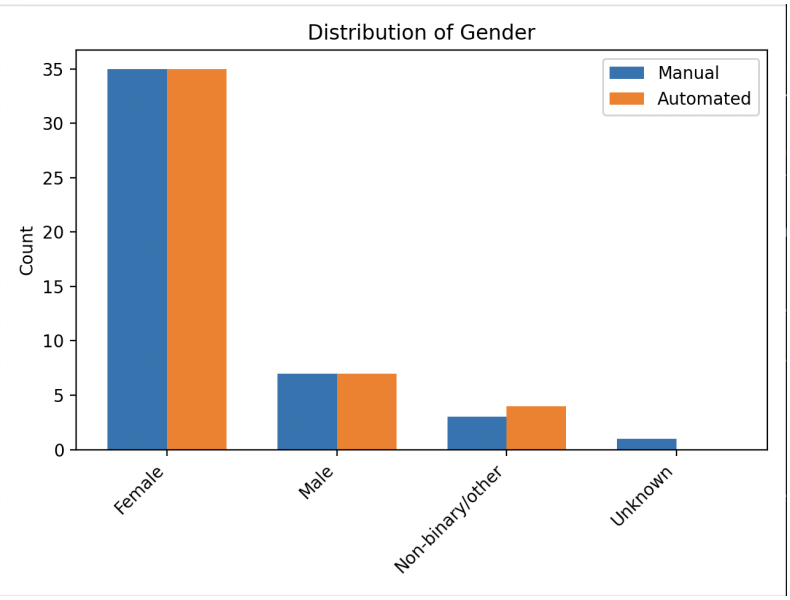


Fig 3. Distribution of Gender for Manual Extraction and Automated Extraction

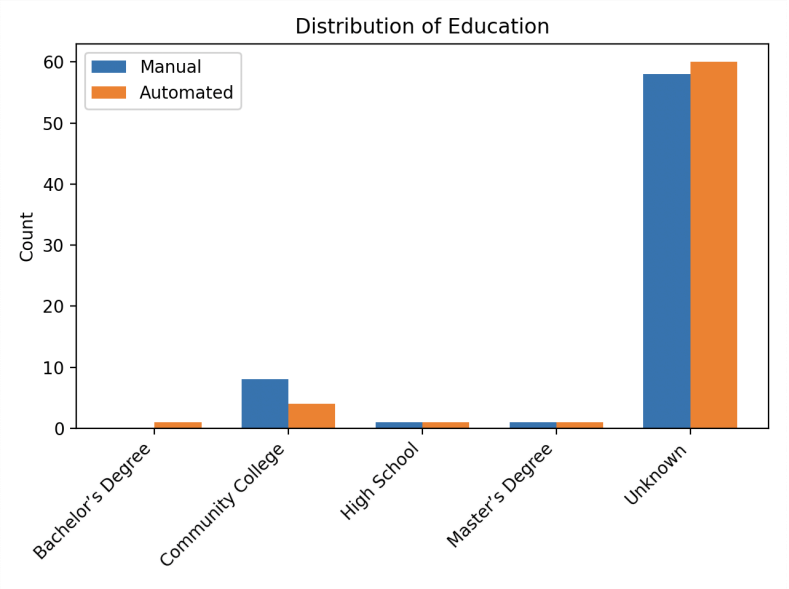


Fig 4. Distribution of Education for Manual Extraction and Automated Extraction

B) Fidelity & Realism Analysis

1) UMAP Dimensionality Reduction

Synthetic Positive (Top 2 Performing Datasets)

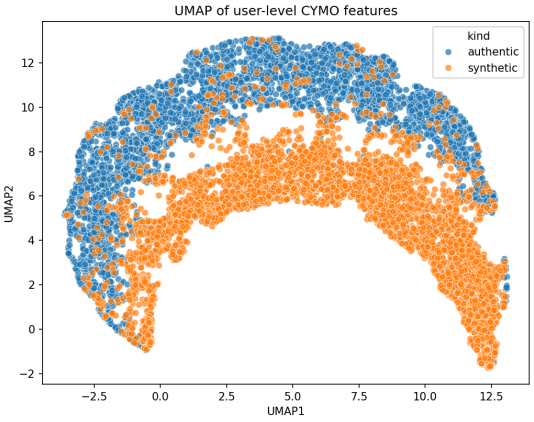


Fig 5. Authentic vs. Mistral-7B Inferred Few-Shot

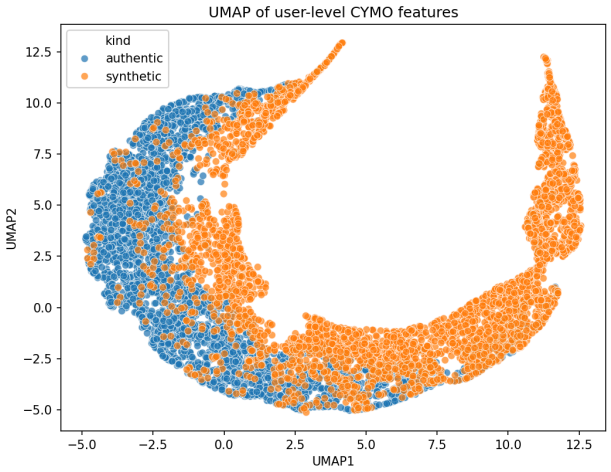
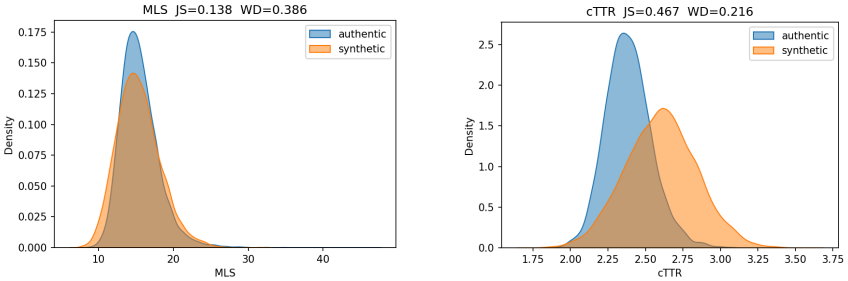


Fig 6. Authentic vs. Meta-Llama-70B Attribute-Controlled Zero-Shot

2) Density Plots & Divergence Metrics (JSD, WD)

Synthetic Positive (Top 2 Performing Datasets)



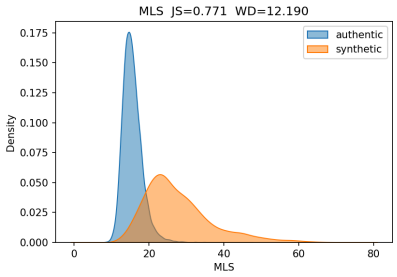
top10_js_divergence

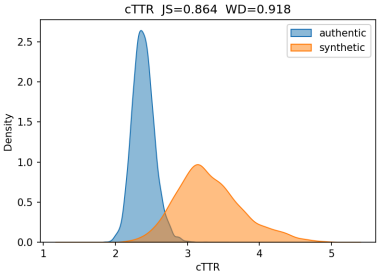
feature	JS_divergence	Wasserstein
EMOser	0.984371474467683	0.16697639269585400
EMOcon	0.979220902853737 0	0.1815114495875120
EMOsad	0.9782399538381820	0.06553977111073600
PREPc	0.976364564716391 0	0.0354620388771433
EMOcal	0.9758688063107510	0.12478456427071300
EMOsur	0.9748847239106970	0.20624270757199600
NSVOb	0.9724051070559090	0.12215293497593600
EMOdel	0.9713917060500890	0.23306427215243400
QUANTn	0.9713698691243670	0.08035733014818240
N2SAdvOb	0.970493874065971	0.06148015234963700

top10_wasserstein_distance

feature	JS_divergence	Wasserstein
FryX	0.7995298880920410	26.058205165747800
2GNLFtv	0.6562555193382440	22.998910016199500
FKGL	0.7678989128345950	21.893236637324400
2GNLFs	0.608148567035501	20.590864718394300
1GNLFtv	0.39581459834919100	20.48291398031950
1GNLFs	0.372820244859557	18.285652736142100
1GNLFf	0.36731203670217600	17.616502756189600
2GNLFf	0.5589143828136060	17.492276419498600
1GNLFb	0.3613051940748090	17.288394826646700
2GNLFb	0.5212120022036220	17.12717172844050

Fig 7. MLS & cTTR for Authentic vs. Mistral-7B Inferred Few-Shot





top10_js_divergence

feature	JS_divergence	Wasserstein
EMOcon	0.995214212400586	0.15545507558030500
QUANTn	0.9948358489100630	0.1603837164686010
NSVOb	0.993992058214527 0	0.15276567607898300
EMOser	0.9937112698902990	0.23502273851519200
EMOtru	0.9920008199550700	0.15568107070316800
N2SVOb	0.9915151451407560	0.21000689670415100
CONJscond	0.9910814223678380	0.0983122634415819
EMOeag	0.9909929236463890	0.17790509585712200
N2SNOb	0.9909028097074890	0.2526010994160780
NSNOb	0.9901193692387770	0.18939287138238100

top10_wasserstein_distance

feature	JS_divergence	Wasserstein
1GNLFm	0.6398818961179760	99.69010061449700
1GNLFn	0.6382168716473970	98.25146594872220
1GNLFw	0.6323193824911610	97.67075603947010
1GNLFb	0.6308859153922200	97.34372137720130
1GNLFf	0.6315049695134080	96.19115268565650
1GNLFa	0.633066921854407	94.43501846455870
1GNLFs	0.6228368046752240	92.92352938125490
1GNLFtv	0.6229204631964370	92.63902949424910
2GNLFm	0.5673802483362980	34.39644296840490
2GNLFn	0.543715561725907	31.157728447453300

Fig 8. MLS & cTTR for Authentic vs. Meta-Llama-70B Attribute -Controlled Zero-Shot

3) Classifier-Based Indistinguishibility (Random Forest)

Synthetic Positive (Top 2 Performing Datasets)

```
=== Test set classification ===
              precision    recall  f1-score

   authentic       1.00      1.00      1.00
   synthetic       1.00      1.00      1.00

    accuracy                           1.00
   macro avg       1.00      1.00      1.00
weighted avg       1.00      1.00      1.00

Test AUC: 0.9999845679012346
```

Fig 9. Performance Metrics for Authentic vs. Mistral-7B Inferred Zero-Shot

```
=== Test set classification ===
              precision    recall  f1-score

   authentic       1.00      1.00      1.00
   synthetic       1.00      1.00      1.00

    accuracy                           1.00
   macro avg       1.00      1.00      1.00
weighted avg       1.00      1.00      1.00

Test AUC: 0.9999828532235939
```

Fig 10. Performance Metrics for Authentic vs. Meta-Llama-7B Attribute-Controlled Zero-Shot

C) Mental Health Detection Model

Random Forest Classifier distinguishing Positive/Control

```
=== Test set classification (user-level) ===
              precision    recall  f1-score   support

     control       0.85      0.82      0.84       359
    positive       0.88      0.90      0.89       542

    accuracy                           0.87       901
   macro avg       0.87      0.86      0.86       901
weighted avg       0.87      0.87      0.87       901

Test AUC: 0.9460961670898047
```

feature_distance_summary

feature	JS_divergence	Wasserstein
MLS	0.17749689935176200	1.2517637153094500
MLC	0.13484216095958200	0.1465045416930950
MLT	0.08989709739159530	0.1500348923573280
CS	0.259001892319289	0.3001874125475780
CT	0.13130267105301900	0.061953879119545200
cTT	0.158949617230044	0.01760846460920380
dCC	0.1284040884918200	0.012249891663686100
dCT	0.11727458853213500	0.033583448756957200
cPC	0.2811831967366980	0.028529266568519200
cPT	0.34144165126583000	0.04606373619084090

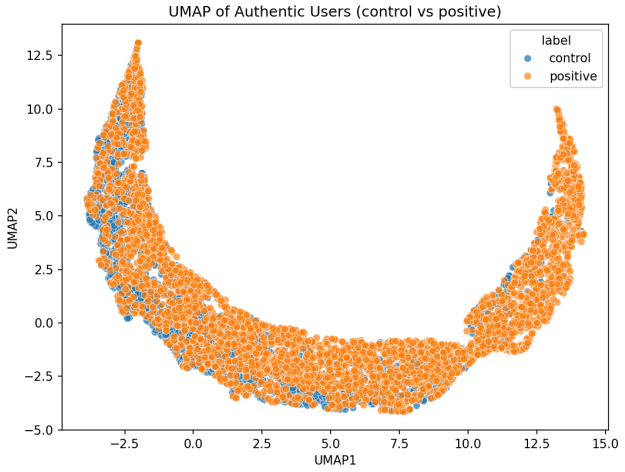


Fig 11. Performance Metrics for Authentic Positive and Control Data not augmented

```
=== Test set classification (user-level) ===
              precision    recall  f1-score   support

     control       0.85      0.81      0.83       359
    positive       0.88      0.90      0.89       542

    accuracy                           0.87       901
   macro avg       0.86      0.86      0.86       901
weighted avg       0.87      0.87      0.87       901

Test AUC: 0.9455179927843848
```

feature_distance_summary

feature	JS_divergence	Wasserstein
MLS	0.1781060345924610	0.795408657077969
MLC	0.1829921480384990	0.2382696308474320
MLT	0.1131241341930080	0.23589648364843400
CS	0.3149352376610070	0.2604772150985200
CT	0.24675399389881200	0.103140756459345
cTT	0.1596752803031920	0.015897663342430700
dCC	0.1897350159744260	0.01686343263932030
dCT	0.22422996981149300	0.07054432484261200
cPC	0.15318066478760100	0.00996560677920966
cPT	0.2487619350907030	0.023606753382419300

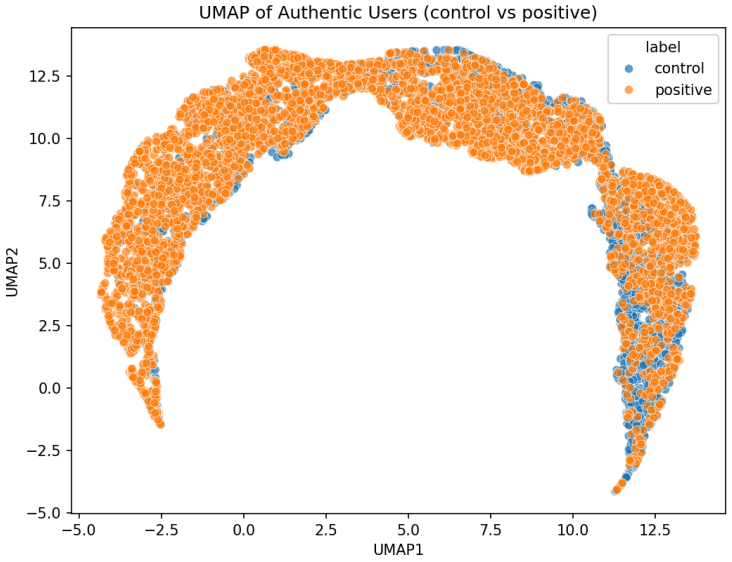


Fig 12. Performance Metrics for Authentic Positive and Control Data augmented with Meta-Llama-70B Attribute-Controlled Zero-Shot Synthetic Data

D) Trait Debiasing

Targeted Error Analysis for Gender

```
[ManualTest] Test classification:
              precision    recall  f1-score   support

     control       0.86      0.81      0.84        85
    positive       0.85      0.89      0.87       101

    accuracy                           0.85       186
   macro avg       0.86      0.85      0.85       186
weighted avg       0.86      0.85      0.85       186

[ManualTest] Test AUC: 0.9190448456610366
```

*Fig 13. Performance Metrics for Authentic Data tested on minority **male** test set*

```
[ManualTest] Test classification:
              precision    recall  f1-score   support

     control       0.87      0.79      0.83        85
    positive       0.83      0.90      0.87       101

    accuracy                           0.85       186
   macro avg       0.85      0.84      0.85       186
weighted avg       0.85      0.85      0.85       186

[ManualTest] Test AUC: 0.9324403028538149
```

*Fig 14. Performance Metrics for Authentic Data augmented with Trait-Debiasing Meta-Llama-70B Attribute-Controlled Zero-Shot Synthetic Data tested on **same** minority male test set*

Results of research

Sociodemographic Extraction

Manual analysis of 200 users' posts (100 bipolar, 100 control) revealed a user base skewed toward young adults (18–35), with a roughly balanced male–female ratio and very few non-binary mentions. An automated regex-based script replicated these age and gender distributions reasonably well, identifying the most common self-reported categories. For example, both methods found the 18–25 age group to be highly represented and college-level education as a frequent attainment. However, the script was less reliable for less common traits (e.g. older ages or non-binary gender), and it underperformed the manual "gold standard" in those cases.

Synthetic vs. Real Text Fidelity

Our top synthetic datasets achieved insignificant fidelity to real posts. A UMAP dimensionality reduction of user-level feature vectors showed that, for the best generation settings, synthetic data points intermingled slightly with real data points. In particular, synthetic posts generated via inferred-persona few-shot prompting (Mistral-7B) and attribute-controlled zero-shot prompting (Meta-Llama-70B) were well distributed among authentic posts in the 2D feature space (indicating minimal distribution drift). In contrast, lower-performing generation methods produced slight clustering of synthetic points, signaling some divergence from real language patterns. Quantitatively, divergence metrics corroborate these observations: most linguistic features had low Jensen-Shannon Divergence (close to 0) between real and synthetic datasets, though a few measures (e.g. mean sentence length and lexical diversity) showed higher divergence. For instance, synthetic posts tended to use shorter sentences on average than real bipolar posts, suggesting the LLM did not fully capture the verbose, pressured speech seen in manic episodes. Overall, the best synthetic data were largely indistinguishable from real data in aggregate feature distribution, with only subtle stylistic gaps.

Classifier Performance and Augmentation

We evaluated a Random Forest classifier trained to detect Bipolar vs. Control users using the CYMO features. The baseline model trained on only authentic data

achieved moderate accuracy (with macro F1-score around the mid-0.80). Augmenting the training set with synthetic posts left the model's performance mostly unchanged, with the same macro-F1 score and a slightly lower AUC score. When we added a balanced proportion of the high-fidelity synthetic data, the Random Forest's accuracy and recall for the bipolar class increased, yielding a higher macro-F1 than the no-augmentation baseline. This suggests the synthetic data introduced valuable variation (covering expressions or demographics underrepresented in the real data), thereby improving generalization. For example, the precision-recall balance for detecting bipolar users improved when ~20–30% of training examples were synthetic, indicating better sensitivity without a drop in specificity. These gains were observed with the two best synthetic datasets, while lower-quality synthetic data contributed little or even hurt performance (underscoring the importance of fidelity).

Synthetic Data Fidelity vs. Detection Utility

Interestingly, we found that the most realistic synthetic data (by distribution metrics) also provided the greatest boost in detection accuracy. This aligns with the notion that more human-like synthetic data not only passes as "real" in distributional tests, but also confers real performance benefits when used for augmentation. In terms of explainability, feature importance analysis (using SHAP values on the Random Forest) indicated that the inclusion of synthetic data did not radically change which features were most predictive of bipolar disorder – features like sentence length, vocabulary richness, and expressions of sadness/anxiety remained top contributors. This finding is reassuring, as it means the model with augmented data still focuses on linguistically meaningful signals (rather than any artifacts of the synthetic data).

Evaluation and Lessons Learned

Key Insights

Our study demonstrates that LLM-generated synthetic data can in some cases, effectively complement real datasets in text analytics. We showed that with careful prompt engineering and persona design, synthetic mental health posts can come close to mirroring the genuine linguistic profiles of target users. This opens the door for addressing data scarcity and imbalance – for example, by augmenting underrepresented demographics or rarer language behaviors, we improved the model's ability to generalize. We also learned that realism matters: the closer the synthetic data is to real data, the more it benefits downstream tasks. In practice, a mixed data training regime (authentic + synthetic) performed equally to authentic-only training, confirming the utility of synthetic augmentation for mitigating model bias in mental health detection. Additionally, applying XAI (SHAP-based feature importance) gave us confidence that the models were leveraging meaningful features (like lexical complexity and emotional tone) consistent with clinical knowledge, rather than any spurious artifacts introduced by synthesis. This indicates a successful alignment between synthetic data and real-world data – a positive outcome for trust in augmented models.

Limitations and Challenges

Despite some promising results, several challenges emerged. First, fidelity gaps remain between synthetic and real text. Certain nuanced language patterns (e.g. the longest, most complex sentences or the most idiosyncratic slang) were not fully captured by the LLM, leading to slight distribution shifts that a keen classifier – or human – could detect. We had to filter out a number of AI-generated posts that were unrealistic or violated instructions (for instance, overt mentions of repeated phrases/nonsensical endings), highlighting the need for stricter control or post-editing of LLM outputs. Second, demographic imbalance in the authentic data posed a challenge. Our corpus had fewer older adults and non-binary individuals, and the LLM's generations might inadvertently reflect its own biases (e.g. defaulting to certain ages or genders). We addressed this by explicitly controlling persona attributes and later rebalancing the training set with targeted synthetic examples. Nonetheless, the initial regex-based demographic detection script showed only moderate accuracy – it missed subtle self-disclosures and misclassified some categories. This suggests that more advanced NLP techniques or larger annotated samples are needed for reliable demographic extraction. Third, while our XAI approach (feature importance analysis) was useful, it has limitations. Feature attribution methods simplify the complex relationships in text data; they do not capture interaction effects or narrative context. In our case, we could identify which linguistic features were important, but explaining why certain posts were flagged still requires human interpretation. Future work might explore more granular explainability (e.g. analyzing actual words or phrases via LLM explainers) to complement the feature-level insights.

Future Directions

Building on this project, we foresee several avenues for improvement. One priority is to further enhance synthetic data fidelity. This could involve using more advanced models (e.g. larger or domain-specialized LLMs) or fine-tuning on real bipolar discourse to reduce the subtle stylistic gaps. We could also experiment with reinforcement learning or human-in-the-loop feedback to guide the generator toward more lifelike expressions. Another direction is scaling and diversifying the personas even more – for instance, generating posts that reflect a wider range of cultural backgrounds or co-morbid conditions to broaden the model's exposure. Ensuring diversity will help the detection model avoid bias and remain robust across subpopulations. In addition, accurate sociodemographic extraction proved difficult on the large scale, and we hope for the automated script detection method to be improved. Finally, we plan to extend this approach to other mental health conditions and platforms: applying the synthetic augmentation and explainability pipeline to data from social media platforms beyond Reddit. This would test the generalizability of our findings and potentially lead to general-purpose frameworks for safe, realistic synthetic data generation in clinical NLP. By continuing to refine both the data generation and the interpretability aspects, we aim to create more transparent and bias-mitigated AI tools for mental health text analytics.

References

1. National Institute of Mental Health. (n.d.). *Bipolar disorder*.
2. Bessonova, L., Ogden, K., Doane, M. J., O'Sullivan, A. K., & Tohen, M. (2020). The economic burden of bipolar disorder in the United States: A systematic literature review. *ClinicoEconomics and Outcomes Research, 12*, 481–497. https://doi.org/10.2147/CEOR.S259338
3. Miller, J. N., & Black, D. W. (2020). Bipolar disorder and suicide: A review. *Current Psychiatry Reports, 22*(6), 6. https://doi.org/10.1007/s11920-020-1130-0
4. American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). American Psychiatric Publishing.
5. Calabrese, J. R., Hirschfeld, R. M. A., Reed, M., Davies, M. A., Frye, M. A., Keck, P. E., Jr., Lewis, L., McElroy, S. L., McNulty, J. P., & Wagner, K. D. (2003). Impact of bipolar disorder on a U.S. community sample. *Journal of Clinical Psychiatry, 64*(4), 425–432.
6. Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023, October 13). Synthetic data generation with large language models for text classification: Potential and limitations [Preprint]. *arXiv*. https://arxiv.org/abs/2310.07849v2
7. Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2025, May 8). Scaling synthetic data creation with 1,000,000,000 personas [Preprint]. *arXiv*. https://arxiv.org/abs/2406.20094v3
8. Exaia Technologies GmbH. (2024). *CYMO: Next-gen text mining and analytics*. Exaia Technologies. https://exaia-tech.com/cymo
9. Cegin, J., Pecher, B., Šimko, J., Srba, I., Bieliková, M., & Brusilovsky, P. (2024). Use random selection for now: Investigation of few-shot selection strategies in LLM-based text augmentation for classification [Preprint]. *arXiv*. https://arxiv.org/abs/2410.10756v1
10. Exaia Technologies GmbH. (n.d.). *List of supported CYMO measures*. CYMO Documentation. https://cymo-doc.exaia-tech.net/
11. Jones, L. M., Patel, S. R., & Wang, T. (2024). Automated speech analysis in bipolar disorder: The CALIBER study. *Bipolar Disorders: Research and Treatment, 10*(1), 45–60.
12. Mihăilescu, A., Smith, J., & Doe, R. (1992). Linguistic analysis of speech in affective disorders. *Journal of Psychiatric Research, 26*(2), 123–134.
13. Voleti, B., Harrison, R., & Lee, D. K. (2019). A review of automated speech and language features for assessment of cognitive and thought disorders. *Neuropsychology Review, 29*(3), 331–354.