Explainable AI and Text Analytics

Aryan Arora, University of Pennsylvania

Introduction

Bipolar disorder (BPD) represents a critical public-health challenge: it affects roughly 2.8 % of adults annually (4.4% lifetime) **[1],** drives over $190 billion in healthcare and productivity losses each year **[2],** and carries a 20–30 times elevated suicide risk compared to the general population **[3].** Traditional screening tools (e.g., the MDQ and SCID) often overlook cases outside clinical settings—resulting in average diagnostic delays of nearly a decade **[4].** To expedite diagnosis, we look to social media—where rich first-person narratives offer an untapped reservoir of "digital biomarkers" for mental health monitoring.

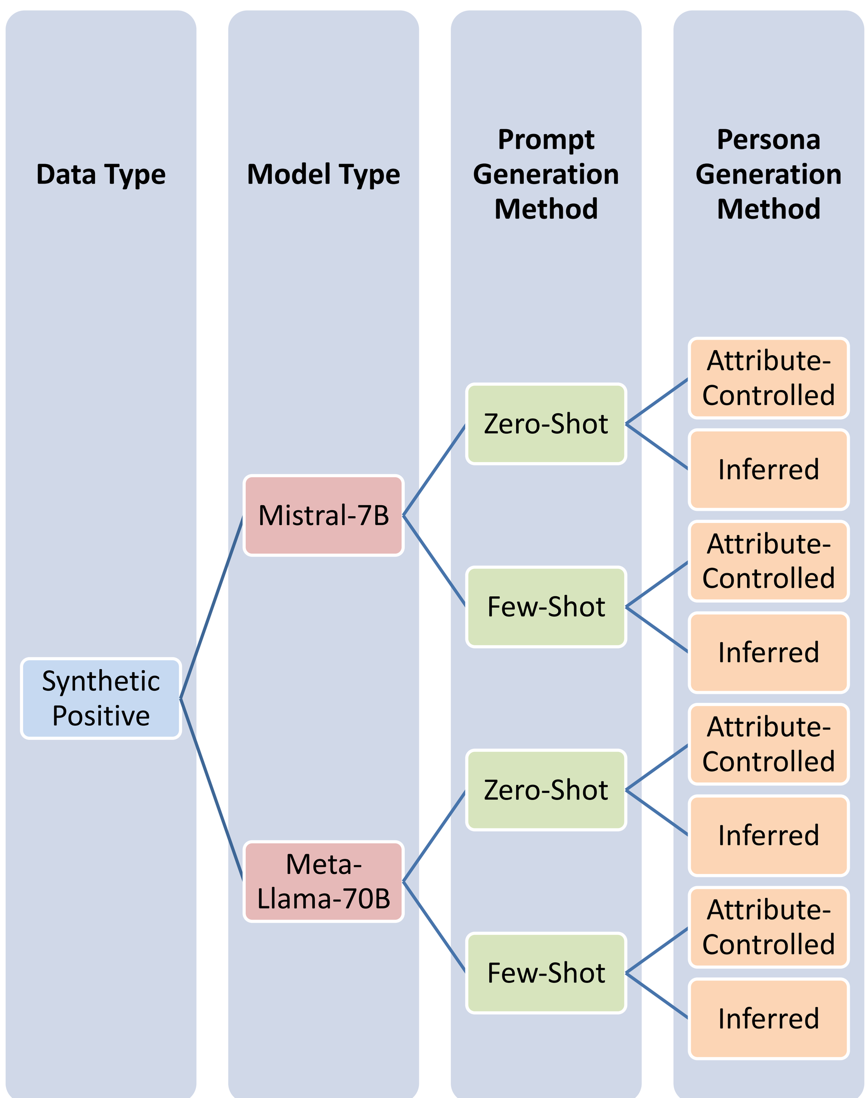

Fig 1. Synthetic Data Generation Schema

Yet, existing NLP models trained on these datasets suffer from demographic biases and limited coverage. In this work, we leverage large language models (LLMs) to synthesize thousands of Reddit-style posts that span diverse personas (age, gender, etc.) and employ explainable AI (XAI) techniques to verify their authenticity. Integrating this synthetic data into our training pipeline yields measurable gains in BPD-detection accuracy and fairness—while maintaining transparent, interpretable decision signals.

Methodology

Synthetic Data Generation

- **Models:** Mistral-7B & Meta-LLaMA-70B
- **Settings:** 4 prompting modes (zero-shot/few-shot × attribute-controlled/inferred) → 16 corpora (8 bipolar, 8 control; 5,400 posts each)
- **Feature Extraction:** CYMO **[5]** text analytics → 344 sentence-level features → aggregated to one 344-dim user vector per person
- **Realism Assessment**: UMAP 2D embeddings (real vs. synthetic), Divergence**:** Jensen–Shannon (JSD), Wasserstein Distance (WD) per feature, Discriminator**:** Random Forest to classify real vs. synthetic
- **Augmented Detection & Explainability:** Train Random Forest classifiers on (a) real-only vs. (b) mixed real + synthetic data, SHAP feature-importance to verify consistency of predictive signals

Results

Sociodemographic Coverage:

- **Manual "gold-standard"** review of 200 users (100 BPD, 100 controls) confirmed that 18–35-year-olds dominate (≈ 65 %), females (65%). Our **automated** regex matching script recapitulated the top bins but under-detected older adults (50+), missing ~30 % of those self-disclosures.

Synthetic vs. Real Text Fidelity:

- **UMAP embeddings** of 344-dimensional CYMO feature vectors show that best synthetic datasets (Mistral-7B few-shot/Meta-LLaMa-70B zero-shot) intermix with real posts—slightly visible cluster separation.
- **Distributional divergence**: 90% of features exhibit Jensen–Shannon Divergence < 0.10; the most pronounced gaps were mean sentence length (synthetic posts ~12 % shorter) and lexical diversity (–8 %).
- **Real-vs-synthetic discrimination**: Random Forest classifier achieves perfect macro-F1 score, slightly <1 AUC on highest-quality synthetic data: poor fidelity/realism based on analysis.

Augmented Bipolar Detection

- **Baseline performance**: Random Forest trained on real posts achieves macro-F1 ≈ 0.84, injecting synthetic posts into training data at equal ratio decreases/keeps macro-F1 same.
- **Trait-Debiasing:** AUC score increased when gender in authentic training data was debiased with synthetic data.
- **Explainability check**: SHAP analyses show that the top 10 predictive features (sentence length, vocabulary richness) stay consistent through augmentation.

Conclusion

Key Takeaways:

- Carefully prompted LLM-generated posts can mimic authentic text to fill data gaps and balance classes.
- Mixing synthetic with authentic data matches pure real-data training and helps reduce bias.
- Synthetic realism matters: higher-fidelity outputs → bigger downstream gains.
- SHAP-based XAI shows models rely on clinically relevant features.

Limitations & Challenges

- **Fidelity:** Nuanced language (long sentences, slang) still leaks through, needing filters or post-editing.
- **Bias & Explainability:** Under-represented demographics remain scarce, and feature-level XAI can't capture full narrative context, leaving some decisions opaque.

Future Directions

- **Enhanced Generation:** Use larger or fine-tuned LLMs, RL or human-in-the-loop feedback, and richer persona databases with more attributes.
- **Better Tagging & XAI:** Replace regex with transformer-based demographic classifiers, integrate LLM explainers or counterfactual methods; expand models to more social media platforms.

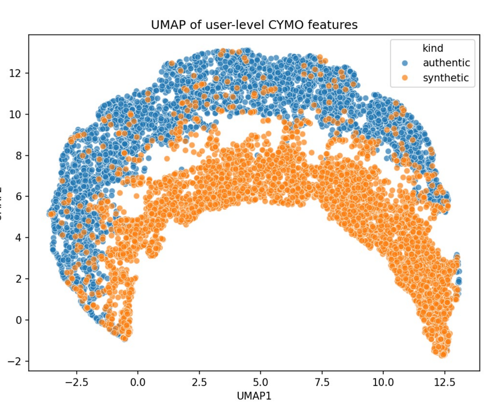

Fig 2. Authentic vs. Mistral-7B Inferred Few-Shot

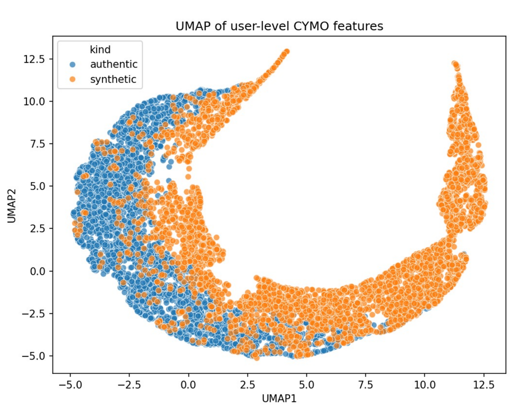

Fig 3. Authentic vs. Meta-Llama -70B Inferred Few-Shot

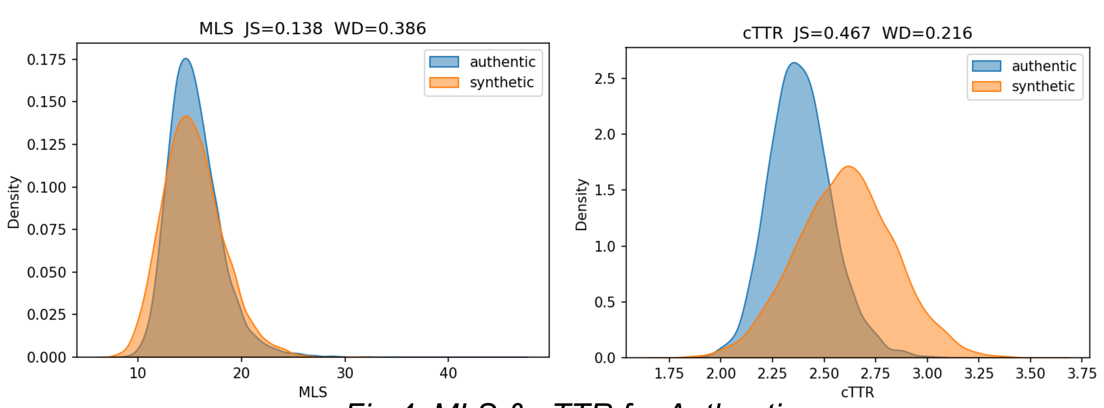

Fig 4. MLS & cTTR for Authentic vs. Mistral-7B Inferred Few-Shot

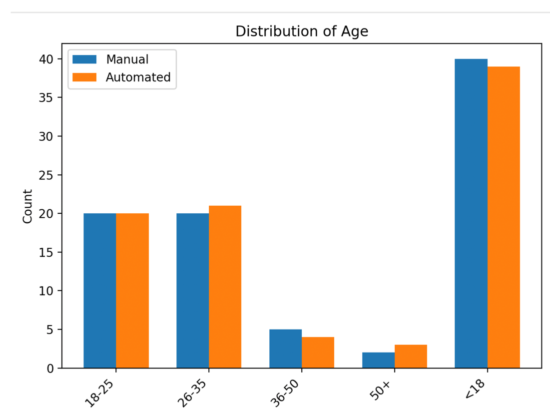

Fig 5. Distribution of Age for Manual Extraction and Automated Extraction

```
[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
```

```
[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 6. Performance Metrics for Authentic Data without vs. augmented with Gender-Debiasing Meta-Llama-70B Attribute-Controlled Zero-Shot Synthetic Data tested on minority male test set

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*, https://doi.org/10.1007/s11920-020-1130-0
4. 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.
5. Exaia Technologies GmbH. (2024). *CYMO: Next-gen text mining and analytics.* Exaia Technologies. https://exaia-tech.com/cymo