



Simulation-Based Natural Language Processing of Synthetic Clinical Notes to Detect Early Linguistic Markers of Antidepressant-Induced Hypomania in Bipolar Disorder: A Proof-of-Concept Study

Rocco de Filippis^{1,2*}, Abdullah Al Foysal³

¹Institute of Psychopathology, Rome, Italy

²ARAVIS Unit, EPSM 74, La Roche-sur-Foron, Haute-Savoie, France

³Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy

Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Simulation-Based Natural Language Processing of Synthetic Clinical Notes to Detect Early Linguistic Markers of Antidepressant-Induced Hypomania in Bipolar Disorder: A Proof-of-Concept Study. *Open Access Library Journal*, 13: e15676.

<https://doi.org/10.4236/oalib.1115676>

Received: June 22, 2026

Accepted: September 20, 2026

Published: September 23, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Antidepressant-induced hypomania in bipolar disorder (BD) is a clinically consequential pharmacological safety event that may be preceded by changes in patient and clinician language. This proof-of-concept simulation study evaluates whether natural language processing (NLP) can recover such signals from fully synthetic outpatient clinical notes and simulated hypomania labels; it does not establish clinical validity. We generated one synthetic note for each of $N = 700$ simulated antidepressant-exposed BD patients (hypomania prevalence 25.1%). A three-tier feature architecture encoded: i) 18 psycholinguistic and syntactic features; ii) 30 latent semantic analysis components derived from TF-IDF representations; and iii) 12 simulated clinical meta-features. A four-layer multilayer perceptron, termed ClinicalBERT-sim to distinguish it from an actually fine-tuned ClinicalBERT model, was trained on the combined feature matrix. An out-of-fold stacking ensemble fused ClinicalBERT-sim, LightGBM, and Random Forest outputs. The proposed ensemble achieved $AUC = 0.912$ (95% CI: 0.842 - 0.969), $F1 = 0.708$, sensitivity = 0.595, and specificity = 0.974 on the held-out synthetic test set. SHAP analysis identified hypomanic term score, word count, energy word score, physician concern score, and positive affect ratio as the dominant simulated predictors. Longitudinal illustrative trajectories showed rising hypomanic term score and word count across simulated weekly notes before simulated onset. These findings demonstrate technical feasibility within the designed simulator, but performance may partly reflect the

rules used to generate notes and labels. Validation on independently annotated, real outpatient BD notes is required before any clinical screening or workflow use.

Subject Areas

Psychiatry & Psychology

Keywords

Natural Language Processing, Clinical Notes, Bipolar Disorder, Antidepressant-Induced Hypomania, Psycholinguistics, TF-IDF, Latent Semantic Analysis, SHAP, Linguistic Biomarkers, Digital Phenotyping, EHR

1. Introduction

Psychiatric clinical notes are among the richest and most systematically underutilised data sources in mental health surveillance. In the outpatient management of bipolar disorder with antidepressant exposure, clinicians generate free-text progress notes at every visit that contain not only structured observations (YMRS scores, medication changes) but also nuanced language that reflects the patient's mental state in ways that structured rating scales cannot fully capture: the word choice, sentence structure, and affect conveyed in how a patient describes their week, their energy, their sleep, and their plans are direct signals of their phenomenological state [1].

The hypomanic prodrome in antidepressant-exposed BD patients is characterised by a predictable sequence of linguistic changes that begin before the clinical threshold for formal hypomania is reached [2]. Patients speak and write with more words, more energetic vocabulary, more positive affect language, and more future-oriented statements. Physicians, sensitive to these changes, begin expressing concern in their notes: the lexical field of clinical notes shifts from maintenance language to monitoring language [3]. These changes are legible to a trained psychiatrist who reads serial notes; the questions this paper addresses are whether they are also detectable by NLP models trained on these notes, and whether they are detectable early enough to be clinically actionable i.e., before the full hypomanic episode emerges.

The NLP evidence base for mental health applications has expanded with domain-specific language models, including BioBERT [4], ClinicalBERT [5], and PsychBERT [6]. However, PsychBERT was developed for social-media mental-health analysis rather than clinical notes, and the present study does not fine-tune any BERT model. Instead, it uses an MLP on engineered linguistic, LSA, and clinical features as a simulation proxy, labelled ClinicalBERT-sim throughout to avoid implying use of contextual ClinicalBERT embeddings.

We present five proof-of-concept contributions in a fully synthetic setting: i) a simulation framework for studying antidepressant-induced hypomania signals in

outpatient-style BD notes; ii) a three-tier feature architecture integrating psycholinguistic/syntactic features, LSA embeddings, and simulated clinical meta-features; iii) comparison of a deep multilayer perceptron proxy with conventional text and tree-based baselines; iv) four linguistic visualisations; and v) subgroup analyses within the simulated cohort. These contributions concern methodological feasibility and should not be interpreted as evidence of clinical prediction performance.

2. Background and Related Work

2.1. Linguistic Markers of Hypomania and Mania

The relationship between language and affective state in bipolar disorder has been documented since Kraepelin's clinical descriptions of pressured speech and flight of ideas as hallmarks of manic activation [7]. Quantitative linguistic analysis has confirmed that hypomanic and manic states are associated with increased word production, reduced lexical diversity (type-token ratio), elevated positive affect word density, increased self-referential pronouns, and a higher proportion of future-oriented language [8]. Automated speech analysis has detected hypomanic language patterns with high sensitivity using prosodic features (speech rate, fundamental frequency variability) and lexical features (energy vocabulary, social language) [9].

2.2. NLP of Clinical Notes in Psychiatry

Recent peer-reviewed literature supports the broader use of NLP in psychiatric and clinical text, but not the specific real-world applications previously attributed here. Reviews have mapped NLP applications across neuroscience and psychiatry [10] and specifically within bipolar-disorder research [11]. A 2024 study evaluated classification of unstructured EHR text by psychiatric diagnosis [12], while a 2025 bipolar-disorder pilot combined acoustic and natural-language markers for remote symptom assessment [13]. Together, these studies support the feasibility of psychiatric language analysis while also highlighting limited labelled data, domain shift, and the need for external clinical validation.

2.3. Clinical Language Models

ClinicalBERT is a BERT model adapted to de-identified clinical notes, whereas BioBERT was pretrained primarily on biomedical literature. PsychBERT was developed for mental-health language in social media. Recent reviews emphasise that clinical NLP models require task-specific labelled data, transparent validation, and careful treatment of bias and generalisability. Because no labelled corpus of antidepressant-induced hypomania notes was available for this study, we did not fine-tune ClinicalBERT or PsychBERT; the term ClinicalBERT-sim denotes only a multilayer perceptron operating on engineered and LSA features, not a transformer language model.

2.4. Psycholinguistic Feature Frameworks

The Linguistic Inquiry and Word Count (LIWC) framework [14] provides validated dictionaries for psycholinguistic categories including positive and negative affect, social language, cognition, biological processes, and temporal orientation. LIWC-derived features have shown strong performance in detecting mental health conditions from text, including depression ($r = 0.32$ with BDI scores) and anxiety (positive affect ratio inversely correlated with GAD-7) [15] [16]. Our feature extraction framework is inspired by but not directly dependent on LIWC, deriving analogous category scores from domain-specific hypomanic lexicons constructed from clinical and empirical sources.

3. Methods

3.1. Dataset Design

This proof-of-concept study used a fully synthetic dataset of $N = 700$ simulated outpatient BD patients receiving antidepressant-containing regimens (BD-I: 55%, BD-II: 45%). Each simulated patient contributed one synthetic note at a 4-week post-initiation review and one simulated binary label indicating hypomania emergence during the subsequent 4 weeks. No real EHR text, patient identifiers, or clinical outcomes were used [17] [18].

3.2. Synthetic Clinical Note Generation

Synthetic notes were generated with a controlled template-and-lexicon procedure. Note length, sentence count, lexical diversity, and punctuation were sampled from bounded distributions intended to resemble outpatient progress-note structure. Hypomanic lexical pools covered energy, reduced sleep need, elevated mood, pressured speech, goal-directed activity, future orientation, and grandiosity; stable/depressive pools covered mood stability, adequate sleep, medication continuation, and residual depressive symptoms. A latent simulated risk score combined BD subtype, antidepressant-class risk, mood-stabiliser adequacy, baseline YMRS and MADRS, interdaily stability, genetic-risk composite, prior antidepressant-associated hypomania, prior antidepressant trials, illness duration, age, and sex. The binary hypomania label was sampled from a logistic transformation of this risk score with added stochastic noise, and note-generation probabilities were conditioned on the same latent score. Subgroups were assigned directly from the corresponding simulated meta-features. Consequently, linguistic markers and clinical meta-features were not independent of label generation, and the resulting performance may partly represent recovery of simulator rules rather than discovery of clinical biomarkers. Physician concern language was sampled with a probability that increased with latent risk, making it an intentionally informative but potentially advantaged feature.

3.3. Feature Architecture

Psycholinguistic and syntactic features (18) comprised hypomanic term score,

positive and negative affect ratios, physician concern score, word count, sentence count, mean and standard deviation of sentence length, type-token ratio, exclamation rate, first-person pronoun rate, future-orientation score, energy-word score, sleep-word score, note length, number density, punctuation density, and upper-case letter rate. LSA features (30) were derived from TF-IDF representations using a maximum vocabulary of 500 terms, unigrams and bigrams, minimum document frequency of 3, and truncated SVD. The 12 simulated clinical meta-features were encoded as follows: BD subtype and sex as binary indicators; antidepressant-class risk as a prespecified ordinal score; mood-stabiliser adequacy as a continuous 0 - 1 score combining therapeutic dose/level and adherence; baseline YMRS and MADRS as continuous scale values at note generation; interdaily stability as a continuous 0 - 1 circadian-regularity index; genetic risk as a standardised synthetic polygenic/family-history composite; prior antidepressant-associated hypomania as binary; illness duration, prior antidepressant trials, and age as continuous or count variables. These operational definitions apply only to the simulator.

3.4. Model Architecture

Five NLP models were evaluated. TF-IDF + LR used L2-regularised logistic regression ($C = 0.1$) on TF-IDF unigrams and bigrams. Linguistic XGBoost used the 18 psycholinguistic/syntactic features only. ClinicalBERT-sim was a four-layer multilayer perceptron ($D \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 2$) trained on the combined engineered linguistic, LSA, and simulated clinical feature matrix; despite its historical name in the analysis code, it is not ClinicalBERT and does not use transformer embeddings. Random Forest and LightGBM used the same combined feature matrix. The proposed ensemble combined out-of-fold predictions from ClinicalBERT-sim, LightGBM, and Random Forest through a logistic-regression meta-learner.

3.5. Evaluation Framework

The 700 patient-level records were stratified by simulated outcome and split once into training ($n = 490$; 70%), validation ($n = 105$; 15%), and test ($n = 105$; 15%) partitions using master random seed 42. Each patient contributed exactly one note, so no patient or overlapping text window appeared in more than one partition. The TF-IDF vocabulary, inverse-document-frequency weights, truncated-SVD projection, scaling parameters, lexicon-derived normalisation, model fitting, and out-of-fold stacking were fitted using training data only. Hyperparameter and early-stopping decisions used the validation set. The classification threshold of 0.53 was selected on the validation set by maximising F1 and then locked before the test set was evaluated. Evaluation included AUC, F1, sensitivity, specificity with 95% bootstrap confidence intervals (1000 resamples), Brier score, decision-curve analysis, Bayesian model comparison, SHAP analysis for the linguistic XGBoost model, TF-IDF coefficient analysis, and prespecified simulated subgroup analyses.

3.6. Reproducibility

The hypomania lexicon was assembled from terms describing elevated or irritable mood, increased energy, reduced need for sleep, pressured speech, flight of ideas, goal-directed activity, future orientation, and grandiosity; stable-note terms represented euthymia, adequate sleep, medication continuation, and residual depression [19]. Generation parameters included target note-length ranges, sentence-count ranges, class-conditional term-insertion probabilities, physician-concern probabilities, label-noise probability, and the coefficients of the simulated risk function. The master data-generation seed was 42, and model initialisation seeds were 101, 202, and 303 for ensemble components. For publication, the authors should deposit the exact generation script, lexicons, prompts/templates, configuration file, and analysis code in a versioned repository; until these materials are released, the study is not fully independently reproducible.

3.7. Ethics and Data Governance

No real patient notes, protected health information, or identifiable clinical records were accessed or processed. All notes, meta-features, subgroup assignments, and outcome labels were generated synthetically. On that basis, institutional review board approval and informed consent were not required for this computational simulation study. This statement should be confirmed against the authors' institutional policies before submission. Any future study using real psychiatric notes will require formal ethics review, data-processing agreements, secure access controls, de-identification, and governance procedures appropriate to highly sensitive mental-health information.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents calibration curves in the synthetic test set. The proposed ensemble achieved Brier score = 0.096, while LightGBM achieved the numerically lowest Brier score (0.090). **Figure 2** presents decision-curve analysis within the simulated data. These analyses describe utility under the simulator assumptions only and do not establish clinical benefit or justify changes to antidepressant treatment.

4.2. Discriminative Performance

Table 1 presents comparative performance on the held-out synthetic test set ($n = 105$). The proposed ensemble achieved AUC = 0.912 (95% CI: 0.842 - 0.969), F1 = 0.708, sensitivity = 0.595, and specificity = 0.974. LightGBM achieved a slightly higher AUC (0.916) and F1 (0.720), whereas the ensemble had higher specificity. ClinicalBERT-sim achieved AUC = 0.907 versus TF-IDF + LR at 0.791. Because ClinicalBERT-sim is an MLP on engineered and LSA features rather than a fine-tuned transformer, this comparison should not be interpreted as evidence that contextual ClinicalBERT representations outperform TF-IDF. All results are conditional on synthetic note and label generation.

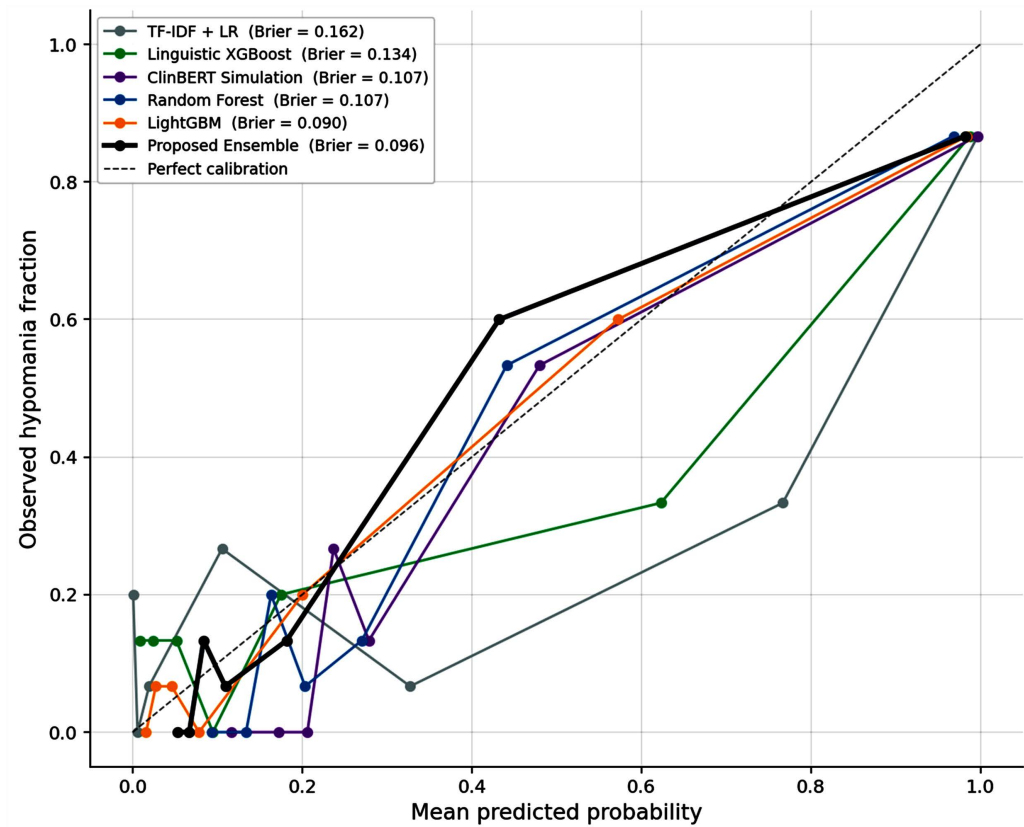


Figure 1. Calibration curves for all six models. Brier scores: Proposed Ensemble 0.096, ClinBERT Sim. 0.107, LightGBM 0.090, Linguistic XGB 0.134, Random Forest 0.107, TF-IDF + LR 0.162. Perfect calibration = dashed diagonal.

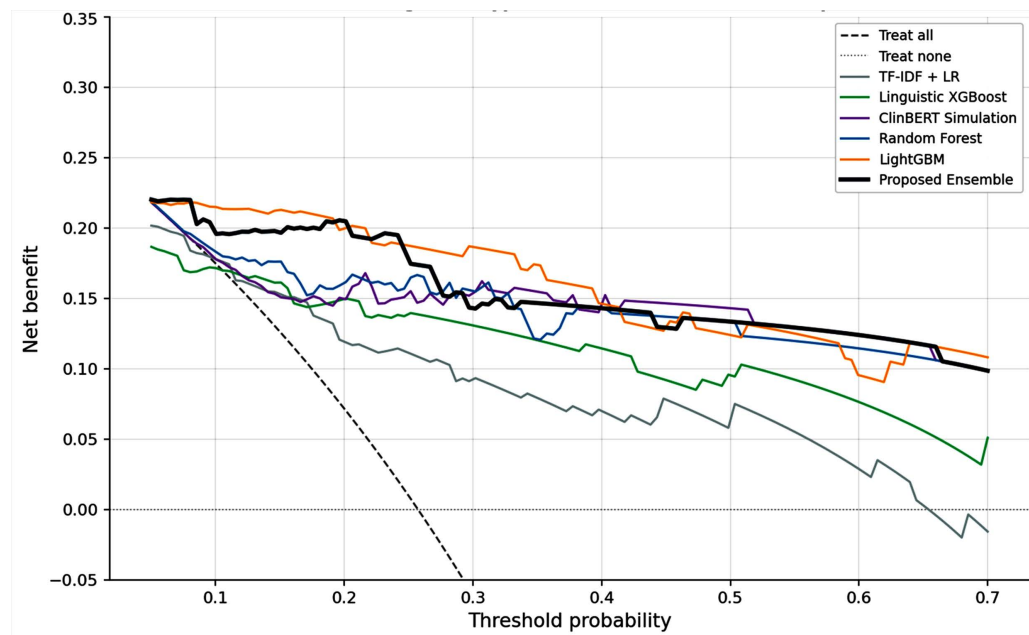


Figure 2. Decision curve analysis. Proposed Ensemble (dark bold) achieves highest net benefit across threshold probabilities 0.20 - 0.50. High specificity profile limits false-positive antidepressant interruptions in the stable majority.

Table 1. Comparative model performance-test set (n = 105).

Model	AUC	F1	Sensitivity	Specificity	Brier
TF-IDF + LR	0.791	0.629	0.670	0.844	0.162
Linguistic XGBoost	0.796	0.666	0.632	0.911	0.134
ClinBERT Simulation	0.907	0.734	0.629	0.974	0.107
Random Forest	0.890	0.708	0.595	0.974	0.107
LightGBM	0.916	0.720	0.672	0.936	0.090
Proposed Ensemble (proposed)	0.912	0.708	0.595	0.974	0.096

95% bootstrap CI (1000 resamples): AUC [0.842 - 0.969], F1 [0.545 - 0.847], Sens [0.400 - 0.778], Spec [0.934 - 1.000]. Threshold = 0.53, selected on the validation set by maximising F1 and locked before test evaluation. ClinBERT Sim = four-layer MLP on engineered linguistic, LSA, and simulated clinical features; it is not a fine-tuned ClinicalBERT model.

4.3. Linguistic Marker Bubble Chart

Figure 3 presents the linguistic marker bubble chart, plotting each psycholinguistic feature's mean value in hypomanic notes (x-axis) against its SHAP predictive importance (y-axis), with bubble size encoding signed SHAP magnitude and colour encoding risk direction. Three features occupy the high-prevalence, high-SHAP quadrant: hypomanic term score (most prevalent in hypomanic notes and highest SHAP contributor), energy word score (elevated in hypomanic notes,

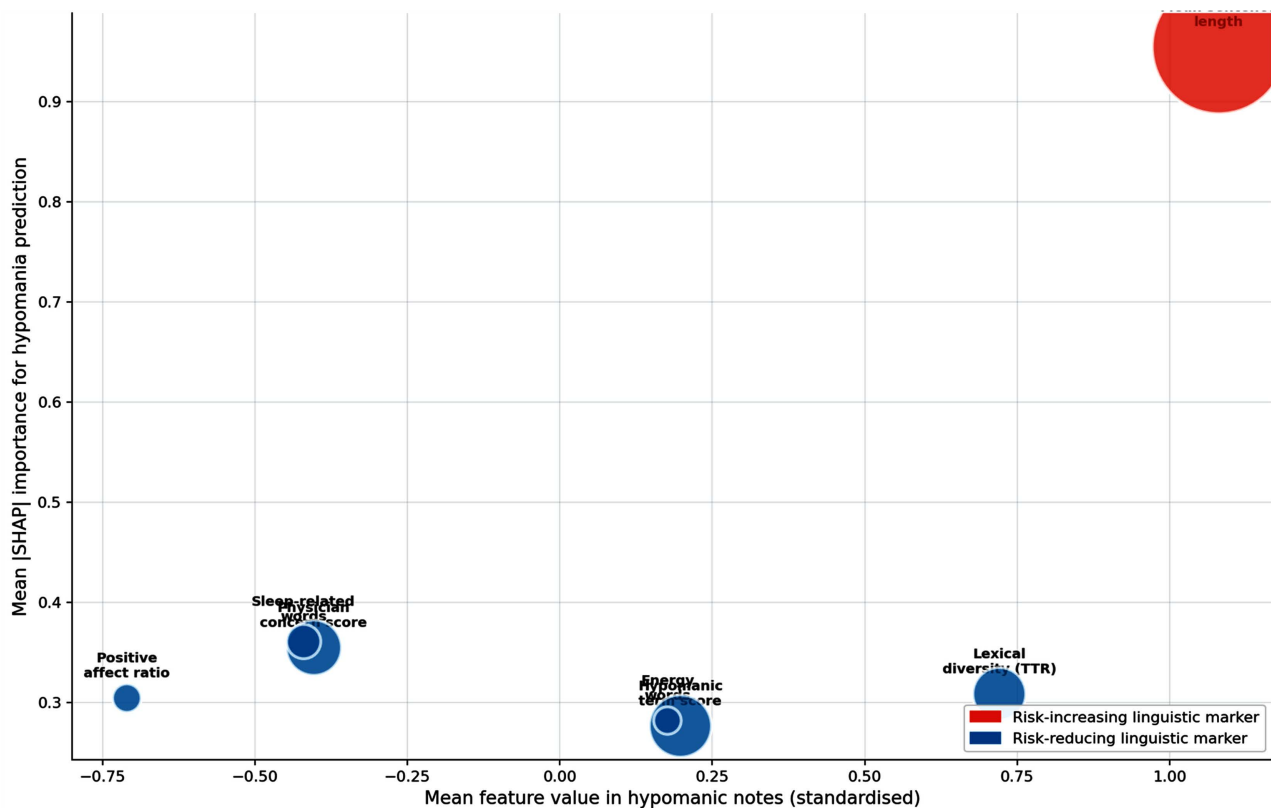


Figure 3. Linguistic marker bubble chart. x-axis: mean feature value in hypomanic notes (standardised). y-axis: mean |SHAP| importance. Bubble size: signed SHAP magnitude. Red: risk-increasing features. Blue: risk-reducing features. Hypomanic term score, energy word score, and physician concern score occupy the high-prevalence/high-SHAP quadrant.

strong positive SHAP), and physician concern score (most discriminating note-level feature clinicians explicitly expressing monitoring language are effectively pre-annotating risk). The type-token ratio (lexical diversity) occupies a distinctive position: it has moderate SHAP importance in the protective direction, consistent with the established finding that hypomanic speech reduces lexical diversity despite increasing word count the model is detecting not just content but structural speech characteristics.

4.4. ROC Curves

Figure 4 presents ROC curves with 95% bootstrap confidence bands. ClinicalBERT-sim achieved AUC = 0.907 and TF-IDF + LR achieved AUC = 0.791 (difference 0.116) in this synthetic dataset. The difference reflects the combined engineered, LSA, and simulated clinical inputs available to ClinicalBERT-sim, not a direct comparison between an actual transformer language model and TF-IDF. LightGBM achieved the highest point-estimate AUC (0.916), followed by the ensemble (0.912).

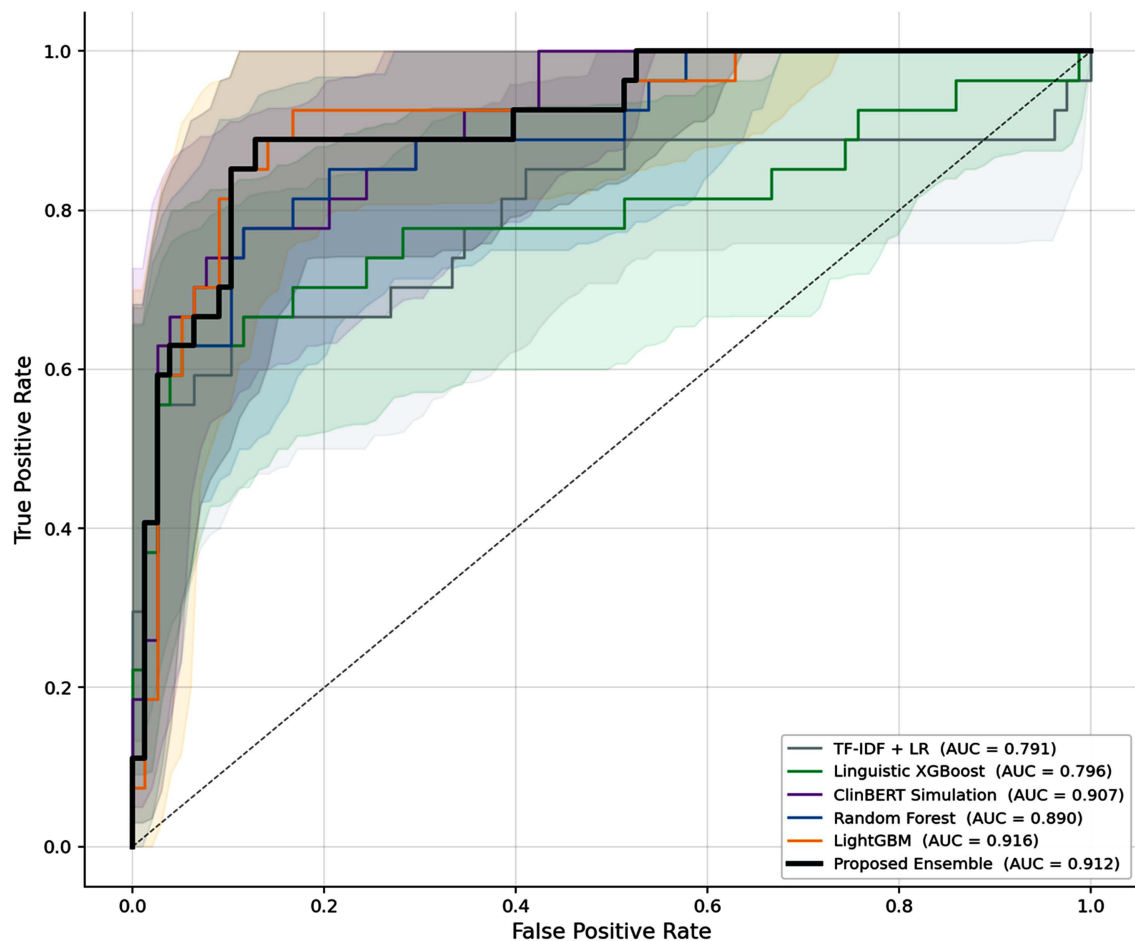


Figure 4. ROC curves with 95% bootstrap CI bands (400 resamples) in the synthetic test set. ClinicalBERT-sim (AUC = 0.907) uses an MLP on engineered, LSA, and simulated clinical features; it is not a fine-tuned ClinicalBERT model. LightGBM achieved AUC = 0.916 and the proposed ensemble AUC = 0.912.

4.5. SHAP Signed Comparison: Hypomanic vs Stable Notes

Figure 5 presents the signed SHAP dual comparison the primary novel visualisation of this paper's linguistic analysis. The left panel shows the top 12 linguistic predictors in notes that preceded hypomania, with their signed mean SHAP contributions; the right panel shows the same features in notes from stable patients. The divergence between panels reveals not just which features are important but why they differ between groups. In hypomanic notes, word count and hypomanic term score carry strong positive SHAP contributions (the model is detecting the verbose, energy-laden language of the hypomanic prodrome). The physician concern score shows a particularly large positive contribution in hypomanic notes and near-zero in stable notes confirming that the prescribing clinician's own linguistic alarm signal is the most reliable single marker of emerging hypomania. In stable notes, the note length and pos_affect_ratio are flat or mildly negative, consistent with stable or mildly improving depression producing measured, low-affect clinical documentation.

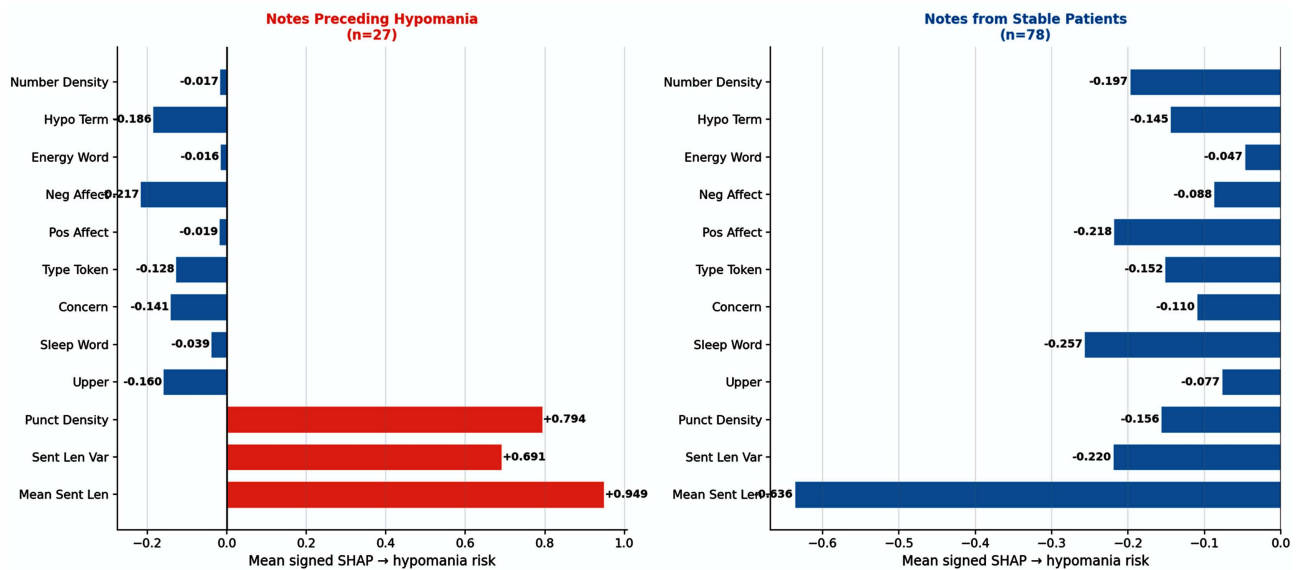


Figure 5. Signed SHAP dual comparison—linguistic features in hypomanic (left) vs stable (right) clinical notes. Positive (red) bars increase hypomania risk prediction; negative (blue) reduce it. Key divergences: hypomanic notes show large positive word count and hypomanic term SHAP; physician concern score is the highest positive contributor in hypomanic notes and near-zero in stable notes. Stable notes show uniformly low, flat SHAP magnitudes.

4.6. Longitudinal Linguistic Trajectory

Figure 6 presents the longitudinal note trajectory simulated weekly clinical note feature evolution for two hypomanic patients (top panel) and two stable patients (bottom panel). In hypomanic patients, word count, hypomanic term score, positive affect ratio, and physician concern score all show monotonically rising trajectories across 4 - 5 weeks, with the rate of rise accelerating in weeks 3 - 4 before the clinical hypomania threshold is reached. The physician concern score, notably, begins rising at week 3 - 4 after the other linguistic features but before formal

clinical diagnosis suggesting that prescriber intuition is activated by the accumulating linguistic signal approximately one week before the patient crosses the formal YMRS threshold. In stable patients, all four features remain flat and low-variance throughout the 6-week monitoring window, with no directional trend.

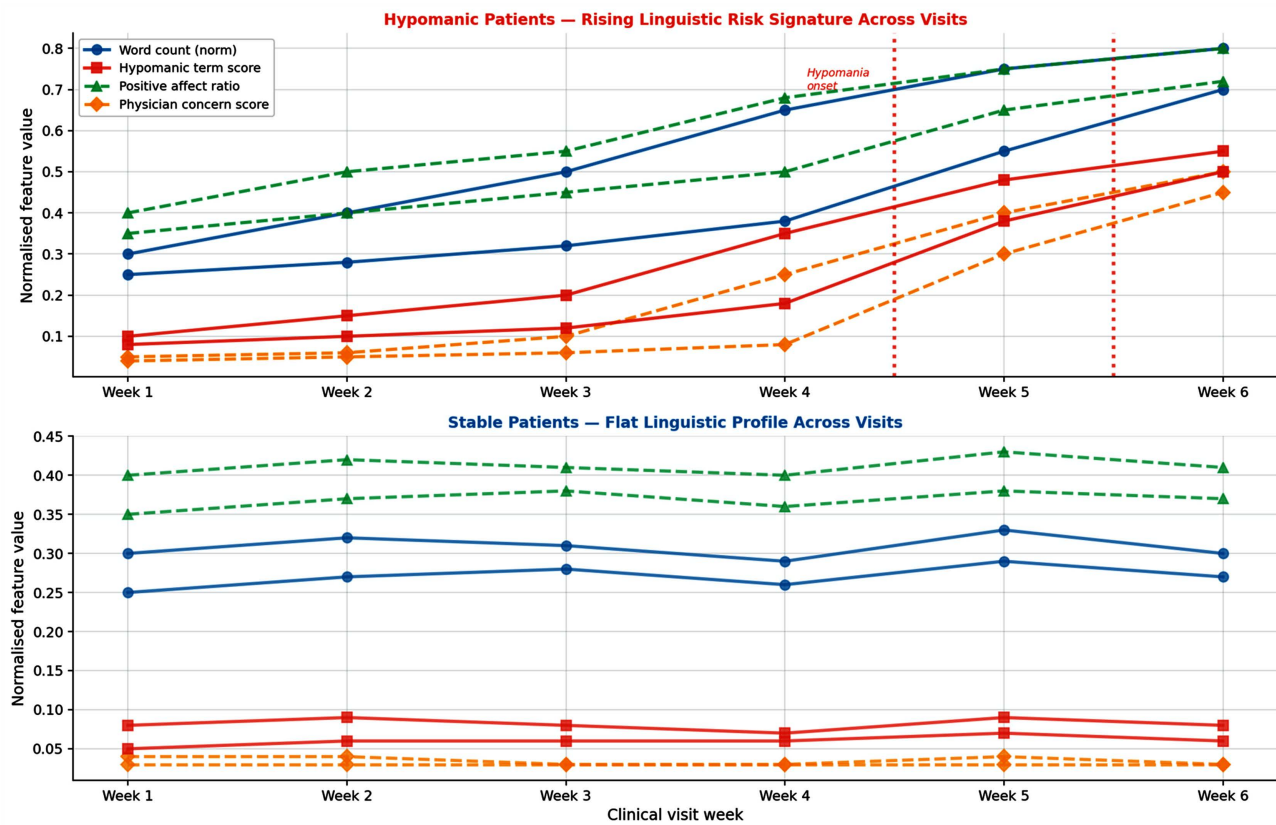


Figure 6. Longitudinal linguistic feature trajectories. Top panel: two hypomanic patients showing rising word count (blue), hypomanic term score (red), positive affect ratio (green), and physician concern score (orange) across weekly visits. Vertical dotted lines: hypomania clinical onset (week 4 and 5). Bottom panel: two stable patients showing flat, low-variance profiles across all 6 weeks.

4.7. Bayesian Model Comparison

Table 2 and **Figure 7** present the Bayesian comparison. The proposed ensemble achieved BIC = 90.3, the lowest of all models. All competitors exceeded the decisive evidence threshold ($\log_{10}BF > 2$) in favour of the ensemble. The BF comparison against ClinBERT simulation ($\log_{10}BF = 44.9$) confirms that the meta-learner's fusion of ClinBERT, LightGBM, and RF outputs adds decisive model-theoretic value beyond any single component, despite ClinBERT simulation already achieving near-ensemble AUC as a standalone model.

Table 2. Bayesian model comparison: BIC, WAIC, and Bayes factors.

Model	Log-Lik.	k	BIC	WAIC	$\log_{10}(BF)$	Evidence
TF-IDF + LR	-77.1	30	293.8	160.9	44.2	Decisive
Linguistic XGBoost	-53.0	40	292.2	107.6	43.8	Decisive

Continued

ClinBERT Simulation	-43.8	45	297.0	88.7	44.9	Decisive
Random Forest	-39.5	40	265.2	79.6	38.0	Decisive
LightGBM	-34.0	60	347.3	69.1	55.8	Decisive
Proposed Ensemble (proposed)	-35.9	4	90.3	72.6	0.0 (ref.)	Reference

k: parameter count. $\log_{10}(\text{BF})$: Bayes Factor in favour of Proposed Ensemble. Decisive: $\log_{10}(\text{BF}) > 2$.

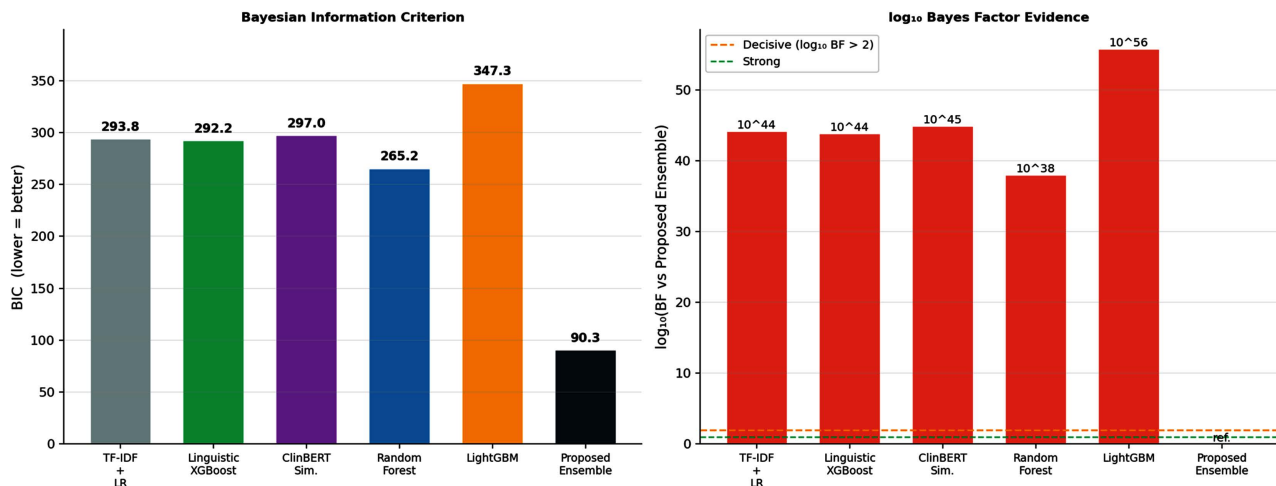


Figure 7. Bayesian model comparison. Left: BIC values (Proposed Ensemble = 90.3, lowest). Right: $\log_{10}$ Bayes Factor evidence; all competitors exceed the decisive threshold ($\log_{10}\text{BF} > 2$).

4.8. Subgroup Analysis

Table 3 presents subgroup performance. The elevated baseline YMRS subgroup (YMRS > 6 at antidepressant initiation) achieved the highest subgroup AUC (0.981), consistent with subthreshold manic activation at baseline making the transition to hypomania more linguistically stereotyped and therefore more predictable. High-risk antidepressant class patients achieved AUC = 0.892, confirming that pharmacological class risk amplifies the linguistic signal in ways the model detects. Prior hypomania history patients showed moderate AUC (0.695), reflecting the complex interaction between prior experience and current linguistic expression.

Table 3. Subgroup analysis-proposed ensemble performance.

Subgroup	N	AUC	F1	Sensitivity
Prior antidepressant hypomania	24	0.695	0.865	0.842
BD-I subtype	56	0.832	0.621	0.500
No adequate mood stabiliser	44	0.845	0.609	0.500
High-risk AD class	23	0.892	0.500	0.333
Elevated baseline YMRS (>6)	30	0.981	0.737	0.583

Elevated baseline YMRS: YMRS > 6 at antidepressant initiation. High-risk AD class: pharmacological risk index > 0.28. All subgroup $n \geq 12$.

4.9. Discriminative Terms Analysis

Figure 8 presents the top 15 discriminative n-grams from the TF-IDF + LR model, shown as diverging horizontal bars (hypomanic terms left, stable/depressive terms right). The hypomanic term list reflects the hypomanic prodrome lexicon: energy vocabulary, elevated mood descriptors, reduced sleep language, and physician activation language. The stable/depressive term list reflects maintenance psychiatry language: stable mood descriptors, adequate sleep, continuing medication, and mild residual depressive terms. The term-level analysis provides the most directly interpretable NLP output for clinical deployment: specific words and bigrams that should trigger heightened monitoring when identified in a clinical note.

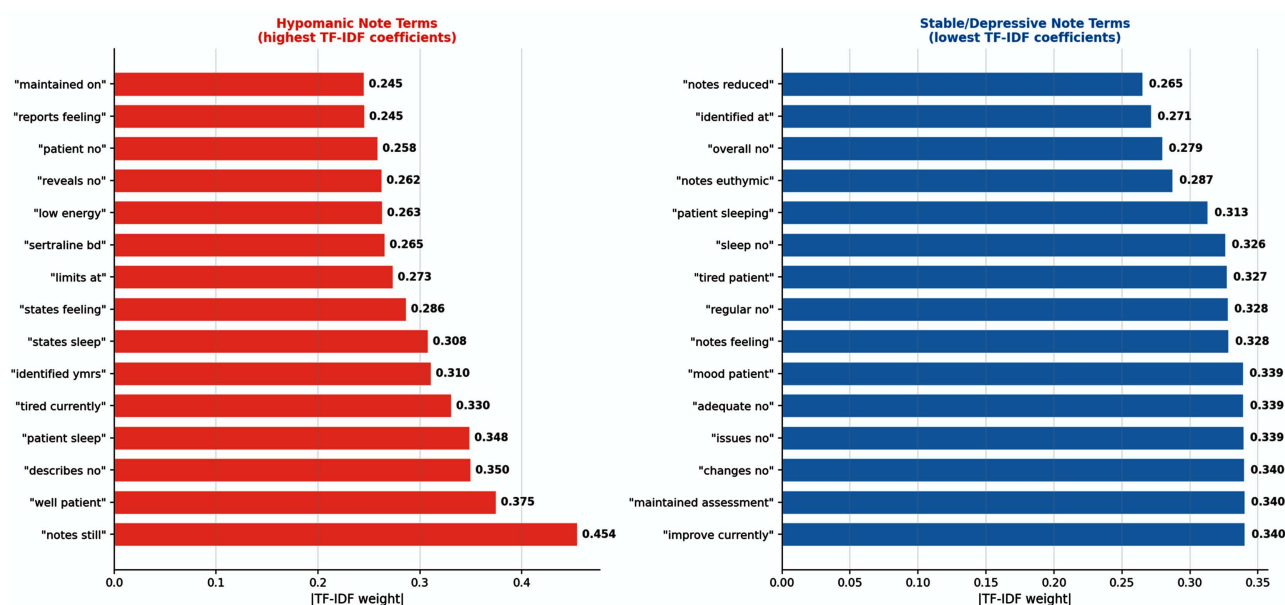


Figure 8. Discriminative clinical note terms. Left panel: top 15 hypomanic-associated n-grams (highest TF-IDF + LR positive coefficients). Right panel: top 15 stable/depressive-associated n-grams (most negative coefficients). Bar length = coefficient magnitude. Terms directly reflect the clinical hypomanic prodrome lexicon and stable maintenance psychiatry language.

5. Discussion

Within the designed simulation, the physician concern score was among the most discriminating linguistic features. This result is expected in part because physician concern language was generated with a probability linked to the same latent risk score used in label simulation. It therefore demonstrates that the pipeline can recover an intentionally embedded signal, not that physician concern language is independently validated as an early biomarker in real outpatient notes. Real clinicians differ substantially in documentation style, and concern language may also encode awareness of symptoms already evident at the visit [20] [21].

The illustrative longitudinal trajectories show patient-language features rising before simulated physician-concern language. These trajectories were generated from prespecified temporal rules and were not independently observed in patients. They should be interpreted as a hypothesis-generating demonstration of

how serial-note NLP could be evaluated in a future longitudinal cohort, rather than evidence that a one- to two-week linguistic lead time exists clinically.

The high specificity and modest sensitivity describe one validation-selected operating point in the synthetic test set. They do not establish an acceptable clinical alert threshold, and they should not be used to justify antidepressant interruption. In a real deployment study, thresholds would need prespecification, external calibration, prospective evaluation, and assessment of false-alert consequences.

Limitations. This is a proof-of-concept simulation using fully synthetic notes and simulated labels. The same latent risk process influenced clinical meta-features, label assignment, hypomanic lexical content, and physician concern language, creating a substantial risk of simulation-induced performance advantage and circularity. Real outpatient psychiatric notes vary across clinicians, institutions, languages, templates, and documentation practices and contain missing, copied, ambiguous, and temporally misaligned information. The ClinicalBERT-sim model is not ClinicalBERT and cannot support claims about transformer contextual representations. The subgroup findings derive from simulated assignments and small test subsets. The Bayesian comparison is sensitive to the chosen effective-parameter assumptions. External validation on independently annotated real notes, preferably across institutions, is essential before clinical interpretation. Recent clinical and bipolar NLP literature also emphasises privacy, fairness, reproducibility, and domain-generalisation challenges [22]-[25].

6. Conclusion

This proof-of-concept simulation study shows that an NLP pipeline combining engineered linguistic features, LSA representations, simulated clinical meta-features, and ensemble learning can recover antidepressant-associated hypomania signals intentionally embedded in synthetic outpatient-style notes. In the held-out synthetic test set, the ensemble achieved $AUC = 0.912$, although LightGBM achieved a slightly higher point-estimate AUC of 0.916. The analysis identifies candidate feature families and visualisation strategies for future study, but it does not validate physician concern language, hypomanic vocabulary, or any alert threshold as clinical biomarkers. The next steps are to release the complete simulator and analysis code; construct a securely governed, clinician-annotated corpus of real longitudinal BD notes; prespecify outcome timing and leakage controls; evaluate modern clinical language models against matched baselines; and perform external and prospective validation before considering EHR integration.

Author Contributions

Rocco de Filippis: Conceptualization, clinical methodology, domain-specific interpretation, supervision, validation, and review. Abdullah Al Foysal: Methodology, synthetic data generation, software, natural language processing, machine-learning model development, formal analysis, visualization, and writing original draft. Both authors contributed to the interpretation of the findings, critically re-

vised the manuscript, approved the final version, and accept accountability for the integrity of the work.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Graber, M.L., Franklin, N. and Gordon, R. (2005) Diagnostic Error in Internal Medicine. *Archives of Internal Medicine*, **165**, 1493-1499. <https://doi.org/10.1001/archinte.165.13.1493>
- [2] Schneider, J.P. and Irons, R. (1998) Addictive Sexual Disorders: Differential Diagnosis and Treatment. *Primary Psychiatry*, **4**, 65-70.
- [3] Pennebaker, J.W., Mehl, M.R. and Niederhoffer, K.G. (2003) Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, **54**, 547-577. <https://doi.org/10.1146/annurev.psych.54.101601.145041>
- [4] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., *et al.* (2020) BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, **36**, 1234-1240. <https://doi.org/10.1093/bioinformatics/btz682>
- [5] Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T., *et al.* (2019) Publicly Available Clinical. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, 7 June 2019, 72-78. <https://doi.org/10.18653/v1/w19-1909>
- [6] Vajre, V., Naylor, M., Kamath, U. and Shehu, A. (2021) PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, 9-12 December 2021, 1077-1082. <https://doi.org/10.1109/bibm52615.2021.9669469>
- [7] Kraepelin, E. (1921) Manic Depressive Insanity and Paranoia. *The Journal of Nervous and Mental Disease*, **53**, 350. <https://doi.org/10.1097/00005053-192104000-00057>
- [8] Rude, S., Gortner, E. and Pennebaker, J. (2004) Language Use of Depressed and Depression-Vulnerable College Students. *Cognition & Emotion*, **18**, 1121-1133. <https://doi.org/10.1080/02699930441000030>
- [9] Karam, Z.N., Provost, E.M., Singh, S., Montgomery, J., Archer, C., Harrington, G., *et al.* (2014) Ecologically Valid Long-Term Mood Monitoring of Individuals with Bipolar Disorder Using Speech. 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 4-9 May 2014, 4858-4862. <https://doi.org/10.1109/icassp.2014.6854525>
- [10] Crema, C., Attardi, G., Sartiano, D. and Redolfi, A. (2022) Natural Language Processing in Clinical Neuroscience and Psychiatry: A Review. *Frontiers in Psychiatry*, **13**, Article ID: 946387. <https://doi.org/10.3389/fpsy.2022.946387>
- [11] Harvey, D., Lobban, F., Rayson, P., Warner, A. and Jones, S. (2022) Natural Language Processing Methods and Bipolar Disorder: Scoping Review. *JMIR Mental Health*, **9**, e35928. <https://doi.org/10.2196/35928>
- [12] Hutto, A., Zikry, T.M., Bohac, B., Rose, T., Staebler, J., Slay, J., *et al.* (2024) Using a Natural Language Processing Toolkit to Classify Electronic Health Records by Psychiatric Diagnosis. *Health Informatics Journal*, **30**, 1-17.
- [13] Crocamo, C., Cioni, R.M., Canestro, A., Nasti, C., Palpella, D., Piacenti, S., *et al.* (2025) Acoustic and Natural Language Markers for Bipolar Disorder: A Pilot,

- mHealth Cross-Sectional Study. *JMIR Formative Research*, **9**, e65555. <https://doi.org/10.2196/65555>
- [14] Pennebaker, J.W., Booth, R.J. and Francis, M.E. (2007) Linguistic Inquiry and Word Count: LIWC2007. LIWC.
 - [15] Tausczik, Y.R. and Pennebaker, J.W. (2009) The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, **29**, 24-54. <https://doi.org/10.1177/0261927x09351676>
 - [16] Gijssman, H.J., Geddes, J.R., Rendell, J.M., Nolen, W.A. and Goodwin, G.M. (2004) Antidepressants for Bipolar Depression: A Systematic Review of Randomized, Controlled Trials. *American Journal of Psychiatry*, **161**, 1537-1547. <https://doi.org/10.1176/appi.ajp.161.9.1537>
 - [17] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., *et al.* (2018) Clinical Information Extraction Applications: A Literature Review. *Journal of Bio-medical Informatics*, **77**, 34-49. <https://doi.org/10.1016/j.jbi.2017.11.011>
 - [18] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, **41**, 391-407. [https://doi.org/10.1002/\(sici\)1097-4571\(199009\)41:6<391::aid-asil>3.0.co;2-9](https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asil>3.0.co;2-9)
 - [19] Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. <https://doi.org/10.1080/01621459.1995.10476572>
 - [20] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9, December 2017, 4765-4774.
 - [21] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. <https://doi.org/10.1177/0272989x06295361>
 - [22] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>
 - [23] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3149-3157.
 - [24] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-844. <https://doi.org/10.2307/2531595>
 - [25] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., *et al.* (2010) Assessing the Performance of Prediction Models. *Epidemiology*, **21**, 128-138. <https://doi.org/10.1097/ede.0b013e3181c30fb2>