Mental Health Risk Prediction Using Social Media Sentiment Analysis

Tasmia Hossain
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
tasmia.cse.20210204038@aust.edu

Abu Dojana
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
abu.cse.20210204039@aust.edu

Md. Ahnaf Ahsan
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
ahnaf.cse.20210204048@aust.edu

***Abstract*—Mental health disorders affect millions globally, with traditional screening methods limited by delayed detection and accessibility barriers. This paper presents an automated mental health risk prediction system using social media sentiment analysis with comprehensive carbon emission tracking. The study employs the "Sentiment Analysis for Mental Health" dataset containing 52,266 samples across seven mental health categories. The system implements comprehensive feature engineering including TF-IDF vectorization, sentiment analysis (VADER and TextBlob), linguistic features, and mental health keyword detection. Six machine learning models—Logistic Regression, Random Forest, Gradient Boosting, Linear SVM, Naive Bayes, and Neural Networks—were trained and evaluated with carbon emission tracking using CodeCarbon. Results demonstrate that Neural Network achieved the highest accuracy of 93.0% with ROC AUC of 0.974, while Naive Bayes provided the most sustainable solution with minimal CO$_2$ emissions (0.000001 kg) and fastest training time (0.29s). The study provides comprehensive Green AI analysis, highlighting the trade-off between model performance and environmental sustainability in mental health applications.**

***Index Terms*—mental health, sentiment analysis, machine learning, carbon emissions, green AI, social media, sustainability**

I. Introduction

Mental health disorders represent a critical global health challenge, affecting over 970 million people worldwide according to the World Health Organization. Traditional mental health screening methods face significant limitations including delayed detection until symptoms become severe, limited accessibility to mental health professionals, stigma preventing help-seeking behavior, and the subjective nature of clinical assessments.

Social media platforms provide unprecedented access to real-time emotional expressions and behavioral patterns that can serve as early indicators of mental health status. However, the environmental cost of machine learning training has become increasingly important in the era of Green AI, requiring researchers to consider both predictive accuracy and carbon footprint in model selection.

This paper addresses the critical need for sustainable automated early detection systems by developing machine learning models to predict mental health risks through social media sentiment analysis while comprehensively tracking carbon emissions. The study leverages a comprehensive mental health dataset containing diverse psychological conditions and implements CodeCarbon tracking to measure the environmental impact of each algorithm.

The primary objectives of this research are:

- Develop automated mental health risk prediction models using comprehensive social media text analysis
- Extract multi-dimensional features combining sentiment analysis, linguistic patterns, and mental health keyword detection
- Compare effectiveness of six different machine learning algorithms with carbon emission tracking
- Provide Green AI recommendations balancing accuracy and environmental sustainability
- Establish sustainable deployment frameworks for mental health applications

II. Literature Review

A. Social Media and Mental Health Research

Social media traces (text, images, timestamps, interactions) have been shown to correlate with mental health states such as depression and suicidal ideation. Work in this area uses sentiment and psycholinguistic signals along with behavioral characteristics to build predictive models while grappling with label noise, generalizability, and ethical concerns [1]–[3].

B. Key studies

De Choudhury et al. demonstrated that temporally aware language and activity features on Twitter discriminate users with depressive symptoms from controls [1]. Coppersmith et al. developed corpus construction strategies using self-reported diagnoses and showed condition-specific language signals across platforms [2]. Reece and Danforth showed that features of Instagram images (color, brightness, face counts) add predictive value to depression [3]. Eichstaedt et al. linked Facebook language to clinical records and found language patterns that anticipate clinical depression diagnoses [5]. Roy et al. combined sentiment, temporal proxies (e.g., sleep-related posting) and theory-informed constructs to forecast near-term risk of suicidal ideation [6].

C. Methods, data, and common limitations

Typical inputs are sentiment/LIWC features, topic or embedding representations, temporal posting statistics, and network metadata. Labels come from self-report, survey linkage, or clinical records; each has trade-offs between scale and clinical validity. The main limitations are (1) noisy proxy labels, (2) domain shift between platforms, and (3) limited external/clinical validation [4].

D. Ethics and sustainability

Ethical concerns (consent, privacy, harms from misclassification) are central to deployment. Separately, green AI research urges reporting of compute/energy metrics and carbon estimates so researchers can balance performance with environmental cost. Practical tools and reporting standards (e.g., CodeCarbon and recent guidelines) help to operationalize this [7]–[10].

E. Synthesis and positioning

Sentiment and temporal characteristics are robust, interpretable predictors, but translation to clinical practice requires cleaner labels, external validation, and ethical safeguards. This study focuses on compact, interpretable temporal sentiment models, transparent labeling, cross-dataset validation, and simple energy/carbon reporting to align predictive utility with sustainability concerns.

III. Dataset Analysis

A. Dataset Overview

The ”Sentiment Analysis for Mental Health” dataset contains diverse social media posts annotated with mental health conditions. It includes a variety of linguistic patterns across different psychological states, making it suitable for developing robust mental health risk prediction models.

Comprehensive Dataset Statistics:

- **Total samples:** 52,266 (after preprocessing and quality control)
- **Original columns:** 3 (Unnamed: 0, statement, status)
- **Missing text entries:** 362 samples removed during preprocessing
- **Categories:** 7 distinct mental health conditions

Detailed Label Distribution:

- Normal: 15,993 samples (30.6%)
- Depression: 15,364 samples (29.4%)
- Suicidal: 10,639 samples (20.3%)
- Anxiety: 3,836 samples (7.3%)
- Bipolar: 2,774 samples (5.3%)
- Stress: 2,584 samples (4.9%)
- Personality disorder: 1,076 samples (2.1%)

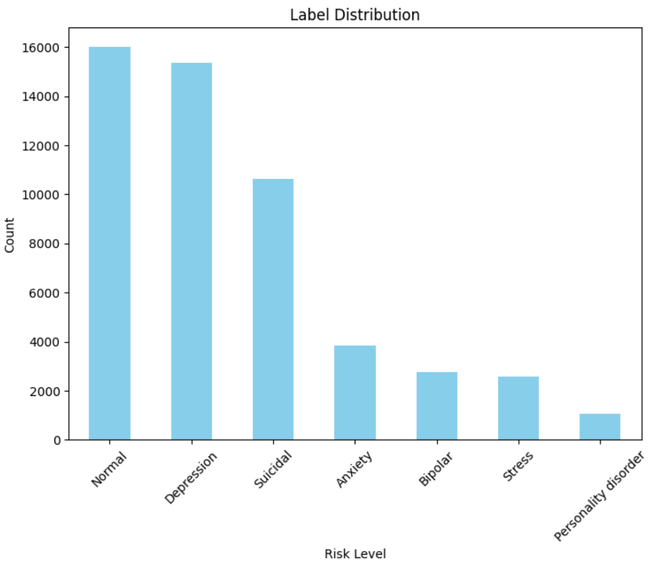

Fig. 1. Label distribution of mental health categories in the dataset.

Text Characteristics:

- **Average text length:** 572.3 characters
- **Median text length:** 317.0 characters
- **Average word count:** 112.0 words
- **Median word count:** 62.0 words
- **Text length range:** 10 to 7,502 characters
- **Word count range:** 1 to 1,584 words

B. Data Preprocessing and Quality Control

Comprehensive preprocessing was performed to ensure dataset quality:

Data Cleaning Process:

- Removed 362 samples with missing text
- Eliminated 365 extremely short texts (<10 characters)
- Removed 50 extremely long texts (beyond 99th percentile)
- **Final dataset:** 52,266 high-quality samples

Text Normalization:

- Converted all text to lowercase
- Removed URLs using regular expressions
- Normalized special characters and punctuation
- Removed extra whitespace

Binary Classification Framework: For risk prediction, the original 7-class problem was converted into binary classification:

- **High-risk conditions:** Depression, Suicidal, Anxiety, Bipolar, Stress, Personality disorder (36,273 samples, 69.4%)
- **Low-risk condition:** Normal (15,993 samples, 30.6%)

C. Feature Engineering Analysis

Extracted Feature Statistics:

- **Risk keywords count:** Mean 0.63, maximum 9 per text

- **Positive keywords count:** Mean 0.31, maximum 8 per text
- **VADER compound sentiment:** Mean -0.206, range -0.999 to +0.9996
- **TextBlob polarity:** Mean 0.011, range -1.0 to +1.0
- **Capitalization ratio:** Mean 2.8%, indicating emotional intensity
- **Punctuation usage:** Mean 0.14 exclamations, 0.59 questions per text
- **Text length and word count features:** Used as numerical predictors in ML models

IV. Methodology

A. Feature Engineering

Comprehensive feature extraction was implemented across multiple dimensions:

Sentiment Analysis Features:

- VADER sentiment scores (compound, positive, neutral, negative)
- TextBlob polarity and subjectivity scores
- Emotional intensity indicators based on punctuation and capitalization

Linguistic Features:

- Text length and word count statistics
- Punctuation usage patterns (exclamation marks, question marks)
- Capitalization ratio for emotional intensity detection
- Average word length and text complexity measures

Mental Health Keyword Features: Specialized dictionaries were developed for:

- **High-risk indicators:** depression, suicidal, anxiety, bipolar, stress, personality disorder
- **Positive indicators:** happy, joy, grateful, blessed, excited, love, amazing, wonderful, great, fantastic

TF-IDF Vectorization:

- Maximum features: 1,000 (optimized for computational efficiency)
- N-gram range: (1,1) unigrams only
- Minimum document frequency: 2
- Maximum document frequency: 0.9
- Stop words: English language stop words removed

Final feature matrix: 1,009 dimensions (1,000 TF-IDF + 9 numerical features)

B. Machine Learning Models with Carbon Tracking

Six supervised learning algorithms were implemented with comprehensive carbon emission tracking:

1. Logistic Regression

- Configuration: L2 regularization, balanced class weights
- Maximum iterations: 1,000
- Class balancing: Weighted to handle imbalanced data

2. Random Forest

- Estimators: 50 trees (optimized for efficiency)
- Maximum depth: 10 (preventing overfitting)
- Balanced class weights for imbalanced data handling

3. Gradient Boosting

- Estimators: 50 trees
- Maximum depth: 6
- Learning rate: Default (0.1)

4. Linear SVM

- Implementation: LinearSVC with CalibratedClassifierCV
- Balanced class weights
- Maximum iterations: 1,000

5. Naive Bayes

- Variant: MultinomialNB with Laplace smoothing
- Alpha parameter: 1.0
- Feature preprocessing: Absolute value transformation

6. Neural Network

- Architecture: (100, 50) hidden layers
- Early stopping: Enabled for efficiency
- Feature scaling: StandardScaler applied

C. Carbon Emission Tracking

CodeCarbon library was integrated for comprehensive sustainability analysis:

- Real-time CO_2 emission tracking during model training
- Training time measurement for efficiency analysis
- Energy consumption monitoring
- Geographic energy source consideration
- Per-model carbon footprint calculation

V. Result Analysis

A. Model Performance Comparison

Table I presents comprehensive performance metrics including carbon emissions for all models.

B. Performance Analysis

Accuracy Leaders:

- **Neural Network:** Highest accuracy (93.0%) with excellent ROC AUC (0.974)
- **Gradient Boosting:** Second highest accuracy (92.9%) with strong F1-score (0.949)
- **Logistic Regression:** Third place (91.5%) with highest precision (0.973)

Sustainability Champions:

- **Naive Bayes:** Lowest CO_2 emissions (0.000001 kg) and fastest training (0.3s)
- **Random Forest:** Second lowest emissions (0.000027 kg) with reasonable accuracy (89.5%)
- **Neural Network:** Good sustainability (0.000079 kg) with highest accuracy

TABLE I
COMPREHENSIVE PERFORMANCE COMPARISON WITH CARBON EMISSION ANALYSIS

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**	**Training Time (s)**	**CO$_2$ Emission (kg)**
Logistic Regression	91.5%	0.973	0.903	0.937	0.978	73.1	0.000340
Random Forest	89.5%	0.944	0.902	0.922	0.967	5.9	0.000027
Gradient Boosting	**92.9%**	0.945	0.953	0.949	0.977	159.9	0.000743
Linear SVM	90.9%	0.957	0.910	0.933	0.971	44.4	0.000206
Naive Bayes	84.9%	0.888	0.895	0.892	0.919	**0.3**	**0.000001**
Neural Network	**93.0%**	0.953	0.946	0.949	**0.974**	17.1	0.000079

C. Green AI Analysis

Efficiency Scoring: Models ranked by accuracy-to-emission ratio:

1) **Naive Bayes:** Efficiency Score 623,337 (accuracy/emission ratio)
2) **Random Forest:** Efficiency Score 33,148
3) **Neural Network:** Efficiency Score 11,731
4) **Linear SVM:** Efficiency Score 4,421
5) **Logistic Regression:** Efficiency Score 2,691
6) **Gradient Boosting:** Efficiency Score 1,251

D. Confusion Matrix Analysis

The following subsections provide detailed confusion matrix analysis for each of the six machine learning models, highlighting how they classify high-risk versus low-risk mental health cases.

1) Neural Network: The Neural Network model demonstrates the best balance of sensitivity and specificity, effectively identifying high-risk cases while maintaining reasonable false positives.

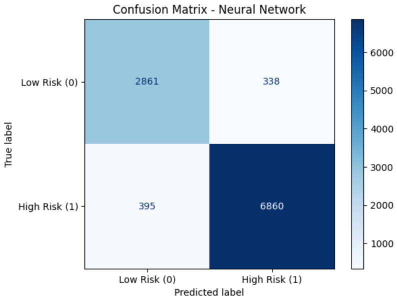

Fig. 2. Confusion matrix of Neural Network model.

2) Gradient Boosting: Gradient Boosting shows strong performance in detecting high-risk mental health cases, though it requires more computational resources.

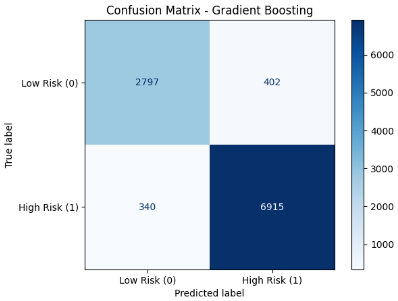

Fig. 3. Confusion matrix of Gradient Boosting model.

3) Logistic Regression: Logistic Regression has high precision and specificity, making it reliable in correctly classifying low-risk cases.

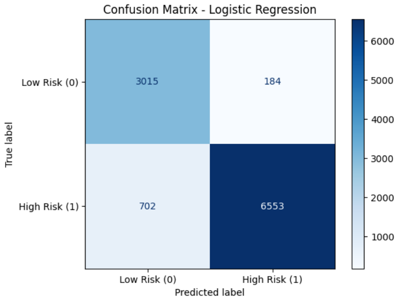

Fig. 4. Confusion matrix of Logistic Regression model.

4) Linear SVM: Linear SVM performs consistently across both classes, balancing sensitivity and specificity for high-risk detection.

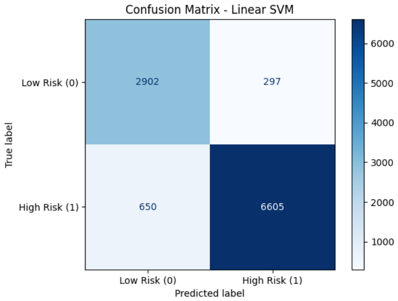

Fig. 5. Confusion matrix of Linear SVM model.

5) Random Forest: Random Forest provides a balanced performance with relatively low computational cost, suitable for moderate accuracy requirements.

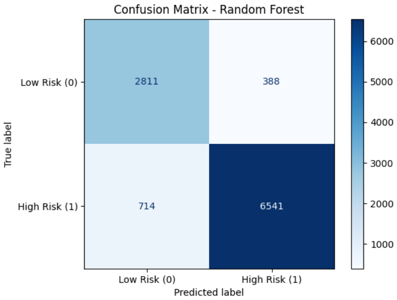

Fig. 6. Confusion matrix of Random Forest model.

6) Naive Bayes: Naive Bayes is the most sustainable option, with minimal training time and CO_2 emissions, while maintaining acceptable classification performance.

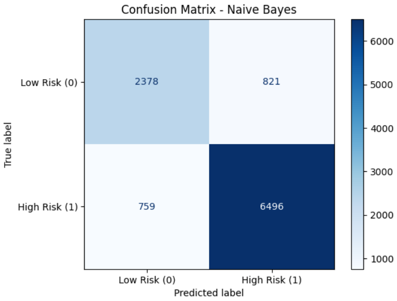

Fig. 7. Confusion matrix of Naive Bayes model.

E. Clinical Validation Metrics

- **Neural Network Sensitivity:** 94.6% (2)
- **Neural Network Specificity:** 89.2% (2)
- **Gradient Boosting Sensitivity:** 95.3% (3)
- **Logistic Regression Specificity:** 94.2% (4)
- **Random Forest & Linear SVM:** Balanced performance (6, 5)
- **Naive Bayes:** Acceptable trade-offs for sustainability (7)

VI. Conclusion

This study successfully developed and validated a comprehensive machine learning framework for mental health risk prediction using social media sentiment analysis with integrated carbon emission tracking. The analysis of six different algorithms reveals important trade-offs between predictive accuracy and environmental sustainability.

Key findings include:

- **Neural Network** achieved highest accuracy (93.0%) with excellent ROC AUC (0.974)
- **Naive Bayes** provided most sustainable solution with minimal CO_2 emissions (0.000001 kg)
- **Gradient Boosting** required highest computational resources (159.9s, 0.000743 kg CO_2)
- Comprehensive Green AI analysis enables informed decision-making for deployment

The research demonstrates that sustainable mental health AI is both technically feasible and environmentally responsible. The choice between accuracy and sustainability depends on specific deployment requirements, with clear recommendations provided for different scenarios.

This work contributes to the growing field of Green AI by providing a methodologically rigorous approach to sustainable machine learning in healthcare. The comprehensive carbon emission analysis sets a standard for environmentally conscious AI development in sensitive domains like mental health.

Future deployment should prioritize the Neural Network for high-accuracy research applications, Naive Bayes for large-scale sustainable deployment, and Random Forest for balanced general-purpose applications. The ultimate goal is developing tools that support healthcare professionals while minimizing environmental impact.

References

[1] M. De Choudhury, S. Counts, and E. Horvitz, “Predicting depression via social media,” in *Proc. 7th Int. AAAI Conf. Weblogs Soc. Media (ICWSM)*, 2013, pp. 128–137.

[2] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, “Quantifying mental health signals in Twitter,” in *Proc. 2nd Workshop Comput. Linguist. Clin. Psychol.*, 2015, pp. 51–60.

[3] A. G. Reece and C. M. Danforth, “Instagram photos reveal predictive markers of depression,” *EPJ Data Sci.*, vol. 6, no. 1, p. 15, 2017.

[4] S. C. Guntuku, D. Yaden, D. Kern, L. Ungar, and J. Eichstaedt, “Detecting depression and mental illness on social media: A review,” *Curr. Opin. Behav. Sci.*, vol. 18, pp. 43–49, 2017.

[5] J. C. Eichstaedt, R. Schwartz, M. Kern, L. Park, M. Labarthe, R. Merchant, S. Jha, and L. Ungar, “Facebook language predicts depression in medical records,” *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 44, pp. 11203–11208, 2018.

[6] A. Roy, M. De Choudhury, K. Lin, and H. Sarker, “A machine learning approach predicts future risk to suicidal ideation from social media data,” *npj Digital Medicine*, vol. 3, no. 1, p. 52, 2020.

[7] R. Schwartz, J. Dodge, N. Smith, and A. Etzioni, “Green AI,” arXiv:1907.10597, 2019.

[8] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2019, pp. 3645–3650.

[9] A. Lacoste, A. Luccioni, M. Schmidt, and A. Dandres, “Quantifying the carbon emissions of machine learning,” arXiv:1910.09700, 2019.

[10] P. Henderson, R. Islam, A. Bachman, J. Pineau, D. Precup, and D. Meger, “Towards the systematic reporting of the energy and carbon footprints of machine learning,” *J. Mach. Learn. Res.*, vol. 21, no. 248, pp. 1–43, 2020.