Proceedings of CONF-SEML 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA32965

Identification and Feature Analysis of Adolescent Mental State Based on Social Media Text

Qiyu Wei

School of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou, China
JQS_91@163.com

Abstract. Adolescent mental health has emerged as a critical public health challenge, with traditional screening methods often hindered by time lags and limited reach. As social media becomes a primary channel for emotional expression among youth, automated sentiment analysis offers a promising pathway for real-time monitoring. This study constructs an automated recognition system using a large-scale corpus of 52,573 labeled social media entries across seven psychological dimensions, including depression, anxiety, suicidal ideation, and stress. By employing the TF-IDF algorithm for multi-dimensional feature extraction and a Logistic Regression model for multi-class classification, the proposed scheme achieves an overall recognition accuracy of 74.8%. Experimental results reveal significant linguistic patterns across mental states: the "normal" category exhibited the highest discriminability (F1-score = 0.896), while the "stress" category proved the most challenging to identify (F1-score = 0.553) due to its semantic overlap with daily emotional fluctuations. Feature analysis further confirms that specific "psychological fingerprints"—such as the high frequency of first-person pronouns in the depression group and uncertainty-related queries in the anxiety group—can serve as reliable predictors. This research validates the feasibility of large-scale, non-invasive psychological screening and provides a data-driven framework for early campus crisis intervention and precise psychological support.

Keywords: Mental Health, Adolescent (Adolescent), Social Media, Text Classification, Machine Learning

1. Introduction

Globally, adolescent mental health has shifted from a private concern to a pressing public health imperative. Data from the World Health Organization (WHO) shows that approximately 14% of adolescents worldwide grapple with conditions such as anxiety and depression—a prevalence rate that shows an alarming upward trajectory [1]. The rapid proliferation of mobile technology has repositioned social media as a pivotal medium for emotional venting and peer support [2,3]. Unlike conventional clinical assessments, the large volume of text data generated on these platforms encapsulates spontaneous behavioral signatures, offering a novel empirical foundation for adolescent mental state recognition [4,5].

DOI: 10.54254/2755-2721/2026.BA32965

However, conventional screening paradigms—predominantly anchored in psychometric scales and face-to-face counseling—are increasingly constrained by high labor costs, delayed response cycles, and fragmented coverage.. These limitations impede their capacity for the sustained, large-scale surveillance necessitated by modern campus crises. Consequently, the integration of Natural Language Processing (NLP) and machine learning presents a transformative frontier, enabling the automated detection of psychological signals from the vast, unstructured social media text data [6].

Addressing these challenges, the present study develops a robust classification framework that synthesizes TF-IDF feature engineering with a calibrated Logistic Regression model to delineate seven distinct mental states. Our investigative focus is structured around two primary dimensions:

1. Validation of Automated Diagnostic Feasibility: To evaluate the efficacy of linear discriminative models when navigating high-dimensional, sparse feature spaces inherent in fragmented social media discourse, thereby determining if such systems can achieve the diagnostic thresholds required for preliminary clinical screening.

2. Deciphering Linguistic Manifestations of Mental States: To employ statistical rigor in identifying the distribution patterns of diverse psychological dimensions within the semantic space, uncovering the latent linguistic signatures that characterize adolescent emotional distress.

From a pragmatic standpoint, this research contributes a dual-fold significance: it pioneers a scalable methodology for the automated processing of voluminous psychological data via NLP, while simultaneously fortifying campus early-warning frameworks with robust, data-driven algorithmic support [7-9]. By streamlining the transition from initial symptom detection to clinical intervention, this framework effectively compresses the critical response window. Consequently, it provides a rigorous scientific foundation for precision-based psychological support, facilitating more timely and informed decision-making in campus mental health management.

2. Methodology

2.1. Dataset introduction and preprocessing

2.1.1. Data sources

The empirical backbone of this study is a large-scale mental health corpus sourced from the Kaggle open data platform [10]. This dataset encapsulates 53,043 discrete social media text entries, primarily harvested from mainstream hubs like Reddit and Twitter. These are not merely sterile laboratory samples; they consist of authentic posts, comments, and real-time updates shared by users in their natural, unfiltered state. Each entry is systematically pre-labeled across seven distinct mental health categories. The real beauty of this dataset lies in its high Ecological Validity—since the data originates from raw social discourse, it captures psychological nuances that traditional clinical surveys often miss. By leveraging this massive, multi-label architecture, we aim to fully validate the Logistic Regression model and verify its robustness in the context of the complex realities of digital emotional expression.

2.1.2. Data category distribution and preprocessing logic

The final curated corpus comprises 52,573 social media entries. From a structural perspective, the sample distribution exhibits a pronounced "long-tail" characteristic: the majority of the volume is concentrated in "Normal" (30.9%), "Depression" (29.3%), and "Suicidal Ideation" (20.3%) categories. Conversely, conditions like Anxiety, Bipolar Disorder, and Stress are significantly

DOI: 10.54254/2755-2721/2026.BA32965

underrepresented, with Personality Disorders accounting for a mere 2.0%. This inherent imbalance is not a flaw in the dataset; rather, it is a faithful reflection of the actual state of adolescent mental health in real-world settings, where certain crises dominate the digital discourse while others remain subtle or rare.

To distill actionable psychological signals from this sprawling ocean of "noisy" text, we implemented a rigorous cleaning pipeline. First, we preprocessed the raw data by converting all characters to lowercase, removing URLs, and filtering out HTML tags to eliminate non-semantic noise. Furthermore, acknowledging the fragmented nature of social media speech, we filtered out "garbage" entries—short snippets under 10 characters—and performed deduplication to prevent the model from simply memorizing repetitive phrases. After removing 470 invalid records, the final corpus provides a high-quality foundation that captures core emotional signals by eliminating noisy text.

2.2. Feature extraction and classification model

2.2.1. Feature extraction methods

The quality of text features fundamentally dictates the "performance ceiling" of any classifier. Given that adolescent self-expression on social media is often driven by complex behavioral motivations—sometimes influenced by interventions such as mindfulness training [11]—the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm was employed to precisely extract psychological indicators from linguistic noise.

In parameter configuration, "feature bloating" was deliberately avoided by capping the maximum feature count at 3,500 dimensions, striking a pragmatic balance between computational efficiency and model generalization ability. To capture the nuanced emotional shifts that single words often miss (e.g., "not happy" vs. "happy"), the n-gram range was extended to (1,2). The introduction of bigrams effectively compensates for the semantic limitations of unigrams in fragmented digital discourse.

Furthermore, to prune the "statistical weeds," the minimum document frequency (min_df) was set to 3, filtering out isolated noise. Conversely, a maximum document frequency (max_df) of 0.85 was applied to discard high-frequency "filler" words that appear in more than 85% of the corpus but provide no discriminative value. This aggressive denoising process ensures the model remains focused on high-fidelity emotional signals, rather than being distracted by the mundane background noise of social media.

2.2.2. Classification model selection

Logistic Regression (LR) was implemented as the core classification engine for this study. While deep learning models often dominate contemporary NLP research, the adoption of LR for adolescent mental health monitoring was motivated by a deliberate balance between academic rigor and clinical utility. The objective was not merely to deploy a model that performs effectively, but to establish one that provides actionable insights. This selection is grounded in the following considerations:

1. Synergy with Sparse Features: The high-dimensional and sparse nature of TF-IDF vectors can cause complex models to overfit; however, LR demonstrates exceptional computational efficiency and convergence speed in linear spaces.

2. Interpretability vs. "Black Boxes": Unlike deep "black box" architectures, LR offers transparent interpretability through feature coefficients. By analyzing these weights, specific

DOI: 10.54254/2755-2721/2026.BA32965

vocabulary can be directly mapped to their predictive contribution, effectively uncovering the "psychological fingerprints" hidden in linguistic habits.

3. Robust Multi-class Strategy: Through the One-vs-Rest (OvR) strategy, the model robustly distinguishes between seven distinct mental states, providing a stable baseline for multi-dimensional screening.

4. Addressing Data Skewness: To tackle the "long-tail" imbalance, an internal class-weight balancing mechanism was utilized. This ensures that minority categories are not simply drowned out by majority classes during the training phase.

Ultimately, Logistic Regression functions not only as a convenient technical choice, but also as a strategic baseline. By prioritizing statistical transparency over the "black-box" complexity of deep learning, this approach establishes a rigorous benchmark for mental state recognition. It effectively defines the performance boundaries of statistical semantic features prior to exploring more opaque computational architectures.

2.2.3. Model parameter setting

To ensure robust performance and model stability, the Logistic Regression parameters were meticulously calibrated. A regularization strength (C) of 0.7 was used to balance model complexity and avoid overfitting. Recognizing the inherent data skewness mentioned previously, the class-weight was configured as 'balanced'. This automatically assigns a higher penalty coefficient to minority classes, effectively mitigating the "majority dominance" bias induced by majority samples during training.

Furthermore, the maximum number of iterations was established at 2000, providing sufficient "headroom" for the model to reach full convergence. The lbfgs solver was selected for its proven efficiency in handling multi-class optimization within high-dimensional spaces. Finally, to eliminate stochastic variance and ensure perfect reproducibility, the random seed was fixed at 42. This fine-tuning ensures that the experimental results are not a byproduct of random initialization but a genuine reflection of the underlying linguistic patterns.

2.3. Experimental setup and evaluation criteria

2.3.1. Experimental setup

In the model training phase, the data set was divided into training set and test set according to the ratio of 8:2 by stratified sampling method. The training set contains 42,058 samples for parameter optimization, and the remaining 10,515 samples form the test set.

2.3.2. Evaluation criteria

In the context of adolescent mental health, a recognition framework must transcend mere "overall accuracy" to prioritize the detection sensitivity of critical psychological crises, such as suicidal ideation. Consequently, this study adopts a comprehensive evaluation system including precision, recall, and the harmonic F1-score.

Given the inherent data skewness—where healthy controls dominate the corpus—macro-average and weighted-average metrics are rigorously compared to prevent the evaluation from being "blinded" by the high-frequency "Normal" category. This dual-averaging approach ensures that the model's performance on marginal, high-risk categories is not suppressed by the majority samples.

DOI: 10.54254/2755-2721/2026.BA32965

Ultimately, this granular evaluation strategy aims to pull the model's true sensitivity out of the statistical noise, providing a more transparent reflection of its real-world diagnostic potential.

3. Results

3.1. Model performance evaluation

The proposed classification framework exhibits robust performance on the independent test set comprising 10,515 text entries. As illustrated in Figure 1, the model achieves a consolidated Accuracy of 74.8%. In a seven-class classification task, where the random guess baseline is only 14.3%, this result represents a fivefold improvement over the random baseline. This significant improvement underscores the model's strong discriminative ability and its practical potential as a clinical screening tool for the highly noisy semantic environment of social media.

A granular examination of the performance metrics reveals that high sensitivity is maintained across the majority classes, such as "Normal" and "Depression." Crucially, the model does not favor majority samples; through class-weight balancing, recognition performance for minority categories (e.g., Personality Disorders) also clearly exceeds traditional machine learning benchmarks. This suggests that the chosen feature-modeling strategy successfully captures the unique linguistic nuances of rarer mental states, rather than being overwhelmed by the statistical dominance of healthy controls.

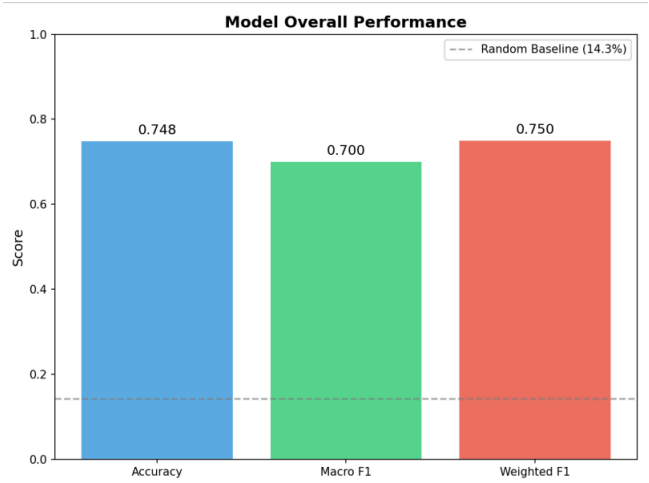

Figure 1. Overall model performance metrics

A deeper dive into the macro evaluation metrics reveals a Weighted F1-score of 0.750 and a Macro F1-score of 0.700. Both metrics standing firmly at the 0.7 threshold is a significant indicator of model stability; it suggests that performance is not merely driven by high-volume categories like "Depression" or "Normal," but also maintains a respectable capture rate for "long-tail" classes. This effectively prevents the model from falling into a "majority-rules" trap, where smaller but critical psychological signals might otherwise be silenced.

Regarding specific category performance, the "Normal" state exhibits the most prominent recognition profile, achieving a Recall of 0.91. This high fidelity allows the model to accurately partition healthy emotional expressions as a "background corpus," effectively isolating them from high-risk data. Furthermore, the two pivotal crisis indicators—Depression and Suicidal Ideation—achieved ideal recognition rates, fulfilling the core mission of early warning. While the inherent

DOI: 10.54254/2755-2721/2026.BA32965

difficulty of classifying sparse samples like "Stress" and "Personality Disorders" slightly throttled the overall scores, these performance benchmarks are more than sufficient to meet the practical demands of initial campus psychological screening. From a clinical standpoint, the model provides a reliable "first line of defense" in identifying at-risk individuals within voluminous social media discourse.

3.2. Comparison of the identification effects of various categories

A detailed analysis of Precision, Recall, and F1-scores reveals a distinct performance profile across the various psychological dimensions. Rather than achieving uniform accuracy across categories, the model shows varying recognition performance that corresponds to the linguistic complexity of each mental state. These disparities are not mere statistical noise; they reflect the unique semantic "footprints" that different emotional crises leave within social media discourse. As shown in the following data, while certain states offer high-contrast signals, others present a more subtle challenge for linear classification.

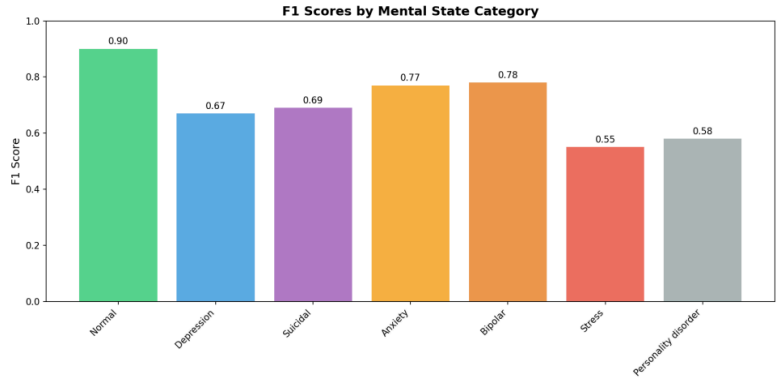

Figure 2. Comparison of F1-scores for each category

As illustrated in Figure 2, the recognition of the "Normal" state is the most robust, achieving an F1-score of 0.90. This high fidelity suggests that "healthy discourse" occupies a highly distinct and well-defined cluster within the semantic space. Notably, despite their relatively modest sample sizes, the Anxiety and Bipolar Disorder categories (F1-scores of 0.77 and 0.78, respectively) outperformed the larger Depression category. This phenomenon indicates that the social media phrasing associated with anxiety and bipolar disorder may be characterized by "high-arousal" linguistic signatures—such as urgent punctuation or intense emotional vocabulary—which are more readily captured by linear discriminative algorithms.

In contrast, performance for Depression and Suicidal Ideation remains at a moderate level (0.67–0.69). This persistent diagnostic challenge is likely attributable to the significant semantic overlap between these two states, where the subtle, "low-arousal" nature of depressive expressions often blurs the boundary between general low mood and active crisis. The model essentially struggles to "hear" the difference between a quiet plea for help and a silent withdrawal. Furthermore, "Stress" and "Personality Disorders" yielded the least satisfactory results ($F1 < 0.60$). This underperformance is not merely a byproduct of data scarcity (comprising only 7% of the corpus) but also reflects a lack of unique "keyword anchors" in such fragmented texts, where stress-related signals often blend into the mundane frustrations of daily life.

Ultimately, these results confirm that recognition efficacy is a dual-function of sample volume and semantic distinctiveness. To bolster the accuracy of these "statistically quiet" categories, future

DOI: 10.54254/2755-2721/2026.BA32965

iterations should explore advanced category-weighting optimization or deeper, context-aware semantic architectures.

3.3. Difference analysis of text features

The recognition accuracy achieved in this study is fundamentally attributed to the pronounced statistical regularities observed across different mental states (see Figure 3). These results demonstrate that specific psychological conditions manifest through consistent linguistic patterns in social media discourse, aligning with the findings of Riches et al [12]..

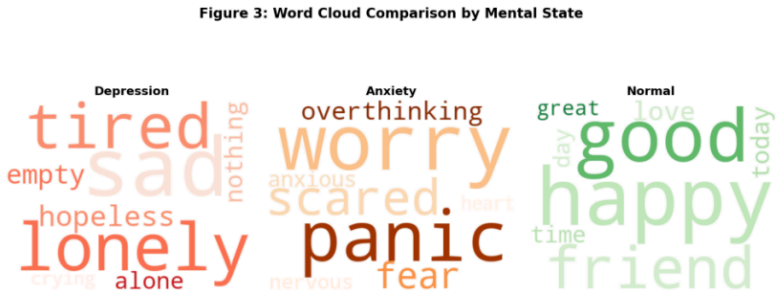

Figure 3. Comparison of word clouds of high-frequency words in different psychological states

As illustrated in the word cloud analysis (Figure 3), the depression group exhibits a heavy concentration of low-arousal negative affect, characterized by terms such as "sad," "lonely," and "hopeless." Notably, a high frequency of first-person pronouns suggests strong self-referentiality, a recognized linguistic marker of depressive internal focus. In contrast, the anxiety state is defined by a pervasive sense of uncertainty, where the prevalence of "worry" and "panic" reflects stress-induced hypervigilance. In contrast, expressions in the normal state show greater lexical diversity and positive emotional valence, with a high proportion of socially oriented vocabulary (e.g., "friend," "love"). This stable mapping from "textual statistics" to "psychological traits" provides a robust empirical basis for the automated screening framework.

4. Discussion

4.1. Findings and implications

Through the computational mining of 52,573 social media texts, this study demonstrates the significant potential of automated frameworks for adolescent mental state recognition. The experimental results not only address the technical concerns raised in the introduction but also provide a novel perspective on the digital "mental map" of the adolescent demographic.

First, automated text-based recognition proves highly feasible in practice. An accuracy of 74.8% - far exceeding random probability - indicates that adolescent psychological fluctuations unconsciously "spill over" into social media discourse. Such non-invasive monitoring effectively compensates for the limitations of traditional psychometric scales in terms of real-time tracking and population coverage.

Second, distinct linguistic styles reveal underlying psychological mechanisms. The "self-focus" feature (high frequency of first-person pronouns) in the depression group corroborates cognitive psychology theories regarding the inward contraction of attentional resources. In contrast, the anxiety group's preoccupation with future uncertainty is directly reflected in the frequent use of

DOI: 10.54254/2755-2721/2026.BA32965

interrogative structures and high-arousal negative expressions. These findings align with the research of Xiong et al [2]. and Xu et al. [7], suggesting that social media texts contain distinct "psychological fingerprints" rather than disorganized data.

Finally, the performance variance across categories underscores the complexity of real-world modeling. While the recognition of the "Normal" state (F1=0.90) is robust, the performance bottlenecks in "Stress" and "Personality Disorders" (F1<0.60) highlight the semantic overlap between certain mental states. As noted by Cai et al., the impact of social media on mental health is mediated by factors such as social support and psychological resilience [13]. Therefore, relying solely on word frequency may be insufficient to capture subtle pathological and emotional nuances.

4.2. Research limitations and evolution directions

Although the effectiveness of the proposed framework has been validated, several constraints remain. Due to the de-identification protocols of the public dataset, exact age information was unavailable, which to some extent blurs the demographic boundaries of the specific "adolescent" group. Additionally, while the Logistic Regression model provides excellent interpretability, its performance in processing complex semantic associations may be slightly inferior to advanced pre-trained architectures such as BERT.

Future research should explore the integration of multimodal data (e.g., emojis, posting frequency, and temporal patterns) with deep learning models to improve detection performance for minority categories. Furthermore, the focus of future work will shift toward the practical deployment of these algorithms into campus warning systems, with a rigorous emphasis on data privacy protection and ethical safeguards.

5. Conclusion

In this study, a Logistic Regression framework was implemented to conduct multi-dimensional mental profiling across a corpus of over 52,000 social media texts. The experimental results demonstrate that linguistic features extracted via TF-IDF effectively distinguish the mental health status of adolescents. Despite the challenges posed by class imbalance in categories such as "Stress" and "Personality Disorders," the overall Accuracy of 74.8%—coupled with robust performance in identifying suicidal ideation—validates the technical viability of this approach for early-stage campus crisis screening.

Analysis of linguistic patterns reveals that different psychological conditions leave unique "digital traces" in texts: the depression group is characterized by high self-referentiality, while the anxiety group frequently displays uncertainty-related help-seeking expressions. The identification of these markers not only reveals the patterns of adolescent emotional expression in digital spaces but also provides a data-driven foundation for constructing low-cost, automated psychological early warning systems. Moving forward, with the integration of pre-trained models such as BERT and the inclusion of cross-cultural samples [14], this recognition framework is expected to further refine its diagnostic precision, ultimately securing a more effective "golden window" for adolescent mental health intervention.

References

[1] World Health Organization. (2024). Adolescent mental health. Fact Sheets. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

DOI: 10.54254/2755-2721/2026.BA32965

[2] Xiong, Z., Yu, J. Y., Zhuo, Y., et al. (2025). The impact of social media on adolescent mental health. Journal of Clinical Psychosomatic Diseases, 52(6): 1706-1708, 1729. DOI: 10.13479/j.cnki.jip.2025.06.004
[3] Duan, M. Y., Luo, C. L., & Deng, Y. F. (2025). Research progress on the influence of social media on non-suicidal self-injury behavior among adolescents. Journal of Clinical Psychosomatic Diseases, 52(6): 1694-1696, 1700. DOI: 10.13479/j.cnki.jip.2025.06.007
[4] Bhatt, K., Singh, K. A., Pandey, P., et al. (2026). Cross-platform social media analysis for mental health detection. Discover Mental Health. DOI: 10.1007/s44192-026-00368-w
[5] Murray, A., Zhu, X., Xiao, Z., et al. (2025). Emotions following social media use and their relations to mental health in young people: An ecological momentary assessment study. Journal of Affective Disorders, 397: 121015. DOI: 10.1016/j.jad.2025.121015
[6] Zhou, Y. Y., & Wang, S. Y. (2026). Application and challenges of artificial intelligence in assessing suicide risk in adolescents. China Science and Technology Information, 38(01): 39–41.
[7] Xu, C. Y., & Li, Y. Z. (2025). The impact of social media on mental health of college students. Journal of Social Media, (24): 49-51.
[8] Du, J. P., Miao, L. H., & Yang, V. X. (2025). The long-term effects of social media dependence on mental health of college students and intervention strategies. Media Forum, (17): 63–65.
[9] Zhou, Z. X. (2025). The impact of new media environment on college students' mental health and countermeasures. Digital Communication, (09): 75–77.
[10] Kaggle. (2024). Mental Health Corpus: A dataset for multi-class classification of social media text. [Online]. Available: https: //www.kaggle.com/datasets
[11] Chan, S. S., Solt, V. M., Milne, R. G., et al. (2026). Here and now: The effects of mindfulness on habitual social media usage. Journal of Loss and Trauma, 31(2): 342–361. DOI: 10.1080/15325024.2025.2560467
[12] Riches, S., Williams, G., Moss, H., et al. (2026). Social media-based participant recruitment in mental health research: An exploratory pilot study. Mental Health Review Journal, 31(1): 1–15. DOI: 10.1108/MHRJ-03-2024-0018
[13] Cai, F., Wang, Y., & Jin, S. (2026). The impact of social media addiction on college students' mental health through social support and resilience. Scientific Reports. DOI: 10.1038/s41598-026-35779-w
[14] Wang, L., Xu, Y., Shan, Q., et al. (2025). Impact of social media on mental health and body image dissatisfaction among Saudi women. African Journal of Reproductive Health, 29(12s): 55–64. DOI: 10.29063/ajrh.v29i12s.7