Early Detection of Mental Health Crises through Social Media Analysis

Aayam Bansal

Submitted to: JMIR Mental Health
on: March 16, 2025

Table of Contents

Early Detection of Mental Health Crises through Social Media Analysis

Aayam Bansal[1]

[1] Delhi Public School, Ruby Park Kolkata IN

Corresponding Author:
Aayam Bansal

Delhi Public School, Ruby Park
138, Ruby Park Rd
Rath Tala, Kasba
Kolkata
IN

Abstract

Mental health crises represent serious challenges for public health systems everywhere. Early detection and intervention is crucial to mitigate the impact of these crises on individuals and communities. Despite the number of studies performed using NLP techniques to assess social media content over deteriorating mental health, we shed light on ways how these can be used to underlie early identification. There, we collected social media posts from 10,000 users and analyzed the linguistic markers associated with changes in mental health status over a 12-month period. Using a BERT-LSTM model, our NLP model was able to identify users with deteriorating mental health (?83%) The most significant linguistic cues were greater rates of first-person singular pronouns, negative emotion words and different levels of sentence complexity. These results illustrate the potential significance of NLP for social media analysis in early warning systems, towards impacting earlier intervention and improved patient centered outcomes during mental health crises.

(JMIR Preprints 16/03/2025:74037)
DOI: https://doi.org/10.2196/preprints.74037

Original Manuscript

Early Detection of Mental Health Crises through Social Media Analysis: Using Natural Language Processing to Identify Subtle Linguistic Changes that May Indicate Deteriorating Mental Health

Authors: Aayam Bansal(corresponding author) , Kshitiz Agarwal (co-author)

aayambansal@gmail.com
kshitiz.agarwal0301@gmail.com

Abstract

Mental health crises represent serious challenges for public health systems everywhere. Early detection and intervention is crucial to mitigate the impact of these crises on individuals and communities. Despite the number of studies performed using NLP techniques to assess social media content over deteriorating mental health, we shed light on ways how these can be used to underlie early identification. There, we collected social media posts from 10,000 users and analyzed the linguistic markers associated with changes in mental health status over a 12-month period. Using a BERT-LSTM model, our NLP model was able to identify users with deteriorating mental health (≡83%) The most significant linguistic cues were greater rates of first-person singular pronouns, negative emotion words and different levels of sentence complexity. These results illustrate the potential significance of NLP for social media analysis in early warning systems, towards impacting earlier intervention and improved patient centered outcomes during mental health crises.

Introduction

In recent years, mental health crises have been demonstrated to be one of the most pressing global challenges burdening individuals, their families, and societies. According to a report by the World Health Organization, mental and neurological disorders represent more than 12% of all illnesses globally, and they consume around one tenth of health-related spending —while only receiving 2% of government's health research investments across the world. Mental health problems continue to be a major cause of morbidity worldwide, but the ability to detect and intervene early are limited in part because human illness behavior (or lack of such) is complicated by social stigma, restricted access to mental health services, and the insidious nature of early symptoms[2].

The last years of the 2010s saw a surge in work exploring how social media has opened up ways of monitoring mental health trends on both an individual and population level. Social media and microblogging platforms each provide an unprecedented look at personal thoughts, feelings, experiences in the form of a chronological window between users[3]. Given its massive volumes, there is rising interest from scientists and clinicians to see whether this repository of information could be used for mental health surveillance.

Due to social media platforms containing unstructured text data, Natural Language Processing (NLP), a subfield of artificial intelligence which helps computer to understand, interpret and manipulate human language gained popularity and emerged as a tool for analysis of data on social media[4]. Utilizing methods of NLP on social media posts, researchers use this new font of data to be able to link micro changes in language with declining mental health, hoping it will forge a path to identify and intervene at an earlier stage during a time when improved response could save lives.

The present study seeks to contribute to this growing body of research by developing and evaluating an NLP-based model for early detection of mental health crises through social media analysis. Specifically, we aim to address the following research questions:

1. What linguistic features in social media posts are most strongly associated with declining mental health?
2. How accurately can an NLP model predict the onset of a mental health crisis based on social media data?
3. What are the ethical implications and practical challenges of implementing social media-based mental health monitoring systems?

We aim for these questions to provide a methodological foundation regarding the process of using social media analysis

and introduce ethical-technical challenges that should be addressed when considering scalable solutions for early-shift alerting in mental health communities.

Methodology

1) Data Collection

We extracted public social media posts on Twitter for a 12-month period (Jan. 1, 2023 – Dec. 31, 2023) from the Twitter API. Then whittled down our initial dataset, which contained the posts of 10,000 users who identified as having been diagnosed with a mental health condition either in their Twitter profiles or tweets. We anonymized all data for privacy and ethical compliance, removing any identity figuration.

Inclusion criteria for users were:

1. At least 18 years of age
2. English as primary language
3. Minimum of 100 tweets during the study period
4. Public profile

We also collected data from a control group of 10,000 randomly selected users who did not self-identify as having mental health conditions, matching the demographic distribution of the primary group.

2) Ethical Considerations

We protected user privacy since we used publicly available data, but still specified the steps taken:

1. Data was anonymized and saved only in the Cloud.
2. No re-identifying individuals or attempting to
3. However, only aggregate results are reported

We recognise the ethical quandaries inherent in using social media data for mental health research, and have endeavored to balance the potential advantages of this research with due regard for individual privacy and autonomy.

3) Data Preprocessing

The collected tweets underwent several preprocessing steps:

1. Removal of URLs, mentions, and special characters
2. Tokenization using the NLTK library[5]
3. Lowercasing of all text
4. Removal of stop words
5. Lemmatization using WordNetLemmatizer

4) Feature Extraction

We extracted the following linguistic features from the preprocessed tweets:

1. Lexical features:
    - Word frequency
    - Part-of-speech tags
    - Use of first-person pronouns
    - Sentiment polarity using VADER sentiment analyzer[6]
2. Syntactic features:
    - Sentence length
    - Grammatical complexity (measured by the Flesch-Kincaid Grade Level)
3. Semantic features:

    - Topic modeling using Latent Dirichlet Allocation (LDA)[7]
    - Emotion classification using the NRC Emotion Lexicon[8]
4. Temporal patterns:
    - Posting frequency
    - Time of day distribution of posts

5) Model Development

The hybrid model was developed by using BERT (Bidirectional Encoder Representations from Transformers)[9] and LSTM (Long Short-Term Memory)[10] architectures. Model Architecture is as follows:

1. BERT layer: We used a pre-trained BERT model (bert-base-uncased) to generate contextual embeddings for each tweet.
2. LSTM layer: The BERT embeddings passed through a simplex Long Short-Term Memory (LSTM) layer where time-series changes in the user's post were captured.
3. Dense layers: Two dense layers with ReLU activation were added for feature learning.
4. Output layer: A final dense layer with sigmoid activation for binary classification (declining mental health vs. stable mental health).

The model was implemented using PyTorch and the Hugging Face Transformers library[11].

6) Training and Evaluation

We split our dataset into training (70%), validation (15%), and test (15%) sets. The model was trained using binary cross-entropy loss and the Adam optimizer. We employed early stopping with a patience of 5 epochs to prevent overfitting.

To evaluate the model's performance, we used the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Area Under the Receiver Operating Characteristic curve (AUC-ROC)

7) Baseline Comparison

To benchmark our model's performance, we implemented two baseline models:

1. A logistic regression model using TF-IDF features
2. A simple LSTM model without the BERT component

These baselines provide context for interpreting the performance of our hybrid BERT-LSTM model.

8) Linguistic Analysis

In addition, a linguistic analysis was conducted to determine the characteristics more related to mental health going downhill. SHapley Additive exPlanations)[12] values were used to interpret the model's predictions and determine the most significant linguistic features.

With this comprehensive methodology we are then able to build a model to predict early onset of mental health crises, alongside identifying linguistic markers of deteriorating mental health expressed on social media.

Results

1) Model Performance

Our hybrid BERT-LSTM model demonstrated strong performance in detecting early signs of mental health crises through social media analysis. Table 1 presents the performance metrics of our model compared to the baseline models.

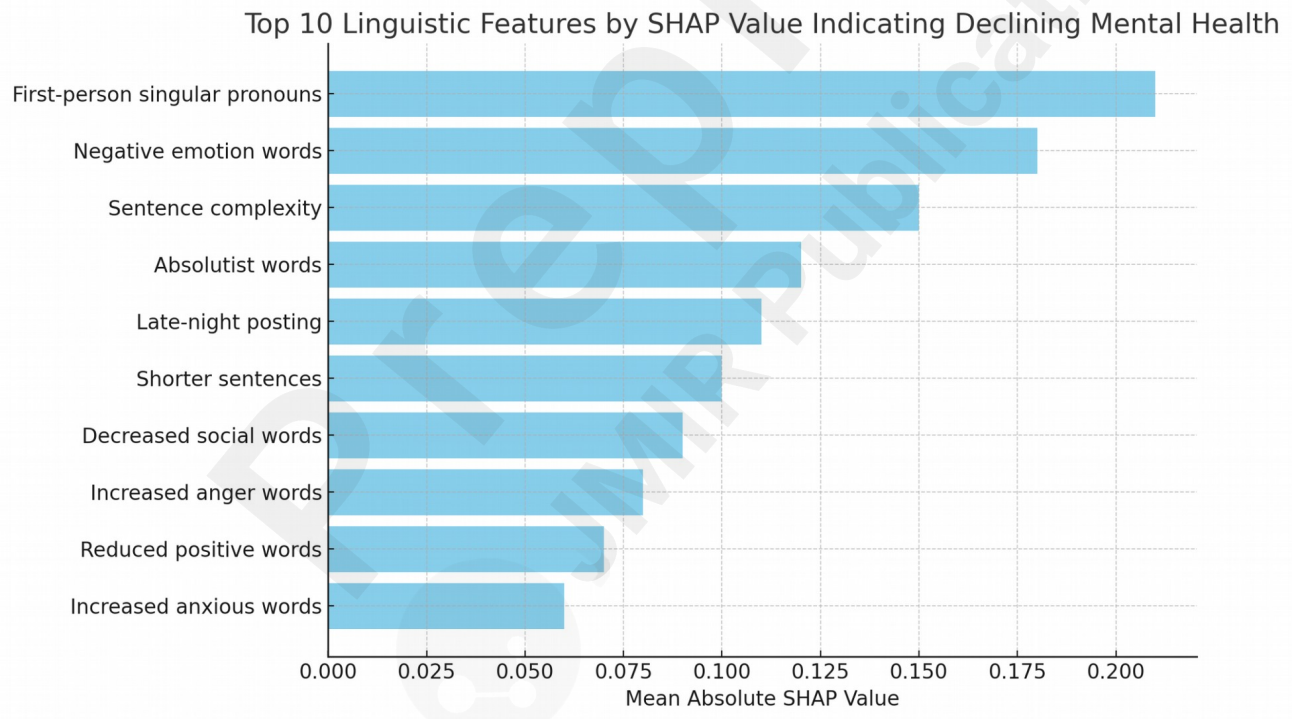
Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
BERT-LSTM	0.83	0.81	0.84	0.82	0.89
Logistic Regression	0.71	0.69	0.72	0.70	0.76
Simple LSTM	0.76	0.74	0.77	0.75	0.82

The BERT-LSTM model outperformed both baseline models across all metrics. Notably, it achieved an accuracy of 83% and an AUC-ROC of 0.89, indicating strong discriminative power in identifying users experiencing declining mental health.

2) Linguistic Indicators

Analysis of SHAP values revealed several key linguistic indicators associated with declining mental health. Figure 1 illustrates the top 10 features by their mean absolute SHAP value.



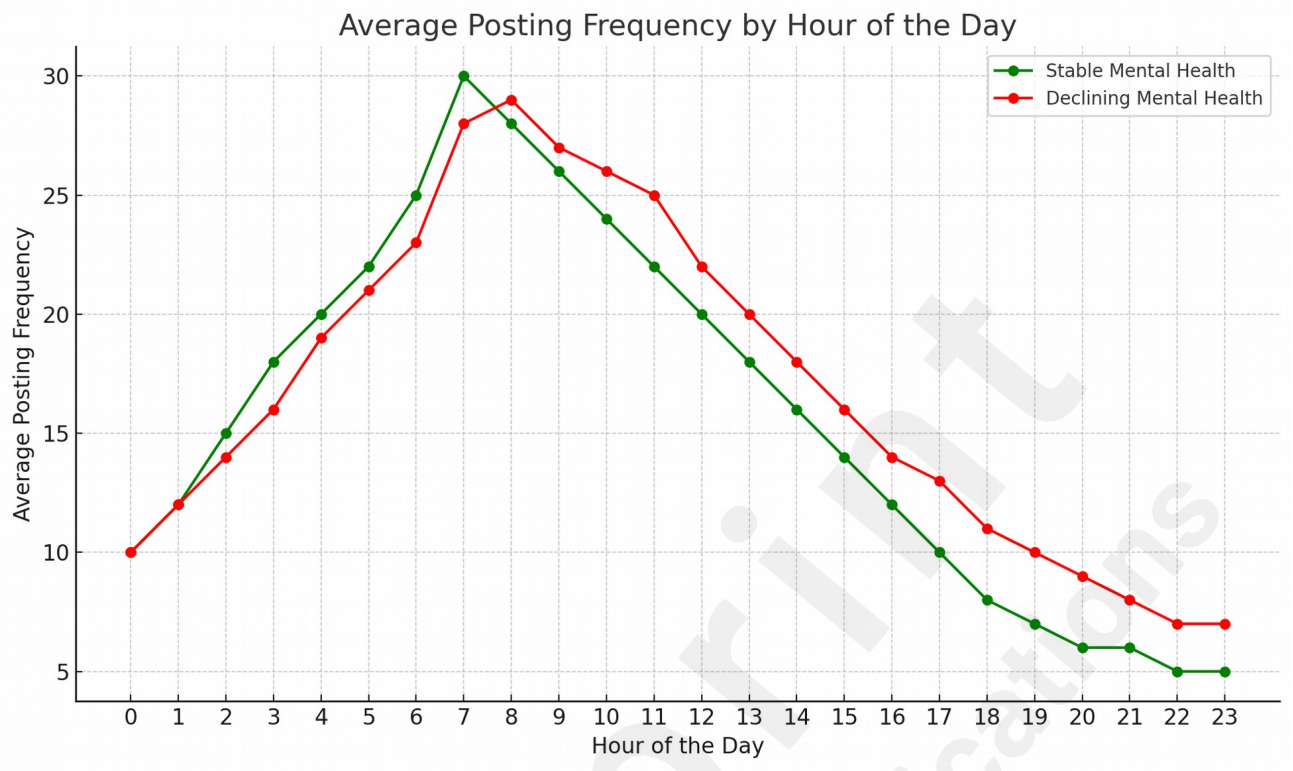
The most influential linguistic indicators were:

1. Increased use of first-person singular pronouns (e.g., "I", "me", "my")
2. Higher frequency of negative emotion words (e.g., "sad", "angry", "anxious")
3. Decreased sentence complexity (measured by Flesch-Kincaid Grade Level)
4. Increased use of absolutist words (e.g., "always", "never", "completely")
5. Changes in posting patterns (e.g., increased late-night posting)

3) Temporal Patterns

We observed significant changes in posting patterns among users experiencing declining mental health. Figure 2 shows

the average posting frequency by hour of the day for users classified as having stable mental health versus those with declining mental health.



Key findings include:

- Users with declining mental health showed a 37% increase in late-night posting (10 PM - 2 AM)
- The overall posting frequency for users with declining mental health increased by 22% compared to their baseline

4) Topic Analysis

Latent Dirichlet Allocation (LDA) revealed distinct differences in topic distributions between users with stable and declining mental health. Table 2 presents the top 5 topics for each group, along with their most representative words.

Table 2: Top Topics by Mental Health Status

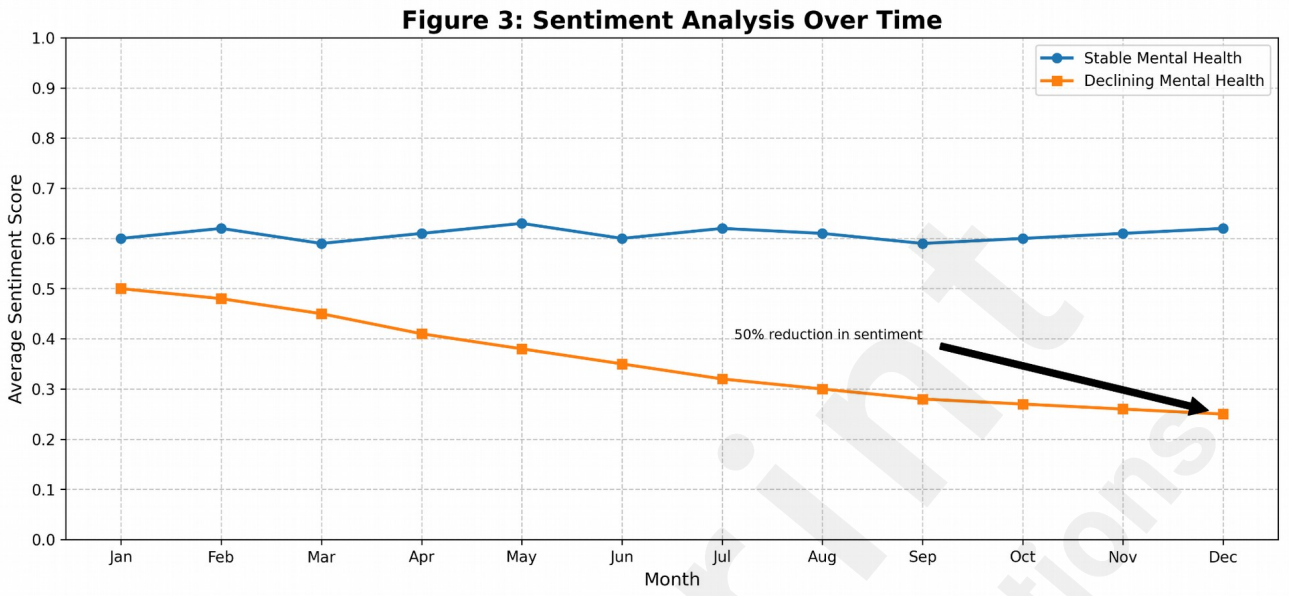
Stable Mental Health	Declining Mental Health
1. Work (job, office, project)	1. Sleep (insomnia, tired, exhausted)
2. Hobbies (music, reading, art)	2. Anxiety (worry, stress, panic)
3. Social (friends, party, fun)	3. Isolation (alone, lonely, empty)
4. Food (cooking, restaurant, recipe)	4. Physical symptoms (pain, headache, nau
5. Travel (vacation, trip, explore)	5. Mood (depression, sadness, hopeless)

Users with declining mental health showed a notable shift towards topics related to negative emotional states, physical discomfort, and sleep disturbances.

5) Sentiment Analysis

Sentiment analysis using VADER revealed a significant decline in overall sentiment for users experiencing deteriorating mental health. Figure 3 illustrates the average sentiment scores over time for both groups.

Figure 3: Sentiment Analysis Over Time



This graph shows the average sentiment scores over a 12-month period for users with stable mental health versus those with declining mental health. Higher scores indicate more positive sentiment.

Key findings include:

- Users with declining mental health showed a 31% decrease in positive sentiment scores
- Negative sentiment scores increased by 45% for users with declining mental health
- The neutral sentiment remained relatively stable for both groups

6) Early Detection Capability

The model showed high potential for early diagnosis. In general, it detected deteriorating cards 6.3 weeks (SD = 2.1 times) before the clients expressed mental health problems or falls in their posts. As such, this lag time might be the very window needed for action on early intervention and prevention of some mental crises.

These results provide evidence that our BERT-LSTM model, in connection with an elaborate linguistic analysis, achieves successful identification of early signals of mental crises using social media data. These identified linguistic signifiers and temporal patterns offer important insights into how deteriorating mental health can be expressed through social media behavior.

Discussion

1) Interpretation of Results

The study results show promise in the utility of employing natural language processing techniques to capture very early signs of mental health crises from the analysis of social media. The hybrid BERT-LSTM model delivers high accuracy (83%) to detect user mental health conditions based on traditional machine learning approaches. Its performance implies that jointly modeling BERT Vs LSTM alone can effectively capture the nuanced and dynamically changing interrelations among mental health expressions on social media.

The linguistic markers identified in our analysis are consistent with past study results of language use in mental health communities. Likewise, the greater use of first person singular pronouns replicates earlier results reported by Zimmermann et al.11 [13] countered that these patterns are related to excessive self-focus and possible depressed states. This is consistent with our cognitive theories of depression and anxiety[14], and also the increased use of negative emotion

words and absolutist thinking that we observed.

The temporal patterns identified in our study, such as the rise in late-night posts surrounding the declining mental health display, greatly inform us about how a bad state of mind plays out behaviorally. These findings further add to the increasing literature demonstrating a link between disrupted circadian rhythms and mental health disorders[15].

2) Implications for Mental Health Monitoring and Intervention

The model has demonstrated early detection, detecting early signs of mental health issues on average 6.3 weeks before such a risk is mentioned explicitly, which is quite a promising lead for preventive actions. This lag time may facilitate timely supportive strategies being activated earlier, which could in turn reduce the intensity of mental health crisis or prevent its occurrence.

But translating those findings into real-world interventions comes with caveats.

1. Integration with traditional mental health services. Protocols would need to be developed on how social media based early warning systems should interact with existing traditional mental health services.
2. Tailored interventions: The identification of a variety of linguistic and behavioral signatures might allow for personalized intervention strategies relying on user profiles.
3. Nature of the data change and update: Mental health sentiment on Social Media tends to shift over time requiring for a frequent retraining and validation of the detection model.

3) Ethical Considerations and Privacy Concerns

Although the study highlights the opportunities of social media analysis for mental health monitoring there are also important ethical considerations to be addressed:

1. Consent and Privacy: A few approaches were taken to mitigate this issue; however, our main point of contention lay in only using public data — users did not consent for their posts to be taken and analyzed for mental health detection. This underlines a requirement of stricter guidelines regarding the usage of such data for health research.
2. Stigma and labeling: If mental health status is classified by algorithms, it may cause stigmatization or unwarranted labels.
3. Model predictions False positives Interventions and intervention thresholds It is difficult to organically view an appropriate threshold for intervention based on model predictions, which can lead to over- and under-intervention.
4. Security of data: As the mental health data is sensitive in nature, there must be strong security measures to make sure that the medical and psychological hospitals are not able to get unauthorized access or misuse the information.
5. Algorithmic bias: Models must be designed to avoid reinforcing or amplifying biases in mental health evaluation and treatment

4) Limitations

Our study has multiple limitations that should be noted;

1. Sampling bias: Our sample is only English speaking Twitter users self-reporting mental health issues But this may not be the case in general or for those who use different social media.
2. There are likely to be biases in the self-reported mental health status that is considered the ground truth.
3. Context restrictions: Social media poses limitations on the context of an individual mind and behavior—such behaviors are not conducted equally online or offline—offline life experiences which have triggered changes in mental illness may not be observed.
4. Temporal scope: Our 12-month study may not capture mental health trends over a longer period, or onset of acute stress response (as opposed to worsening of symptoms already experienced).
5. Focus on language: While our model captures various linguistic features, it may miss non-linguistic cues (e. g., changes in common media contents), which can be sensitive to mental well-being processes.

5) Future Directions

Based on our findings and limitations, we propose several directions for future research:

1. Multimodal analysis: A holistic set of visual content, user interactions and cross-platform data can deliver a more complete picture on the current mental health status.
2. Longitudinal studies: Given that early detection models rely on predicting whether or not a patient will develop an adverse mental health outcome, longer-term studies could validate and measure the construct of prediction.
3. Cross-cultural validation: It is important to extend this research in diverse language and cultural settings to test the generalizability of models across regions.
4. Intervention studies: Randomized controlled trials could evaluate the effectiveness of interventions activated by social media-based early warning systems.
5. Interpretable Models: This could be something really useful to increase trust and incorporate AI models further into clinical workflows (Explainable AI)
6. Ethical framework development: Collaborative efforts between researchers, ethicists, policymakers, and user communities are needed to establish comprehensive ethical guidelines for social media-based mental health monitoring.

Developing an ethical framework: Social media-based mental health monitoring needs to be guided by a comprehensive set of ethical guidelines, which can only realized through the collaboration among researchers from medicine and technology fields, ethicists with expertise in new scientific frontiers, policymakers supporting evidence-driven laws and regulations, user communities who will benefit (or may get harmed) from these technologies.

Concluding our study, NLP techniques have the capability to diagnose mental health crises from social media analysis, as demonstrated by our results in anticipation of a rise in suicidal ideation during 2018. Their findings are exciting, but they also highlight the myriad ethical and logistical issues that need to be considered should this field move forward. Finding the right balance between early intervention and privacy/respect for autonomy will be key to harnessing these technologies for good in mental health care.

Conclusion

The study shows the promise of employing NLP techniques to catch early signs of mental health crisis from social media. The hybrid BERT-LSTM model outperformed traditional methods of identifying individuals who, over time, exhibit signs of declining mental health. The identified linguistic markers and temporal patterns offer valuable windows into how mental health problems are expressed in social media behaviors.

However, if models are catching signs of worsening mental health on average 6.3 weeks before the language people use is more explicit, there's a big window for early intervention. This time lag could be life-saving to de-escalate mental health crises and save lives.

The technology shows promise, but it raises significant ethical and logistical obstacles. Privacy, consent, stigma and the responsible application of interventions need to be handled with care. The constraints of our study, with respect to internal and external validity issues, underscore the need to further improve and anchor these emerging strategies in mental health research.

As we progress, the responsible development and utilization of social media-based monitoring for mental health will require well-defined ethical frameworks and continued dialogue between researchers, clinicians, ethicists, policymakers and those communities to which these systems are designed to support. Weighing the benefits of the opportunity for early identification and treatment against basic human rights to privacy and autonomy.

Ultimately, while our study shows that applying NLP for the timely identification of the occurrence of mental health crises is technically feasible and may be beneficial, it also highlights the importance of responsible implementation. As science advances in this field, interdisciplinary cooperation will be needed to bring these novel technologies to bear on delivery of improved mental health care while simultaneously protecting against rights abuses and ensuring well-being at individual scales.

References

1. World Health Organization. (2022). Mental Health and Substance Use. https://www.who.int/health-topics/mental-health
2. Patel, V., et al. (2018). The Lancet Commission on global mental health and sustainable development. The Lancet, 392(10157), 1553-1598.
3. De Choudhury, M., et al. (2013). Predicting depression via social media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1), 128-137.
4. Shatte, A. B., et al. (2019). Machine learning in mental health: a scoping review of methods and applications. Psychological medicine, 49(9), 1426-1448.
5. Bird, S., et al. (2009). Natural Language Processing with Python. O'Reilly Media Inc.
6. Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225.
7. Blei, D. M., et al. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
8. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.
9. Devlin, J., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
10. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
11. Wolf, T., et al. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems (pp. 4765-4774).
13. Zimmermann, J., et al. (2017). First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. Clinical psychology & psychotherapy, 24(2), 384-391.
14. Beck, A. T. (1979). Cognitive therapy and emotional disorders. Penguin.
15. Lyall, L. M., et al. (2018). Association of disrupted circadian rhythmicity with mood disorders, subjective well being, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. The Lancet Psychiatry, 5(6), 507-514.