Université d'Ottawa | University of Ottawa

An Exploration of Depression Detection using NLP

Paper: Depression Detection on Social Media with Large Language Models

Presented by: Sabrina Cai , Kelvin Mock, Jenifer Yu
2025-03-10

Agenda

Background

- Problem Statement
- Challenges in Traditional Approach
- Related Works
- Research Goals

Methodologies

- Outlining the Framework – DORIS
- Flow of DORIS

Modeling

- Model Design
- Post-Processing – Annotation
- Post-Processing – Mood Course
- Training and Predicting
- Explanability Model

Agenda

Experiment

- Setup
- Dataset
- Process

Evaluation

- Metrics & Results

Summary

- Conclusions
- Future Work

Research Background

Why is depression detection so important?

Problem Statement

Why is depression a major concern?

- WHO reports 5% of adults suffer from depression.
- Stigma and underdiagnosis are major barriers.

Why study social media for depression detection?

- Users express genuine emotions online.
- Potential for large-scale, low-cost monitoring.

Challenges in Traditional Approach

Limitations of Hospital-Based Diagnosis

- Expensive and time-consuming.
- Many individuals do not seek help.

Challenges in Depression Detection

- Requires professional medical knowledge.
- Needs high accuracy and explainability.

Limitations in Existing AI/ML Approaches

- Traditional classifiers lack medical interpretability.
- LLM-based methods are explainable but lack accuracy.
- Advent LLMs are sensitive to small prompt variations.

Research Goals

What this study aims to solve

- Automate depression detection
- Combine medical expertise with AI
- Maintain high Accuracy + Explainability

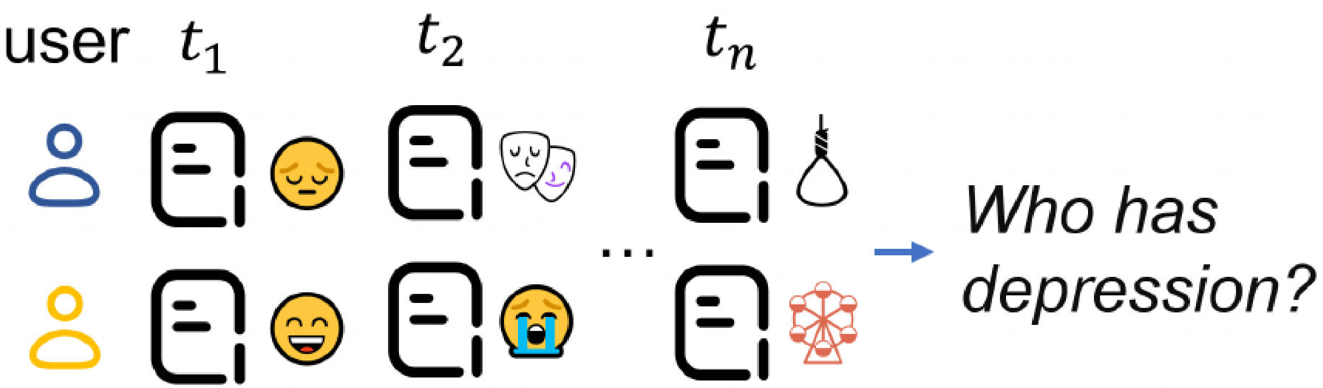


Figure 1: Illustration of depression detection on social media [1]

----

[1] Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024).

Related Works

- **Early Studies:** Sentiment analysis & keyword detection.
- **Feature Extraction**: LIWC, TF-IDF, LDA, etc.
- **Traditional ML**: SVM, Logistic Regression, etc.
- **Deep Learning:** CNN, RNN, PLM & BERT-based models.
- **LLMs in Mental Health:** Improved explainability (but lack accuracy).
- **Fine-tuning**: LLM + embedding models → accuracy

Proven Advantages

✓ Lower concealment potential → reliable diagnosis
✓ Lower cost 🤑
✓ Wider medical coverage
✓ LLM's language generalization ability → interpretability

Methodologies

How is NLP applied in this domain?

Outlining the Framework

- DORIS = **D**iagn**O**stic C**R**iteria-Guided Mood H**IS**tory-Aware
- Aim: To enhance detection accuracy using **DSM-5 criteria**.
- (a widely-used scale aligning with medical knowledge)

- History of a user's post on social media: $P = \{P_1, P_2, \dots, P_n\}$
- Corresponding timestamp: $t_1, t_2, \dots, t_n$
- Features:
  1. Depression Symptom
  2. Post History Representation
  3. Mood Course Representation
- Binary Classification Problem
- LLM + text embedding models

A. Depressed mood
B. Loss of interest/pleasure
C. Weight loss or gain
D. Insomnia or hypersomnia
E. Psychomotor agitation or retardation
F. Fatigue
G. Inappropriate guilt
H. Decreased concentration
I. Thoughts of suicide

Figure 2: Symptoms of depression in DSM-5 [1]

----

[1] Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024).

Flow of DORIS

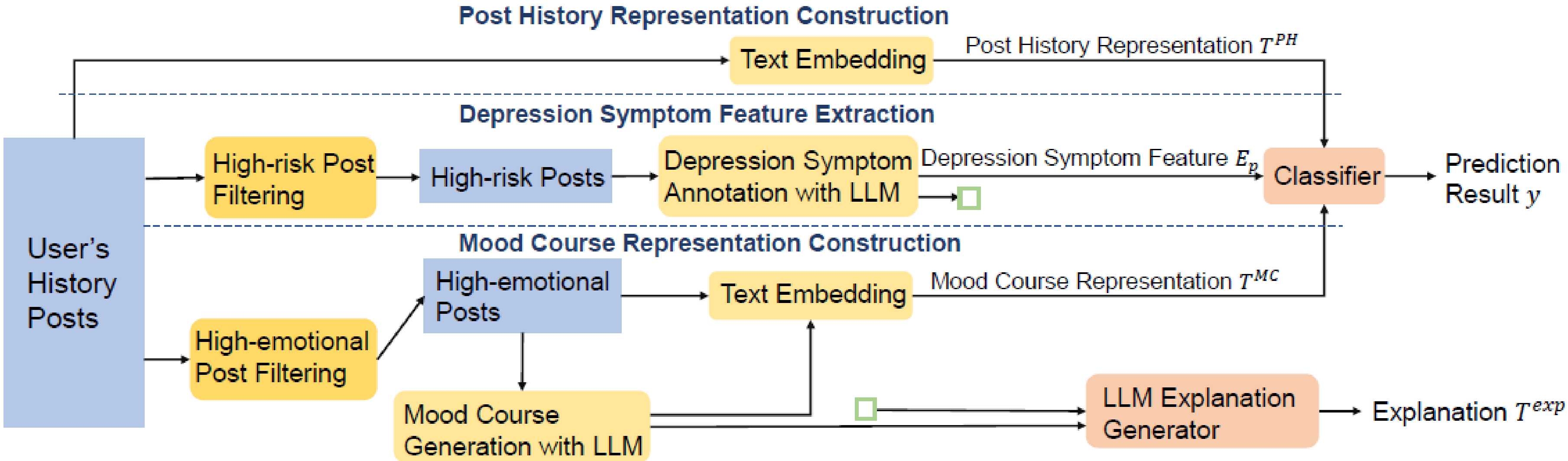


Figure 3: Illustration of DORIS Framework [1]

----

[1] Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024).

Modeling

How is DORIS working?
How can we design a model to achieve the goal?

Model Design

- Leverages LLM's semantic understanding strength
- Automates human annotations

Text → LLM → 9-d vector

- Feature Representation: 0 or 1
- Computation: Annotate only some selected **high-risk** texts
- Symptom Template: containing 1st person textual expressions

  I have lost interest, feel indifferent, bored, unconcerned…
- Text Embedding Model:
  - ❖ For Symptom Template: $H_i = Encoder\left(T_i^{DC}\right),\ \forall\, i\ \in \{A\ \ldots\ I\}$
  - ❖ For each post: $H_p = Encoder(p)$

Post-Processing: Annotation

- Average similarity: post vs symptom template
- Depression Risk Level: $Sim_p = mean\left(Sim\left(H_p, H_i\right)\right), \forall\, i \in \{A \dots I\}$
- Annotate selectively top k-% of the Sim_p score

- **Diagnostic Critical Feature**:
  - ❖ Averaging all symptom vectors
  - ❖ $F_u^{DC} = \frac{1}{N} \sum_{p=1}^{N} E_p$, N = total posts

A. Depressed mood
B. Loss of interest/pleasure
C. Weight loss or gain
D. Insomnia or hypersomnia
E. Psychomotor agitation or retardation
F. Fatigue
G. Inappropriate guilt
H. Decreased concentration
I. Thoughts of suicide

Figure 2: Symptoms of depression in DSM-5 [1]

----

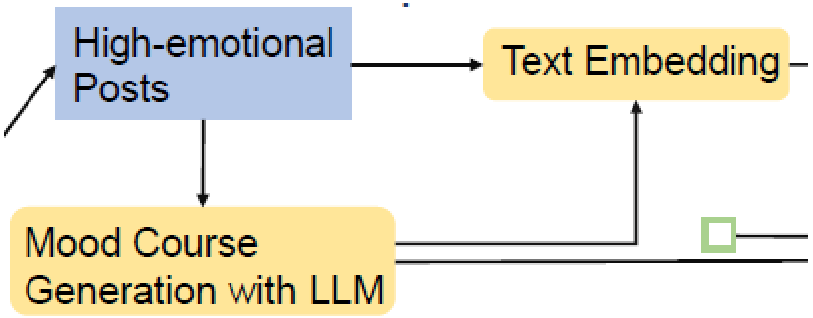
[1] Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024).

Post-Processing: Mood Course

- Temporal pattern and progression of emotional states
- Categorize Emotions (DSM-5):

Anger	Disgust	Anxiety	Happiness	Sadness

- Create a template/corpus of emotional expressions
  Sadness: I am sad, sorrowful, melancholic, in pain, lost, …
- Construct a Representation with pretrained embedding model
  - ❖ Per Post
  - ❖ Per Emotion Template



Post-Processing: Mood-Course

Pre-trained Embedding Model

- Each Emotion: **Retrain** posts within top m-% similarity
- **Post Filtering**: Form a union set of high emotional content
- **Label a user**'s mood course: Intersection of historical posts
  - ❖ Leveraging the strength of LLM (by prompting) = text
  - ❖ Averaging the embedding of each post from a user

Training and Predicting

Components

- Feature: $Concat(F^{MC} + F^{PH}, F^{DC})$
- Classifier: **Gradient Boosting** Ensemble (Decision Trees)
- Iteratively adds a **decision tree** → select an optimal split
- Minimizes a loss function with negative gradient
- Outputting a Prediction after *M* iterations – **sign** of ensemble

✓ Automatically performs feature interactions
✓ Effectively fuses the components in the resulting Feature

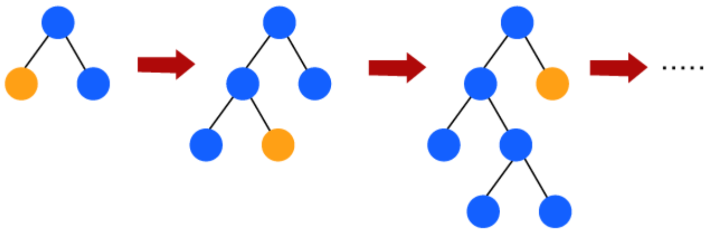


Figure 4: An Illustration of Gradient Boosting [2]

----

[2] Santhosh Sachin (March 2024)

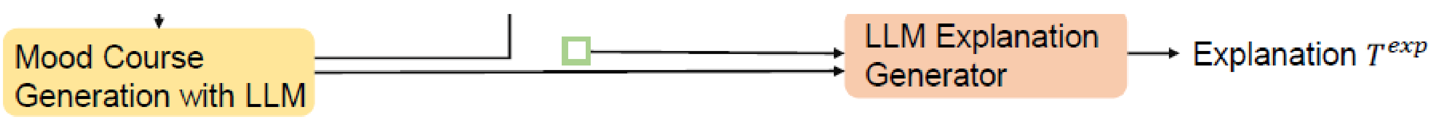
Explanability Model

Recent Works:

× Directly analyze raw texts with LLM → accuracy discrepancies

This Paper's Approach:

Traditional Classifier	Generate mood course descriptions → Stably Accurate
LLM	Annotate symptoms → Explainable



- Final Output = explanatory output + system's annotation + user's mood course descriptions + classification results

----

[1] Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024).

Explanability Model – Sample Prompt

"Assuming you are a psychiatrist specializing in depression. Here is a user's mood course: T^{MC}; below are posts from this user displaying symptoms of depression and the types of symptoms exhibited: ...; this user has been determined by an automated depression detection system to be depressed/normal. Please consider the user's mood course and posts to generate an explanation for this judgment."

Figure 5: An example of a prompt to LLM [1]

Experiment

How can we prove the proposed methodology valid?

Experiment Setup

Implementation Details:

- Low-resource embedding model: gte-small-zh
- LLM (GPT-3.5-Turbo) for annotation
- Gradient Boosting Trees (GBT) for classification

Dataset:

- SWDD: 1,000 depressed vs. 19,000 control users
- Realistic ratio simulating actual prevalence

Baselines & Metrics:

- Compared with traditional (TF-IDF + XGBoost), deep learning (HAN), and LLM-based methods
- Metrics: Precision, Recall, F1, AUROC, AUPRC

Overall Performance Result

Key Findings:

- DORIS outperforms all baselines on all metrics
- Improvement of 0.036 in AUPRC – significant for imbalanced data

Category	Method	Precision	Recall	F1-score	AUROC	AUPRC
Traditional Method	TF-IDF+XGBoost	0.3644	0.4300	0.3945	0.9023	0.4303
Deep Learning-Based Methods	HAN	0.5702	0.6500	0.6075	0.8929	0.5864
	Mood2Content	0.7216	0.7000	<u>0.7106</u>	<u>0.9537</u>	<u>0.7774</u>
PLM-Based Methods	FastText	<u>0.7467</u>	0.5600	0.6400	0.9441	0.6255
	gte-small	0.6359	0.6526	0.6200	0.9499	0.6959
	BERT	0.6667	0.6400	0.6531	0.9481	0.7102
	MentalRoBERTa	0.7326	0.6300	0.6774	0.9423	0.6880
LLM-Based Methods	ChatGPT	0.0875	0.7100	0.1559	0.6603	0.0767
	MentalLLama	0.0899	<u>0.7800</u>	0.1612	0.6821	0.0811
Our Method	DORIS	**0.7596**	**0.7900**	**0.7596**	**0.9715**	**0.8134**

Table 1: Performance of DORIS and baselines.
The best scores are in bold, and second best scores are underlined.

Ablation and Hyperparameter Studies

Ablation Study:

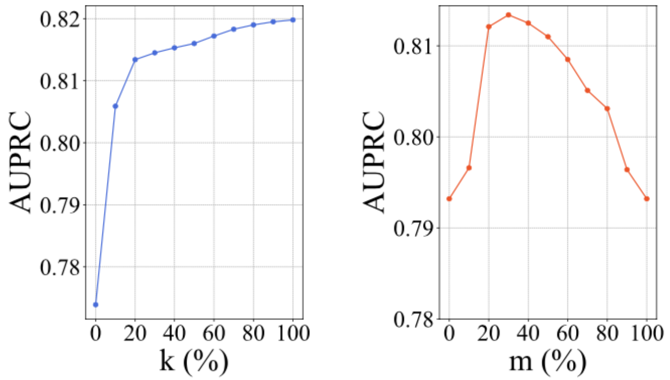
- Removing any component (Diagnostic, Mood Course, or Post History) decreases performance

	F1-score	AUROC	AUPRC
Full Design	0.7596	0.9715	0.8134
w/o DC Feature	0.6867	0.9679	0.7739
w/o MC Representation	0.7415	0.9660	0.7932
w/o PH Representation	0.7200	0.9660	0.7817

Table 2: DC denotes diagnostic criteria, MC denotes mood course, and PH denotes post history.

Hyperparameter Study:

- Parameter k: high-risk text filtering
- Parameter m: emotional intensity filtering
- Optimal settings balance filtering with information retention



(a) The impact of k. (b) The impact of m.

Figure 6: Results of hyperparameter study.

Case Study

Diagnostic Criteria Feature Construction

Post Time	Text	Symptoms
April 23rd, 2015	The incident from two years ago, I'll never forget it for the rest of my life.	None
April 30th, 2015	Driven to the brink of agitation, consumed by an overwhelming restlessness.	Psychomotor agitation
May 3rd, 2015	I'm going back to school tomorrow, and I haven't touched volleyball in a month.	None
May 8th, 2015	The bitterness in my heart, no one understands. Even if I cry until tears run dry.	Depressed mood
…	…	…

Mood Course Representation Construction

Post Time	Text with High Emotional Content
April 30th, 2015	Driven to the brink of agitation, consumed by an overwhelming restlessness.
May 8th, 2015	The bitterness in my heart, no one understands. Even if I cry until tears run dry.
…	…
July 16th, 2015	I really dislike people who lack integrity, and I find it very repulsive.
July 22th, 2015	I just want to find a place to hide and cry my heart out right now.

Mood Course Description

The individual's social media expressions from April to July 2015 depict a concerning trajectory of emotional distress. Initially, they report intense agitation and restlessness, suggesting acute emotional turmoil. As weeks progress, there's a deepening sense of isolation and misunderstood bitterness, indicating potential depressive symptoms. Episodes of wanting to withdraw socially and expressions of disdain towards perceived moral failings in others hint at both a heightened sensitivity to interpersonal dynamics and a profound sense of disillusionment. The culmination of these feelings into a desire for seclusion and emotional release by mid-July underscores a significant emotional struggle, possibly pointing towards depressive episodes or acute stress reactions. This mood course, characterized by fluctuations in emotional intensity and themes of alienation, distress, and a quest for solace, warrants professional attention for potential mental health interventions.

Generated Explanation

Based on the user's mood course and specific posts, the judgment of depression appears to be accurate. The reported symptoms—psychomotor agitation, a persistently depressed mood, unintentional weight loss, and a desire for isolation—are hallmark indicators of depression. These symptoms not only cover a range of depressive disorder criteria, including mood disturbances, physical changes, and social withdrawal, but also span several months, suggesting a sustained period of emotional distress rather than a transient phase. The pattern and intensity of symptoms described align with clinical understanding of depressive episodes, warranting a diagnosis of depression rather than a normal fluctuation of emotions.

Conclusion & Future Directions

What's Next for DORIS?

Conclusion & Future Directions

Conclusion:

- DORIS integrates DSM-5 medical criteria with NLP, improving accuracy & interpretability.
- Promising results, but needs further validation.
- Dependency on LLMs and Computational Cost
- AI Ethics Concerns: Privacy & Bias.

Future Work:

- Deploy in clinical settings for real-world validation.
- Expand to other mental health conditions: **Our project!**

References

- Lan, X., Cheng, Y., Sheng, L., Gao, C., & Li, Y. (2024). Depression detection on social media with large language models. Tsinghua University. Retrieved from https://arxiv.org/abs/2403.10750
- Santhosh S. (2024) Introduction to Gradient Boosting Machines (GBM): A Powerful Ensemble Technique. https://www.linkedin.com/pulse/introduction-gradient-boosting-machines-gbm-powerful-ensemble-sachin-fhdlc/
- Li, Zehan and Zhang, Xin and Zhang, Yanzhao and Long, Dingkun and Xie, Pengjun and Zhang, Meishan (2023). Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281. https://huggingface.co/thenlper/gte-small
- GPT-3.5-Turbo Legacy Model: https://platform.openai.com/docs/models/gpt-3-5-turbo

References

- Yang, Kailai and Zhang, Tianlin and Kuang, Ziyan and Xie, Qianqian and Ananiadou, Sophia (2023). MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. arXiv preprint arXiv:2309.13567. https://arxiv.org/pdf/2309.13567, https://github.com/SteveKGYang/MentalLLaMA
- Tianqi Chen, Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754v3 [cs.LG]. https://arxiv.org/pdf/1603.02754
- Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, Wei Gao (2023). Depression detection on online social network with multivariate time series feature of user depressive symptoms. Elsevier Ltd.

Thank you!

Any Questions?