Mental Health Crisis Detection on Reddit

Arya Vachhani
Maan Kumawat
Manya Agrawal

THE PROBLEM

According to WHO's 2025 report **OVER 1 BILLION** were living with a mental health condition

Almost every year the global economy suffers from a loss of of **$ 1 TRILLION**

Negative effects on, **QUALITY OF LIFE, RELATIONSHIP AND SOCIETAL PRODUCTIVITY**

Delay in early detection can increase cost of care by about **30%** in high cost treatment cases

PROBLEM STATEMENT

How might we leverage longitudinal social media data to detect subtle psychological shifts and predict mental health crises before they escalate?

LITERATURE

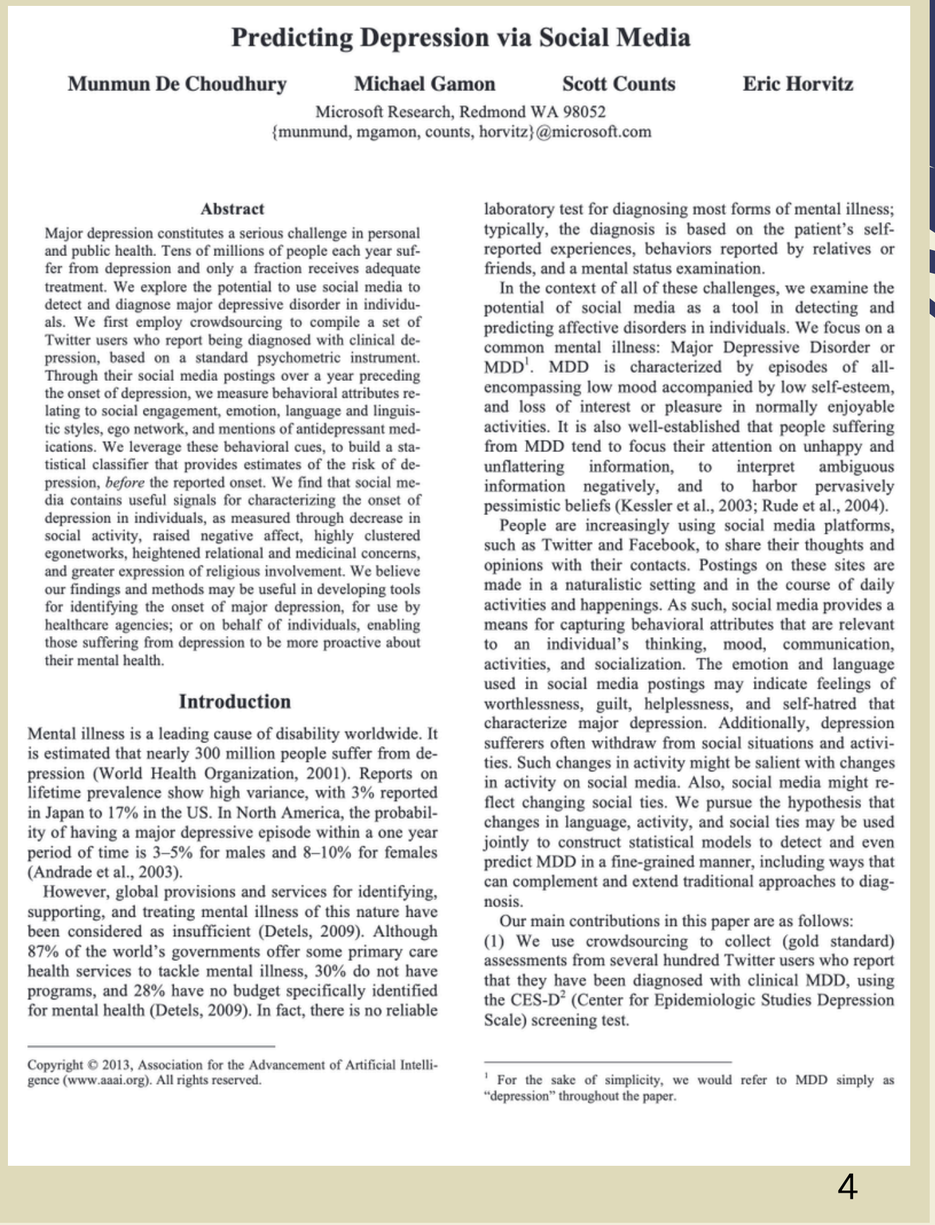
Foundational Work: De Choudhury et al. (2013)

Methodology: Built an SVM classifier using crowdsourced Twitter data screened via the CES-D scale to extract behavioral, emotional, and linguistic features

Performance metric(s) reported: Accuracy, Precision, Recall

Result: Achieved 70% accuracy and 0.74 precision in predicting the likelihood of depression ahead of its reported onset

Limitation: Reliance on retrospective self-reports, potential noise in crowdsourced data, and limited generalizability due to small sample sizes



Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

Abstract

Major depression constitutes a serious challenge in personal and public health. Tens of millions of people each year suffer from depression and only a fraction receives adequate treatment. We explore the potential to use social media to detect and diagnose major depressive disorder in individuals. We first employ crowdsourcing to compile a set of Twitter users who report being diagnosed with clinical depression, based on a standard psychometric instrument. Through their social media postings over a year preceding the onset of depression, we measure behavioral attributes relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. We leverage these behavioral cues, to build a statistical classifier that provides estimates of the risk of depression, *before* the reported onset. We find that social media contains useful signals for characterizing the onset of depression in individuals, as measured through decrease in social activity, raised negative affect, highly clustered egonetworks, heightened relational and medicinal concerns, and greater expression of religious involvement. We believe our findings and methods may be useful in developing tools for identifying the onset of major depression, for use by healthcare agencies; or on behalf of individuals, enabling those suffering from depression to be more proactive about their mental health.

Introduction

Mental illness is a leading cause of disability worldwide. It is estimated that nearly 300 million people suffer from depression (World Health Organization, 2001). Reports on lifetime prevalence show high variance, with 3% reported in Japan to 17% in the US. In North America, the probability of having a major depressive episode within a one year period of time is 3–5% for males and 8–10% for females (Andrade et al., 2003).

However, global provisions and services for identifying, supporting, and treating mental illness of this nature have been considered as insufficient (Detels, 2009). Although 87% of the world's governments offer some primary care health services to tackle mental illness, 30% do not have programs, and 28% have no budget specifically identified for mental health (Detels, 2009). In fact, there is no reliable laboratory test for diagnosing most forms of mental illness; typically, the diagnosis is based on the patient's self-reported experiences, behaviors reported by relatives or friends, and a mental status examination.

In the context of all of these challenges, we examine the potential of social media as a tool in detecting and predicting affective disorders in individuals. We focus on a common mental illness: Major Depressive Disorder or MDD[1]. MDD is characterized by episodes of all-encompassing low mood accompanied by low self-esteem, and loss of interest or pleasure in normally enjoyable activities. It is also well-established that people suffering from MDD tend to focus their attention on unhappy and unflattering information, to interpret ambiguous information negatively, and to harbor pervasively pessimistic beliefs (Kessler et al., 2003; Rude et al., 2004).

People are increasingly using social media platforms, such as Twitter and Facebook, to share their thoughts and opinions with their contacts. Postings on these sites are made in a naturalistic setting and in the course of daily activities and happenings. As such, social media provides a means for capturing behavioral attributes that are relevant to an individual's thinking, mood, communication, activities, and socialization. The emotion and language used in social media postings may indicate feelings of worthlessness, guilt, helplessness, and self-hatred that characterize major depression. Additionally, depression sufferers often withdraw from social situations and activities. Such changes in activity might be salient with changes in activity on social media. Also, social media might reflect changing social ties. We pursue the hypothesis that changes in language, activity, and social ties may be used jointly to construct statistical models to detect and even predict MDD in a fine-grained manner, including ways that can complement and extend traditional approaches to diagnosis.

Our main contributions in this paper are as follows:
(1) We use crowdsourcing to collect (gold standard) assessments from several hundred Twitter users who report that they have been diagnosed with clinical MDD, using the CES-D[2] (Center for Epidemiologic Studies Depression Scale) screening test.

[1] For the sake of simplicity, we would refer to MDD simply as "depression" throughout the paper.

LITERATURE

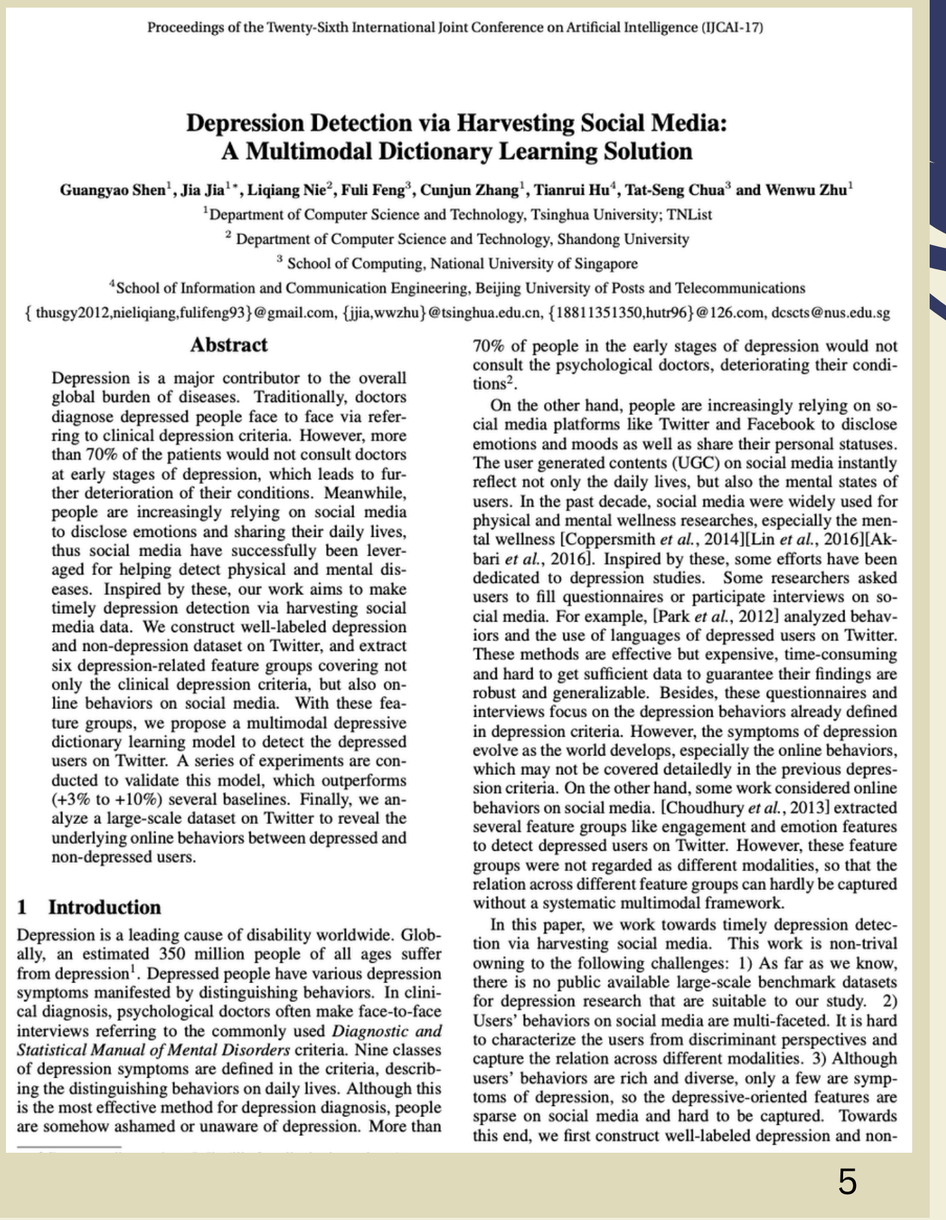
Shen et al. (2017)

Methodology: Integrated social behavior metrics with visual features from posts and utilized Linguistic Inquiry and Word Count (LIWC) and LDA topics for textual analysis. Incorporated domain-specific keywords to enrich the dataset

Performance metric(s) reported: Accuracy, Recall, Precision, F1-Measure

Result: Outperformed baseline models by 3-10%, achieving around an 85% F1-Measure in detecting depressive users

Limitation: The initial labeled dataset was small, and applying heuristic rule-based labeling to large datasets may introduce noise



Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution

Guangyao Shen[1], Jia Jia[1*], Liqiang Nie[2], Fuli Feng[3], Cunjun Zhang[1], Tianrui Hu[4], Tat-Seng Chua[3] and Wenwu Zhu[1]
[1]Department of Computer Science and Technology, Tsinghua University; TNList
[2] Department of Computer Science and Technology, Shandong University
[3] School of Computing, National University of Singapore
[4]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications
{ thusgy2012,nieliqiang,fulifeng93}@gmail.com, {jjia,wwzhu}@tsinghua.edu.cn, {18811351350,hutr96}@126.com, dcscts@nus.edu.sg

Abstract

Depression is a major contributor to the overall global burden of diseases. Traditionally, doctors diagnose depressed people face to face via referring to clinical depression criteria. However, more than 70% of the patients would not consult doctors at early stages of depression, which leads to further deterioration of their conditions. Meanwhile, people are increasingly relying on social media to disclose emotions and sharing their daily lives, thus social media have successfully been leveraged for helping detect physical and mental diseases. Inspired by these, our work aims to make timely depression detection via harvesting social media data. We construct well-labeled depression and non-depression dataset on Twitter, and extract six depression-related feature groups covering not only the clinical depression criteria, but also online behaviors on social media. With these feature groups, we propose a multimodal depressive dictionary learning model to detect the depressed users on Twitter. A series of experiments are conducted to validate this model, which outperforms (+3% to +10%) several baselines. Finally, we analyze a large-scale dataset on Twitter to reveal the underlying online behaviors between depressed and non-depressed users.

1 Introduction

Depression is a leading cause of disability worldwide. Globally, an estimated 350 million people of all ages suffer from depression[1]. Depressed people have various depression symptoms manifested by distinguishing behaviors. In clinical diagnosis, psychological doctors often make face-to-face interviews referring to the commonly used *Diagnostic and Statistical Manual of Mental Disorders* criteria. Nine classes of depression symptoms are defined in the criteria, describing the distinguishing behaviors on daily lives. Although this is the most effective method for depression diagnosis, people are somehow ashamed or unaware of depression. More than 70% of people in the early stages of depression would not consult the psychological doctors, deteriorating their conditions[2].

On the other hand, people are increasingly relying on social media platforms like Twitter and Facebook to disclose emotions and moods as well as share their personal statuses. The user generated contents (UGC) on social media instantly reflect not only the daily lives, but also the mental states of users. In the past decade, social media were widely used for physical and mental wellness researches, especially the mental wellness [Coppersmith *et al.*, 2014][Lin *et al.*, 2016][Akbari *et al.*, 2016]. Inspired by these, some efforts have been dedicated to depression studies. Some researchers asked users to fill questionnaires or participate interviews on social media. For example, [Park *et al.*, 2012] analyzed behaviors and the use of languages of depressed users on Twitter. These methods are effective but expensive, time-consuming and hard to get sufficient data to guarantee their findings are robust and generalizable. Besides, these questionnaires and interviews focus on the depression behaviors already defined in depression criteria. However, the symptoms of depression evolve as the world develops, especially the online behaviors, which may not be covered detailedly in the previous depression criteria. On the other hand, some work considered online behaviors on social media. [Choudhury *et al.*, 2013] extracted several feature groups like engagement and emotion features to detect depressed users on Twitter. However, these feature groups were not regarded as different modalities, so that the relation across different feature groups can hardly be captured without a systematic multimodal framework.

In this paper, we work towards timely depression detection via harvesting social media. This work is non-trival owning to the following challenges: 1) As far as we know, there is no public available large-scale benchmark datasets for depression research that are suitable to our study. 2) Users' behaviors on social media are multi-faceted. It is hard to characterize the users from discriminant perspectives and capture the relation across different modalities. 3) Although users' behaviors are rich and diverse, only a few are symptoms of depression, so the depressive-oriented features are sparse on social media and hard to be captured. Towards this end, we first construct well-labeled depression and non-

LITERATURE

Case Study: Mansoor & Ansari AI Model

Methodology: Developed a multimodal deep learning pipeline combining BERT, LSTM, and multi-head attention to analyze multilingual textual and visual social media data

Performance metric(s) reported: Accuracy, Precision, Recall, F1 Score, AUC-ROC

Result: 89% Accuracy with detection of 7.2 days prior to crisis.

Limitation: The study was limited by only 12 months of data and a notable rate of false positives due to sarcasm in social media text

Order Article Reprints

Open Access Article

Early Detection of Mental Health Crises through Artificial-Intelligence-Powered Social Media Analysis: A Prospective Observational Study

by Masab A. Mansoor [1,*] and Kashif H. Ansari [2]

[1] Louisiana Campus, Edward College of Osteopathic Medicine, Monroe, LA 71203, USA
[2] East Houston Medical Center, Houston, TX 77049, USA
* Author to whom correspondence should be addressed.

J. Pers. Med. **2024**, *14*(9), 958; https://doi.org/10.3390/jpm14090958

Submission received: 14 August 2024 / Revised: 31 August 2024 / Accepted: 5 September 2024 / Published: 9 September 2024

(This article belongs to the Special Issue Ehealth, Telemedicine, and AI in the Precision Medicine Era)

Download Browse Figures Versions Notes

Abstract

Background: The early detection of mental health crises is crucial for timely interventions and improved outcomes. This study explores the potential of artificial intelligence (AI) in analyzing social media data to identify early signs of mental health crises. Methods: We developed a multimodal deep learning model integrating natural language processing and temporal analysis techniques. The model was trained on a diverse dataset of 996,452 social media posts in multiple languages (English, Spanish, Mandarin, and Arabic) collected from Twitter, Reddit, and Facebook over 12 months. Its performance was evaluated using standard metrics and validated against expert psychiatric assessments. Results: The AI model demonstrated a high level of accuracy (89.3%) in detecting early signs of mental health crises, with an average lead time of 7.2 days before human expert identification. Performance was consistent across languages (F1 scores: 0.827–0.872) and platforms (F1 scores: 0.839–0.863). Key digital markers included linguistic patterns, behavioral changes, and temporal trends. The model showed varying levels of accuracy for different crisis types: depressive episodes (91.2%), manic episodes (88.7%), suicidal ideation (93.5%), and anxiety crises (87.3%). Conclusions: AI-powered analysis of social media data shows promise for the early detection of mental health crises across diverse linguistic and cultural contexts. However, ethical challenges, including privacy concerns, potential stigmatization, and cultural biases, need careful consideration. Future research should focus on longitudinal outcome studies, ethical integration of the method with existing mental health services, and developing personalized, culturally sensitive models.

Keywords: artificial intelligence; mental health; crisis detection; social media analysis; early intervention

1. Introduction

Mental health crises represent a significant global health challenge with far-reaching impacts on individuals, families, and communities [1]. Early detection and intervention are crucial in mitigating the severity and duration of these crises, yet traditional identification methods often fall short in providing timely support [2]. In recent years, the ubiquity of social media has created a novel opportunity for monitoring and analyzing behavioral patterns that may indicate emerging mental health concerns [3].

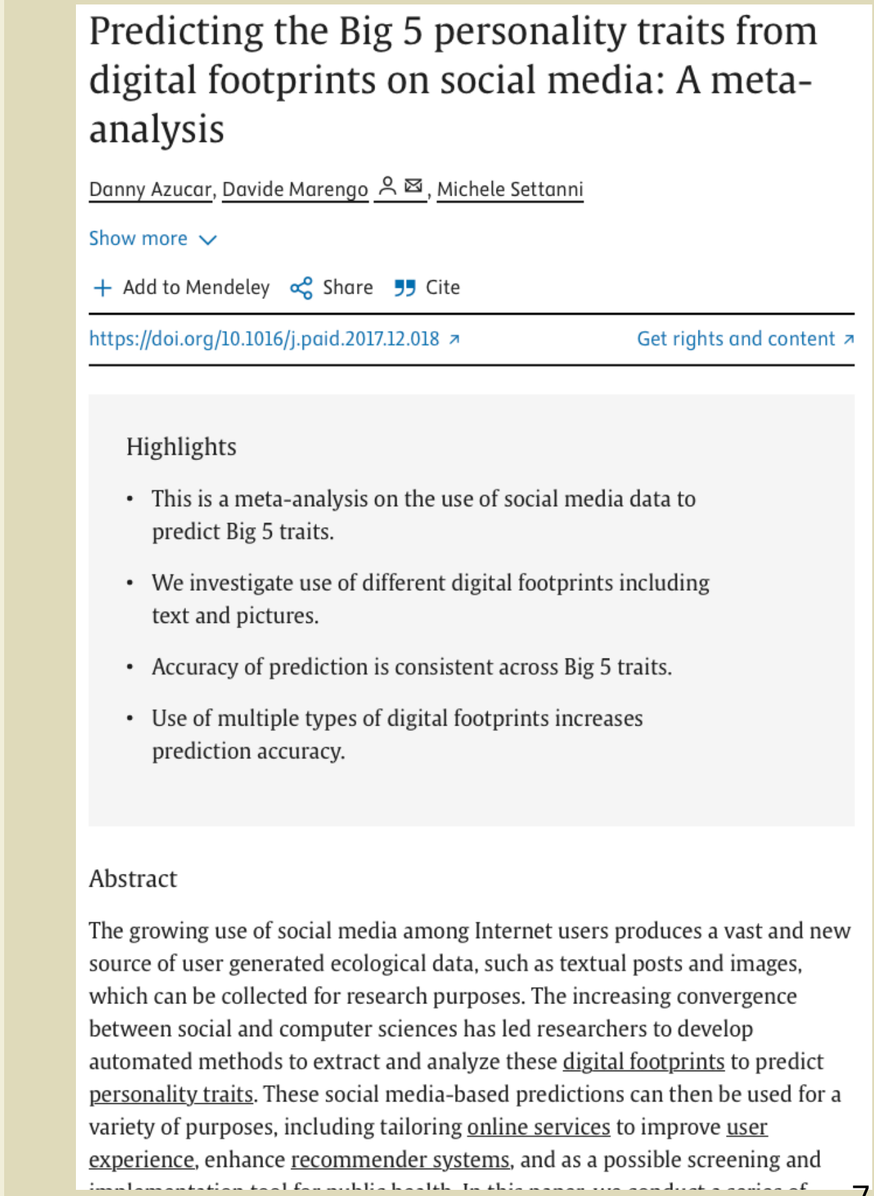
LITERATURE

Azucar, Marengo, and Settanni (2018)

Methodology: Conducted a meta-analysis combining multiple studies to determine the predictive power of social media digital footprints (text, likes, activity) on the Big 5 personality traits. Analysed 28 studies where 19 studies obtained their samples from Facebook, 5 from Twitter, 3 from the Sina Weibo micro-blogging site, and 1 article used a combined sample from Instagram and Twitter

Performance metric(s) reported : Pearson's r (correlation coefficient)

Result: Found significant predictive correlations matching real-world behavior, ranging from r=0.29 (Agreeableness) to r=0.40 (Extraversion)



Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis

Danny Azucar, Davide Marengo, Michele Settanni

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.paid.2017.12.018 Get rights and content

Highlights

- This is a meta-analysis on the use of social media data to predict Big 5 traits.
- We investigate use of different digital footprints including text and pictures.
- Accuracy of prediction is consistent across Big 5 traits.
- Use of multiple types of digital footprints increases prediction accuracy.

Abstract

The growing use of social media among Internet users produces a vast and new source of user generated ecological data, such as textual posts and images, which can be collected for research purposes. The increasing convergence between social and computer sciences has led researchers to develop automated methods to extract and analyze these digital footprints to predict personality traits. These social media-based predictions can then be used for a variety of purposes, including tailoring online services to improve user experience, enhance recommender systems, and as a possible screening and

LITERATURE

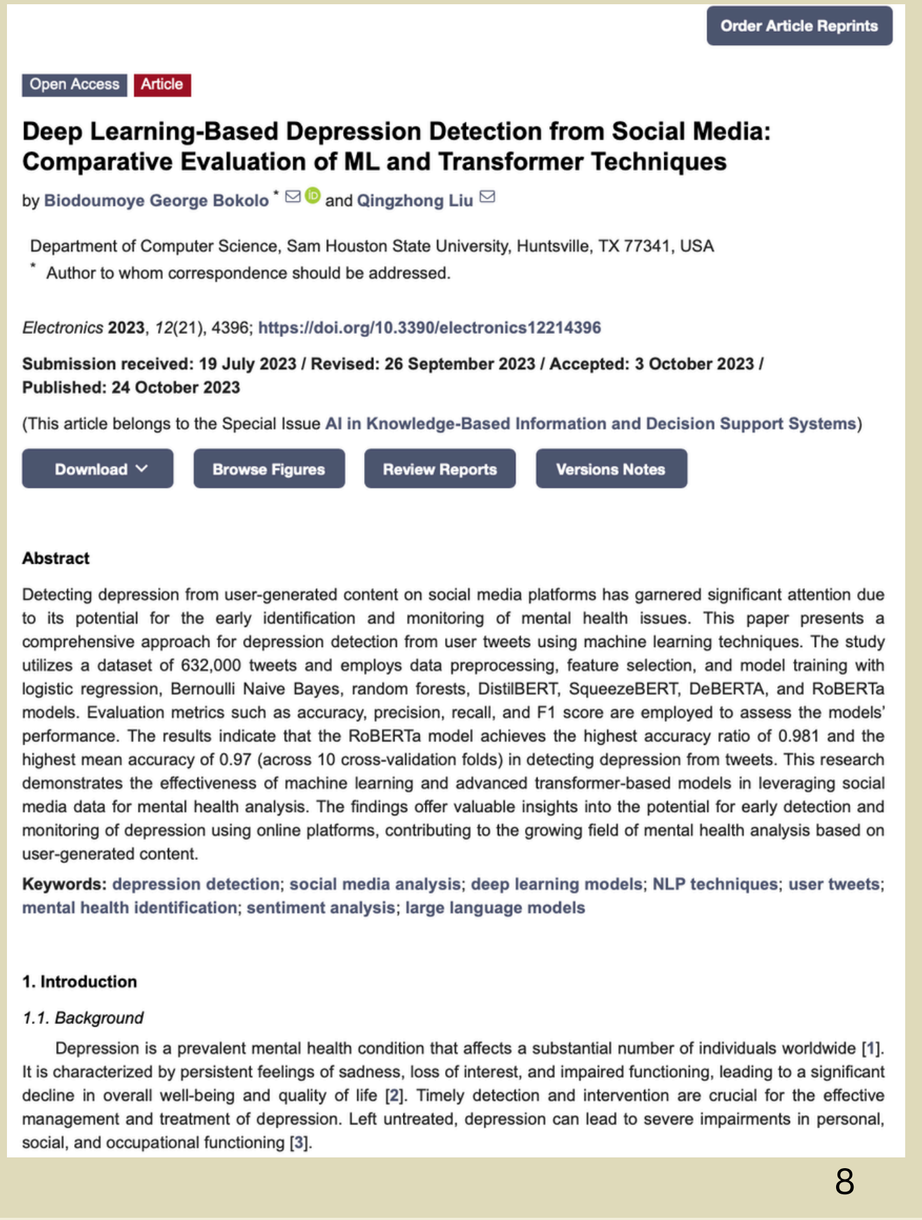
Bokolo and Liu (2023)

Methodology: The study utilises a dataset of 632,000 tweets and employs data preprocessing, feature selection, and model training with logistic regression, Bernoulli Naive Bayes, DistilBERT, SqueezeBERT, DeBERTA, and RoBERTa models

Performance metric(s) reported: Accuracy, Precision, Recall, F1 Score

Result: The fine-tuned RoBERTa model achieved the best performance with an outstanding 98.1% accuracy and 0.98 F1 Score

Limitation: The primary issue lies in the dataset's lack of clinically reliable labels and actual depression data, rendering such high accuracy scores misleading for practical mental health crisis prediction



Order Article Reprints

Open Access Article

Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques

by Biodoumoye George Bokolo * and Qingzhong Liu

Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA
* Author to whom correspondence should be addressed.

Electronics **2023**, *12*(21), 4396; https://doi.org/10.3390/electronics12214396

Submission received: 19 July 2023 / Revised: 26 September 2023 / Accepted: 3 October 2023 / Published: 24 October 2023

(This article belongs to the Special Issue AI in Knowledge-Based Information and Decision Support Systems)

Download Browse Figures Review Reports Versions Notes

Abstract

Detecting depression from user-generated content on social media platforms has garnered significant attention due to its potential for the early identification and monitoring of mental health issues. This paper presents a comprehensive approach for depression detection from user tweets using machine learning techniques. The study utilizes a dataset of 632,000 tweets and employs data preprocessing, feature selection, and model training with logistic regression, Bernoulli Naive Bayes, random forests, DistilBERT, SqueezeBERT, DeBERTA, and RoBERTa models. Evaluation metrics such as accuracy, precision, recall, and F1 score are employed to assess the models' performance. The results indicate that the RoBERTa model achieves the highest accuracy ratio of 0.981 and the highest mean accuracy of 0.97 (across 10 cross-validation folds) in detecting depression from tweets. This research demonstrates the effectiveness of machine learning and advanced transformer-based models in leveraging social media data for mental health analysis. The findings offer valuable insights into the potential for early detection and monitoring of depression using online platforms, contributing to the growing field of mental health analysis based on user-generated content.

Keywords: depression detection; social media analysis; deep learning models; NLP techniques; user tweets; mental health identification; sentiment analysis; large language models

1. Introduction

1.1. Background

Depression is a prevalent mental health condition that affects a substantial number of individuals worldwide [1]. It is characterized by persistent feelings of sadness, loss of interest, and impaired functioning, leading to a significant decline in overall well-being and quality of life [2]. Timely detection and intervention are crucial for the effective management and treatment of depression. Left untreated, depression can lead to severe impairments in personal, social, and occupational functioning [3].

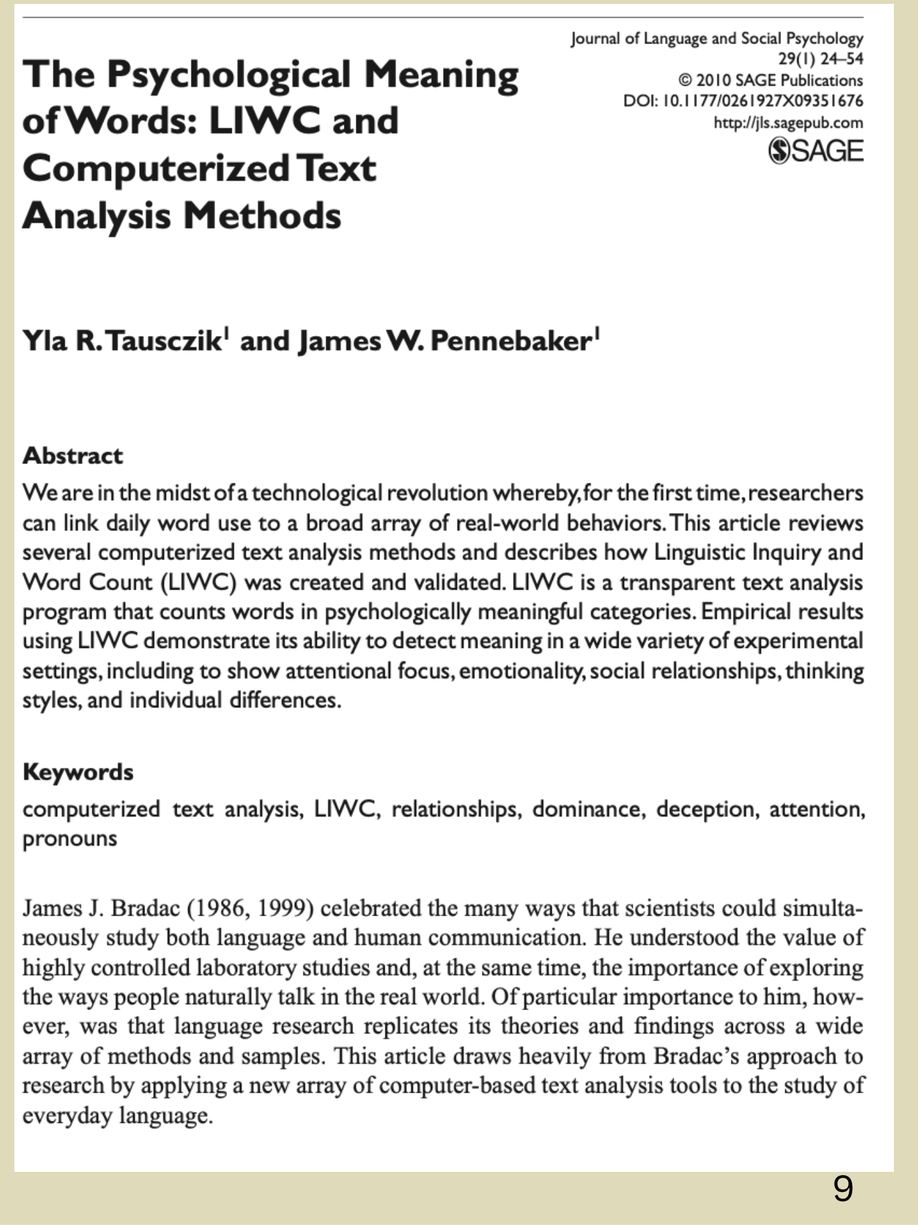
LITERATURE

Tausczik et al. (2010)

Methodology: Reviewed the development and psychometric validation of the Linguistic Inquiry and Word Count (LIWC) program, which automatically categorizes and counts function and emotion words

Performance metric(s) reported: Pearson's r, Cronbach's alpha, Inter-judge agreement percentage

Result: Demonstrated that LIWC reliably identifies attentional focus, emotionality, and individual differences in natural language use and ranked important categories within LIWC



Journal of Language and Social Psychology
29(1) 24–54

DOI: 10.1177/0261927X09351676
http://jls.sagepub.com
SAGE

The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

Yla R. Tausczik[1] and James W. Pennebaker[1]

Abstract

We are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors. This article reviews several computerized text analysis methods and describes how Linguistic Inquiry and Word Count (LIWC) was created and validated. LIWC is a transparent text analysis program that counts words in psychologically meaningful categories. Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences.

Keywords

computerized text analysis, LIWC, relationships, dominance, deception, attention, pronouns

James J. Bradac (1986, 1999) celebrated the many ways that scientists could simultaneously study both language and human communication. He understood the value of highly controlled laboratory studies and, at the same time, the importance of exploring the ways people naturally talk in the real world. Of particular importance to him, however, was that language research replicates its theories and findings across a wide array of methods and samples. This article draws heavily from Bradac's approach to research by applying a new array of computer-based text analysis tools to the study of everyday language.

OUR NOVEL CONTRIBUTION

Longitudinal Data

Utilized four years of real Reddit behavioral data for realistic, in-depth analysis.

Use of reliable techniques

Using clinically sound and proven tools like Linguistic Inquiry and Word Count (LIWC)

Accounting for user timeline

Studies did not account for user timeline and how the crisis rose over a time

DATASET OVERVIEW

Title - The Dataset: Reddit Mental Health Data (2019)

Source: Publicly available on Kaggle

Nature: Textual social media posts scraped across 5 core mental health subreddits

Why this dataset? It captures unsolicited, raw longitudinal thought patterns over time

Ethics: The dataset authors collected this using the official Reddit API. To protect privacy, all personal identifiers are omitted. It is approved purely for academic research.

Size: ~1.8M total raw posts scraped covering January 2019 through August 2022, structured in monthly batches from 85,000 unique users.

Unnamed: 0	author	created_utc	score	selftext	subreddit	title	timestamp
0	erbush1988	1556632225	4	Hello all. incorrect or uneducated i out their issues. I think th e house we are renting an erything in order as it is or felt like she had betrayed around 11pm and by midn hat she wants for her own l gets flustered and overwl familiar with some of the n Please help me help her.	Anxiety	anxiety. How ca	2019-04-30 23:50:25
1	weeblybeebly	1556631109	4	e, the clammy palms, the But I digress. m inside my head to outsi red I'd try. Trying to be ha e morning and remember Well, I feel like shit. y to standing. Tensed all wave blasting through gla cket of warm water thrown nust think I'm so awkward. oming. The doctors appoi ack inside with my negati y to be something your no rogress I made. It was an the only thing real to me. I Maybe a combination his much earlier. It's *actu ite of anxious isn't calm, it to take up room among yo	Anxiety	Anxiety's Kryptor	2019-04-30 23:31:49
2	logicminds	1556630422	1	nan	Anxiety	to avoid actually	2019-04-30 23:20:22
3	kweesnaw	1556629580	2	l script of the Clonidine an	Anxiety	exor, is it okay to	2019-04-30 23:06:20

PREPROCESSING STAGE 1

Issue to resolve

Raw data was an unordered dump of 1.8M isolated posts with noisy deletions.

User Timelines

- Filtering: Automatically dropped standard deleted/removed posts to prevent models from learning false noise artifacts. ([deleted] posts)
- Timestamp Interpolation: Parsed raw UNIX Epoch times (created_utc) into standard Python DateTime objects.
- Temporal Grouping: Algorithmically sorted the entire dataset by author and then by timestamp (ascending).

PREPROCESSING STAGE 2

Issue to resolve

Advanced Transformer models require punctuation (?, !) to understand tone. However, dictionary-based sentiment tools crash if punctuation touches the target word (e.g., matching "sad" against "sad,")

User Timelines

- Built a parallelized cpu-core multiprocessor script mapping 2 distinct streams.
- Stream A (Raw Text): Lowercases text and removes URLs only.
- Stream B (Clean Text): Uses regex [^a-zA-Z0-9\s] to permanently strips out emojis, and removes all non-alphanumeric special characters.

PREPROCESSING STAGE 3

Issue to resolve

Pure text strings cannot be fed into Random Forests or XGBoost.

3.1 : Semantic Embeddings

Passes Stream A text through a fast transformer model (all-MiniLM-L6-v2) to create a 384-dimensional mathematical vector (embedding) of the post's core meaning.

3.2: Linguistic Features

Calculates average sentence length, exclamation density, and posts between midnights.

3.3: Psychological Features

Approximates LIWC (Linguistic Inquiry and Word Count) categories by matching Stream B text against pre-compiled dictionaries for anxiety, sadness, anger, negative emotion, etc.

3.4: Feature Merging

Combining activity (post time), emotion (keywords), subtext (semantic embeddings) and creating a dataset with all these features.

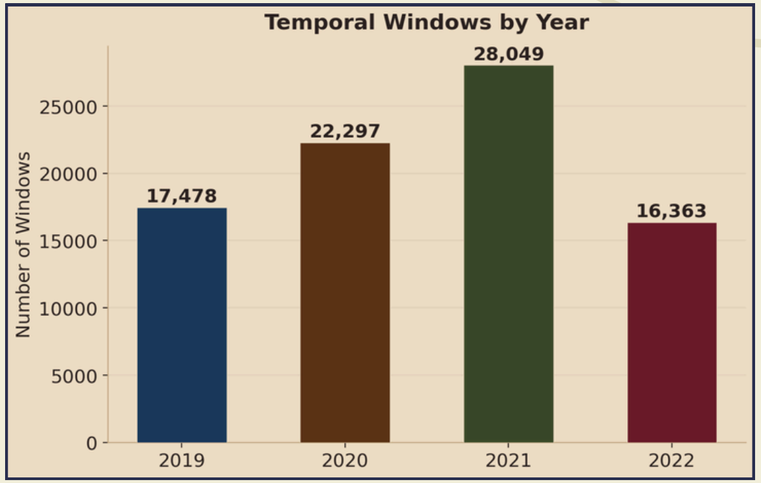
PREPROCESSING STAGE 4

Issue to resolve

A single angry post doesn't indicate a crisis. Natural variance creates massive localized noise.

Post Rolling Window

- Grouped chronologically sequential text into strict blocks of size = 10 posts, with a stride = 5.
- Inside each block, calculated the Intra-Window Mean and Variance (Translating into "Emotional Volatility").
- Deltas (Change): Calculates the exact difference in feature values between Window T and Window T-1 (e.g., did they become more anxious than their own baseline?)
- Embedding Drift: Calculates the cosine distance between the mean embedding vector of Window T and Window T-1.



PREPROCESSING STAGE 5

Issue to resolve

We aren't predicting suicide in real-time; we want to predict the onset before it happens.

Post Rolling Window

- Identify the very first timestamp a user posts in SuicideWatch (the Crisis timestamp).
- Strictly isolated the last 3 progression windows. Recent NLP academic literature on Reddit mental health utilize a 10-to-30 post context window, as this is the exact temporal boundary.
- Tagged these 3 localized progression windows as Label 1 (Pre-Crisis). All other distant history and normal users were tagged as Label 0 (Control).

V1 MODEL

What we did

- Collapsed all temporal windows into a single row per user.
- Each user's entire Reddit history was summarized into:
  - Mean linguistic features
  - Statistical feature distributions
  - Average embeddings
  - Behavioral activity patterns

Additional Behavioral Features

- Posts per week
- Average time gap between posts
- Late-night posting ratio

Labeling

- Users who ever posted in r/SuicideWatch → Crisis User (1)
- Others → Non-Crisis User (0)

Observation

The user-level dataset was almost balanced:

- 48% crisis users
- 52% non-crisis users

REMOVED UNDERSAMPLING

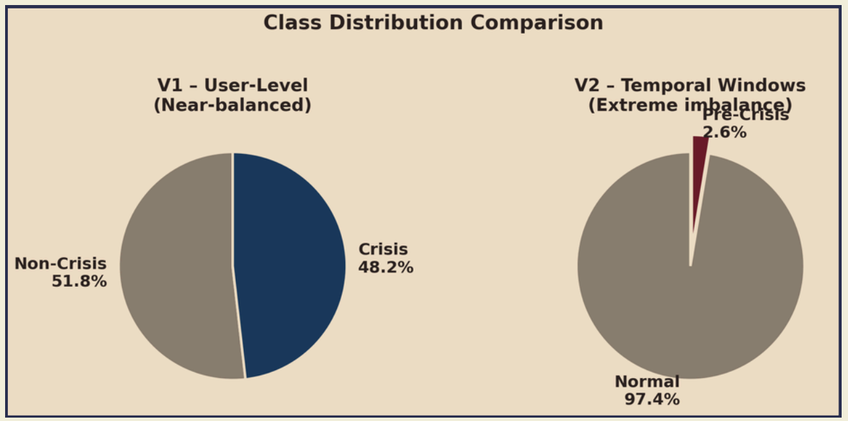
Initially, we planned to undersample the majority class to create class balance.

Problem

The data was already nearly balanced at user level.
Undersampling would:

- Remove thousands of valuable crisis users
- Reduce training information
- Hurt model performance

So, we removed undersampling entirely.



Result

Performance improved significantly because:

- More data was retained
- The model learned richer behavioral patterns

MODEL SELECTION- XG BOOST

Advantages

- Handles high-dimensional features effectively
- Works well with imbalanced data
- Robust against noisy features
- Captures non-linear relationships
- Computationally efficient

Parameters Used

max_depth = 6

learning_rate = 0.05

n_estimators = 300

We intentionally avoided heavy hyperparameter tuning because V1 was mainly a sanity-check experiment.

THRESHOLD OPTIMIZATION

By default, XGBoost predicts positive cases only if probability ⩾ 0.50

Optimal thresholds depend on:

- Class distribution
- Business objective
- Risk sensitivity

Our Approach:

We tested thresholds from: 0.01 → 0.99

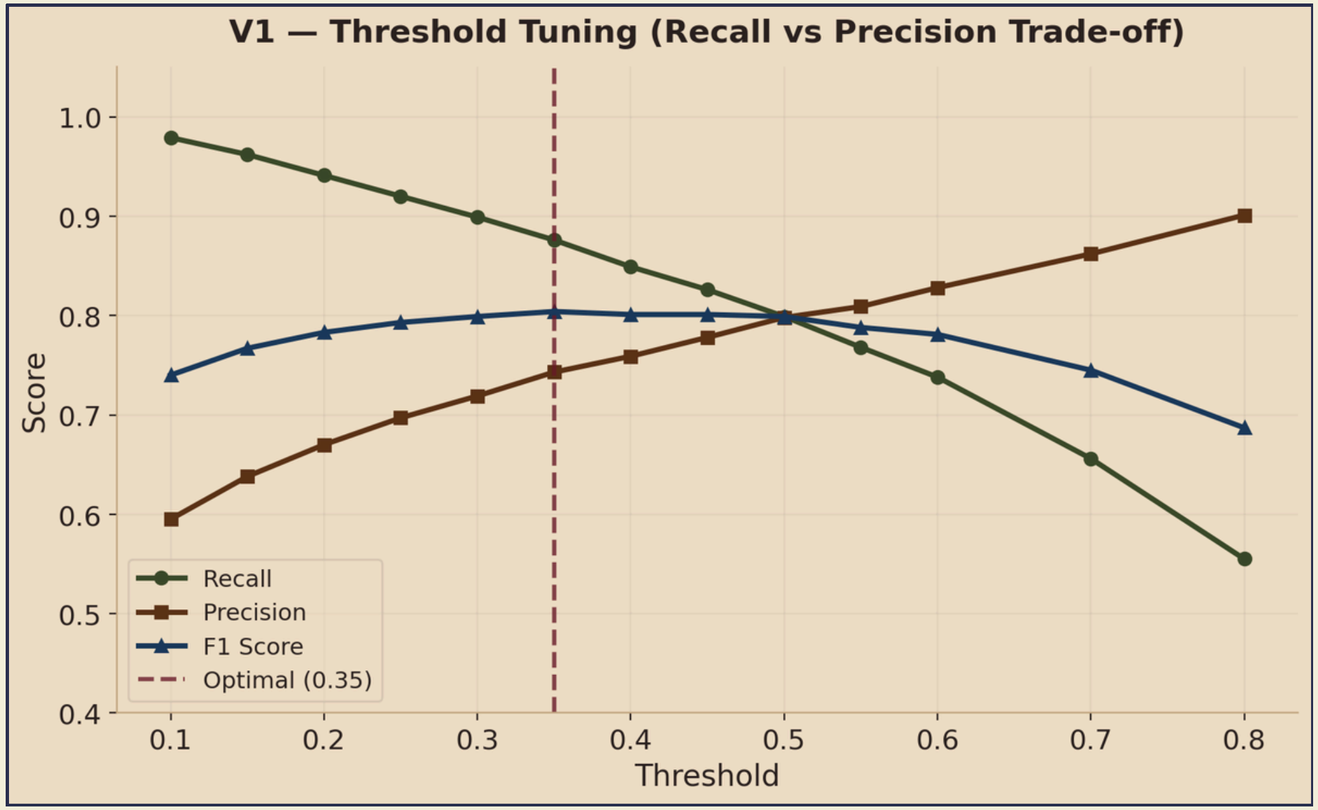
Best Threshold for V1:

Threshold = 0.35

Recall = 87.6%

Precision = 74.3%

Best F1 Score achieved



V1 RESULTS

0.887

AUROC

We can tell crisis users from non-crisis users

87.6%

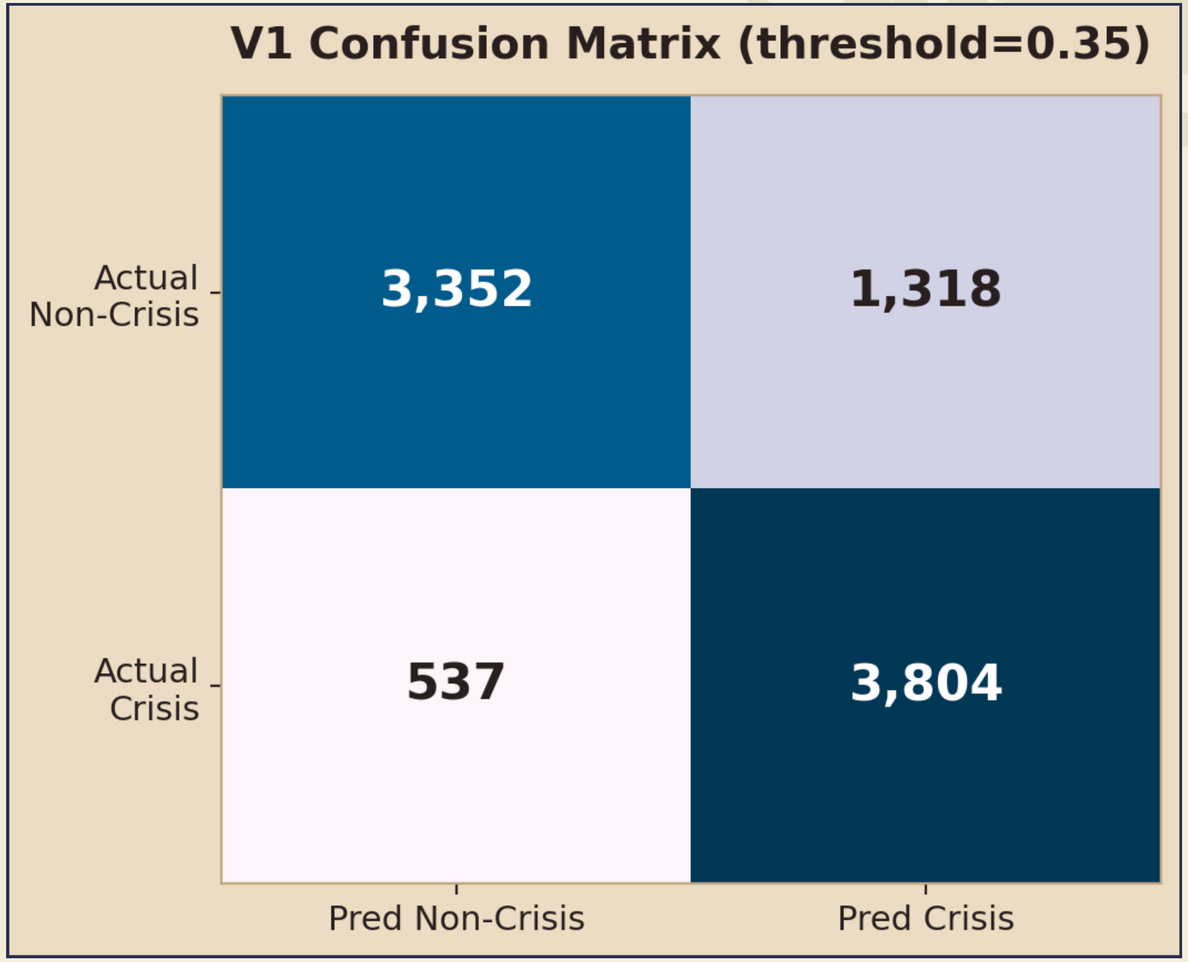
RECALL

Out of 4,341 crisis users, we correctly catch 3,804

74.3%

PRECISION

Out of all users the model flagged as “crisis”, 74.3% were truly crisis users.



V2 Temporal Early Detection Model

Predicting crisis before users reach a severe stage

Key Difference from V1

- Instead of using the full user history, we use recent temporal windows only
- All post-crisis data was removed.
- Using post-crisis behaviour would leak future information into training and create unrealistic performance.
- This ensures:
  Realistic deployment conditions
  Proper early warning behavior

Main Challenge in V2

At temporal window level:
Only **2.59%** windows were positive
This means:
Most windows represent normal behavior
Very few represent pre-crisis behavior

Problem

Without handling imbalance:
The model predicts almost everything as normal
Crisis windows get ignored

Solution

Used:
scale_pos_weight = 37.6
Missing a crisis window is treated as 37.6× more costly than a false alarm.

Why We Prioritized Recall

Precision vs Recall Tradeoff

In Medical & Crisis Systems

False negatives are much more dangerous than false positives

False Positive

User receives unnecessary support message

False Negative

A vulnerable user is missed

Potentially severe consequences

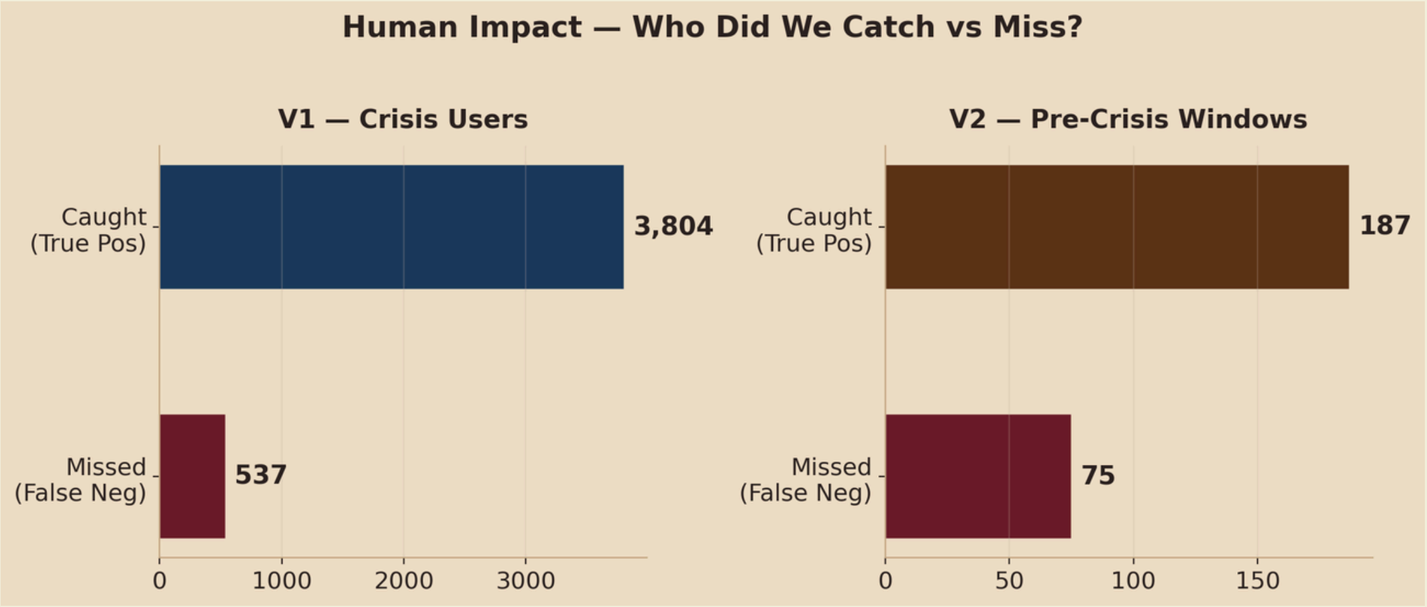
Decision

We intentionally optimized for higher recall.

Threshold = 0.06

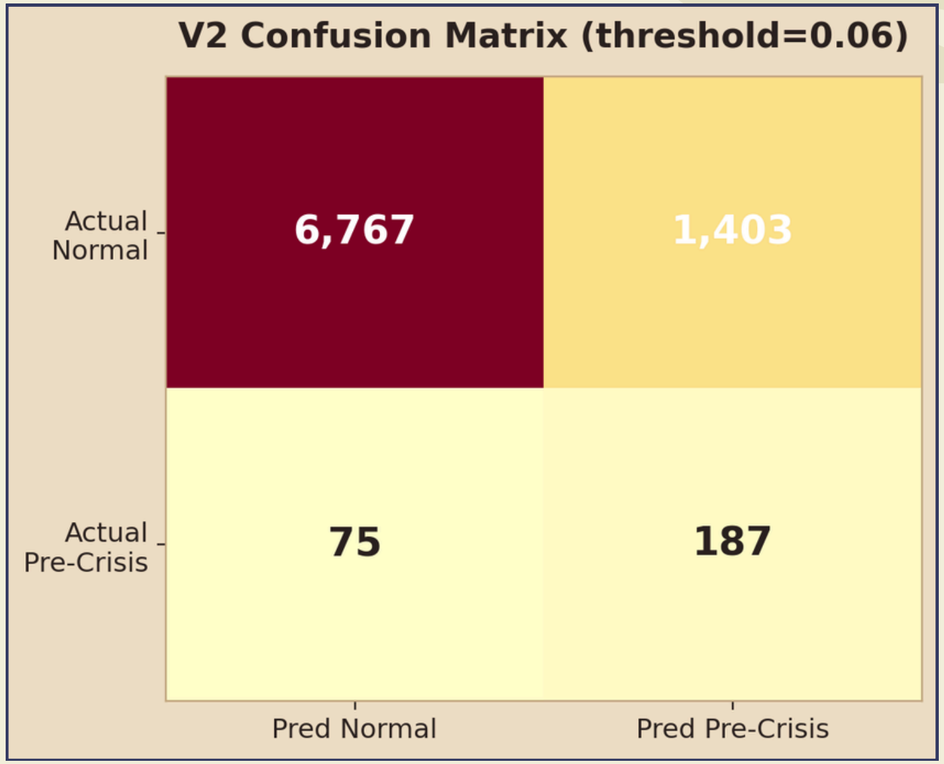
This helped us detect:

71.4% of pre-crisis windows



V2 RESULTS

0.884	71.4%	11.8%
AUROC	RECALL	PRECISION



Interpretation

The model is highly effective at ranking crisis windows correctly.
Although precision is lower:
False alarms are acceptable in mental health systems
High recall is more important for safety-critical applications

Key Achievement

Successfully detects most crisis windows before escalation.

TF-IDF VS TRANSFORMER EMBEDDINGS

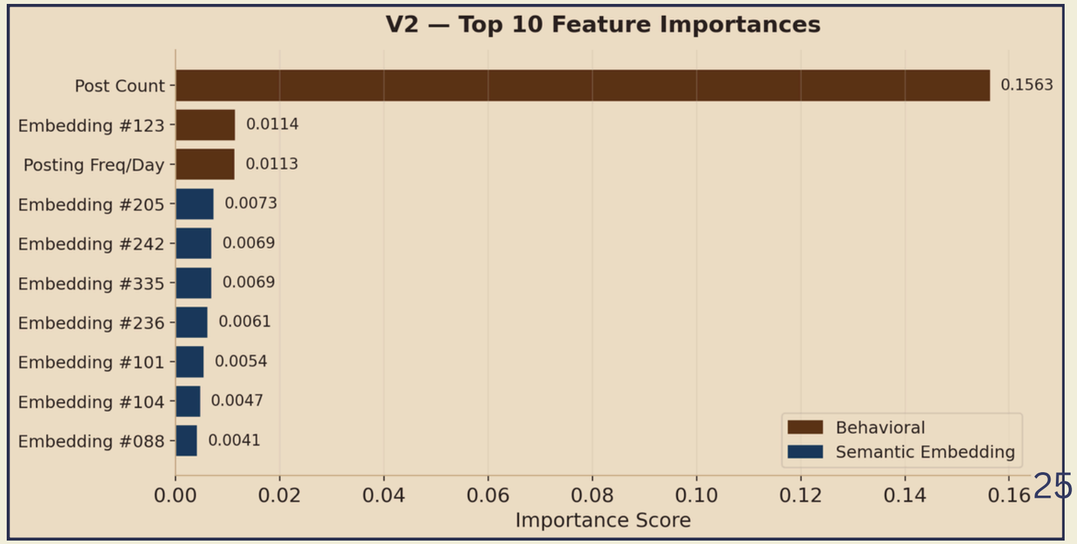
We compared:
Traditional TF-IDF features vs Transformer embeddings

TF-IDF
Counts words
No semantic understanding

Transformer Embeddings
Understand meaning and context
Capture emotional tone and intent
Better for subtle psychological patterns

Result
Transformer embeddings outperformed TF-IDF across all metrics.

METRIC	EMBEDDINGS	TF-IDF	WINNER
RECALL	0.876	0.767	EMBEDDINGS
PRECISION	0.743	0.687	EMBEDDINGS
F1 SCORE	0.804	0.725	EMBEDDINGS
AUROC	0.887	0.847	EMBEDDINGS



CHALLENGES FACED

Extreme Class Imbalance

Very few positive crisis windows.

Noisy Social Media Data

Informal language, slang, inconsistent text.

Computational Limitations

Large-scale embedding generation and temporal processing required significant resources.

Ethical Concerns

Privacy

False positives

Responsible deployment

Deployment Proposal: Supporting Student Well-being

01 **Dedicated Student Community**

A private Reddit space tailored for Plaksha students, fostering open communication and mutual support.

02 **System Functionality**

Monitor behavioral patterns
Identify potentially vulnerable users

03 **Ethical & Privacy-First Approach**

Emphasizing human oversight, strict moderation, and transparent privacy policies to build trust and ensure safety.

References and Sources

Paper 1 - https://ojs.aaai.org/index.php/ICWSM/article/view/14432/14281

Paper 2 - https://hcsi.cs.tsinghua.edu.cn/Paper/Paper17/IJCAI17-SHENGUANGYAO.pdf

Paper 3 - https://www.mdpi.com/2075-4426/14/9/958

Paper 4 - https://www.cs.columbia.edu/~julia/papers/azucaretal2017.pdf

Paper 5 - https://www.mdpi.com/2079-9292/12/21/4396

Paper 6 - https://www.cs.cmu.edu/~ylataus/files/TausczikPennebaker2010.pdf

Dataset:

Dataset source: https://www.kaggle.com/datasets/entenam/reddit-mental-health-dataset

Github Repository:

Repository - https://github.com/wherearemyavocadoes/MLPR-Final-Project.git