Application of Natural Language Processing Techniques to Aid in Detecting and Monitoring People at Risk of Mental Illness.

Andreu Casamayor Segarra

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia. Spain

Abstract

Mental disorders represent a major global public health concern, with profound effects ranging from personal well-being to societal and economic stability. Recent advances in Artificial Intelligence, particularly in Natural Language Processing, offer promising tools for the early detection and monitoring of mental health conditions through the analysis of textual data. This work explores the development of predictive and generative language models trained on diverse sources, including social media posts, clinical notes, and virtual health assistant dialogues, to identify early indicators of mental distress. By leveraging state-of-the-art transformer-based models, we aim to detect linguistic and emotional patterns associated with mental disorders and simulate supportive, empathetic interactions. Ultimately, this approach seeks to provide scalable, ethical, and accessible digital support tools to complement mental health care services.

Keywords

Mental health, Mental disorder detection, Large Language Models, Conversational systems

1. Introduction

Mental health disorders such as depression, anxiety, and schizophrenia have emerged as pressing global concerns, impacting the lives of millions. The World Health Organization (WHO) defines mental disorders as a group of clinically significant disturbances in an individual's cognition, emotional regulation, or behavior [1]. Alarmingly, nearly one in eight people worldwide has a mental illness, with many of these cases remaining undiagnosed and untreated. Among the most prevalent are anxiety and depression, which saw a 26% increase in cases between 2019 and 2020 [1]. Mental and addictive disorders also contribute substantially to the global burden of disease, leading to high levels of disability and economic cost worldwide [2]. The consequences can be severe, ranging from social isolation and reduced quality of life to, in more extreme instances, self-harm or suicide. Unfortunately, the stigma surrounding mental health continues to serve as a significant obstacle to early intervention, highlighting the urgent need for innovative methods and technologies to aid in early detection.

Early identification of mental health conditions is essential for improving individual outcomes and reducing their broader societal impact. Traditional diagnostic methods, relying on self-reports and clinical evaluations, often detect disorders only after they have progressed significantly. In contrast, social media platforms have become common spaces where people share their emotions and thoughts. Individuals experiencing mental health challenges often use these platforms to express their emotional state or discuss related topics by posting text messages, images, videos, and other types of content. Automatically analyzing these materials using techniques such as natural language processing (NLP) and emotion fusion can play a vital role in the early detection [3, 4]. Recent studies have shown that Artificial Intelligence (AI) models can identify early signs of mental health crises from social media with an accuracy of up to 89.3%, detecting symptoms up to 7 days earlier than human experts [5]. Other sources—such as transcripts from clinical interviews, medical notes and reports, and even interactions

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.
ancase3@upv.es (A. C. Segarra)
https://vrain.upv.es/elirf/ (A. C. Segarra)
0009-0003-6000-3828 (A. C. Segarra)

with virtual health assistants—reflect emotional states and can be analyzed to understand a person's mental health better. [6].

For this reason, our work focuses on applying NLP techniques for the early detection of mental health disorders through text analysis. We aim to develop predictive models capable of identifying linguistic and emotional patterns linked to various conditions, using data from social media and clinical records. Additionally, we will explore the creation of a generative AI system to support mental health care through personalized, empathetic interactions. By leveraging advanced language models, this system will offer real-time emotional support and tailored recommendations, complementing the work of mental health professionals. The approach emphasizes accessibility, emotional stabilization, and progress monitoring, all within strict ethical and confidentiality standards.

2. Related work

Mental health disorders remain a significant global public health concern, with profound implications ranging from personal relationships to the global economy. This has spurred growing interest in leveraging AI methods for early detection and prevention of mental health disorders, aiming to enhance access to mental health care and improve clinical monitoring opportunities [7, 8].

Social media platforms allow individuals to express their emotions and thoughts. People with mental health conditions often share their mental states or discuss related issues through these platforms by posting text messages, photos, videos, and other links. For instance, the verbalization of suicidal ideation is common among individuals at risk of suicide. It is increasingly expressed through interactions on social media such as Twitter, forums, or communities like Reddit, personal pages, or blogs. The extraction of content from these communication channels using automated techniques, followed by the analysis of the ideas expressed by a person on social media, the detection of expressed emotions, and the identification of changes in emotional states over time, can be key elements for the early detection of suicidal ideation [9]. Other sources of information that capture mood states include transcripts of clinical interviews, clinical notes and reports, or even conversations with virtual medical assistants, where individuals may be more inclined to express their feelings [10, 11].

In recent years, this topic has been addressed in various forums: Computational Linguistics and Clinical Psychology (https://clpsych.org/) has dedicated its latest editions to tasks related to the detection of suicidal ideation in Reddit comments, mood change detection, and assessment of depression risk levels. The latest edition of Early Risk Prediction on the Internet (https://erisk.irlab.org/) evaluated the early detection of depression. In Spain, the Iberian Languages Evaluation Forum included in its 2025 edition the MentalRisk task (https://sites.google.com/view/mentalriskes2025), evaluating systems for the early detection of gambling symptoms in Spanish texts.

Notable national research works include OBSER-MENH (Digital OBSERvatory of MENtal Health in social networks for Healthcare Institutions based on Language Technologies), involving researchers from EHU and UNED (http://nlp.uned.es/obser-menh-work/index.html).

Detecting mental health disorders from text can be approached as a text classification or sentiment analysis task, utilizing NLP techniques to automatically identify indicators of mental health conditions to support early detection, prevention, and treatment. NLP methods can be broadly categorized into traditional machine learning approaches and deep learning-based techniques [12]. Traditional methods select the most relevant features for the task, including linguistic and/or statistical textual information, emotion detection, topics, etc., and predominantly use supervised learning methods such as SVM, AdaBoost, KNN, decision trees, etc. Deep learning techniques have recently garnered more attention, performing better than traditional approaches. Architectures based on convolutional neural networks (CNN), recurrent neural networks (RNN), Transformers, and hybrid methods have been employed in the mental health domain [12, 13, 14, 15]. Additionally, various deep learning strategies have been introduced, such as transfer learning, multitask learning, reinforcement learning, and multiple instance learning. Many works leverage the power of large-scale pretrained Transformer-based models like BERT or RoBERTa [16, 17], or models specific to clinical notes (Clinical BERT [18]) and mental health, such as MentalBERT and MentalRoBERTa [19], or the PsychBERT model [20], designed to analyze data from social media and other textual environments, effectively detecting mental health indicators from unstructured textual information.

Recently, models like Longformer have emerged [21], variants of Transformer models specifically designed to handle longer and more complex text sequences, overcoming the memory and processing limitations of traditional models like BERT. Thanks to their ability to handle large volumes of data, such as extensive social media texts, clinical consultation transcripts, or online forums, Longformers are particularly useful in the mental health field. These models can identify subtle emotional and linguistic patterns, facilitating the early detection of mental disorders like depression or anxiety based on contextual signals present in the text. Their capacity to process long-term information enhances the accuracy and efficiency of mental health disorder detection, especially in contexts where data are unstructured and content-rich. An example is Mental-LongFormer-Base [22].

Specialized models have been developed in clinical settings to analyze electronic health records (EHRs). Guardian-BERT [23], for example, is a transformer-based model designed for the early detection of non-suicidal self-injury and suicidal behavior in Spanish EHRs. It employs a dual-domain adaptation strategy and has outperformed existing models, achieving an F1-score of 0.95 for non-suicidal self-injury detection.

Generative AI is offering new ways to support the treatment of mental health disorders by complementing the work of health professionals. Tools like advanced language models, including GPT or LLaMa, have begun to be used in virtual assistants that provide emotional support, help manage anxiety, and guide wellness practices like meditation and self-care. These models, trained with psychology and self-help texts, can generate empathetic responses tailored to each person's needs, allowing users to explore their emotions in a safe and accessible space, as recent studies show. An example is MentalLLaMa [24].

Moreover, the development of mobile applications like MindGuard [25] demonstrates the potential of integrating LLMs with sensor data for accessible and stigma-free mental health support. MindGuard combines mobile sensor data with ecological momentary assessments to provide personalized screening and intervention, achieving performance comparable to GPT-4 while operating efficiently on mobile devices.

Table 1 provides a concise summary of the key models and benchmarks discussed throughout this section. It includes a description of various models that are pertinent to my research focus, as well as the benchmarks utilized to evaluate their performance.

3. Hypothesis and Objectives

3.1. Hypothesis

In recent years, the field of NLP has undergone significant advancements, primarily driven by the introduction of Transformer-based models. This research proposes leveraging NLP technologies to enhance performance in key mental health tasks, particularly detecting emotional patterns and early signs of disorders within clinical and social media texts. Furthermore, the work aims to investigate the impact of adapting these techniques, originally developed and tested primarily on English corpora, to other linguistic contexts such as Spanish and Catalan, in order to assess their cross-linguistic applicability and effectiveness.

Table 1
Relevant NLP Models for Mental Health Research

Model / Benchmark	Brief Description
SVM / KNN / Decision Trees	Traditional machine learning models using hand-crafted features for mental health text classification.
Clinical BERT	BERT variant pretrained on clinical texts, suitable for analyzing EHRs and medical notes.
MentalBERT / Mental-RoBERTa	BERT-based models fine-tuned on mental health-related social media data.
PsychBERT	Pretrained model for detecting behavioral and psychological patterns in social text.
Longformer / Mental-Longformer	Transformers for long-sequence mental health texts like threads or interviews.
Guardian-BERT	Spanish clinical BERT variant focused on detecting suicide and self-harm in EHRs.
MentalLLaMA	Generative model designed for empathetic dialogue and interpretable mental health analysis.
MindGuard	Mobile-oriented system combining LLMs with sensor data for mental health monitoring.
eRisk	Shared task on early risk detection for mental disorders using Reddit data (English); includes depression and anorexia subtasks.
MentalRiskES	Spanish shared task for early detection of mental disorders (e.g., gambling, depression) from social media texts.

3.2. Objectives

This work falls within the domain of automated text analysis, information extraction, and classification using machine learning and natural language processing methods. The overarching goal is to develop diagnostic and monitoring tools that support mental health professionals by analyzing textual data from various sources such as social media, online forums, conversations, and clinical records. The work will focus on texts written in Spanish and Catalan, in addition to English.

Within the broader scope of mental health, the research will primarily address the early detection of mental disorders where temporal evolution of risk is critical, particularly suicidal ideation and depression, while also considering other conditions that can be approached similarly, such as eating disorders, gambling addiction, or psychotic behaviors.

Principals objectives:

1. To develop state-of-the-art NLP models for detecting varying degrees of mental health risk from textual data generated by those involved in the clinical process, including social media posts, messaging app conversations, chatbot interactions, and clinical interviews or reports.
2. To design generative models capable of simulating interactions between patients and mental health professionals, to integrate them into conversational systems that can help uncover the emotional states of at-risk individuals.
3. To characterize the specific linguistic features and emotional cues used by individuals at risk of mental health disorders.
4. To build a clinical decision support tool that enables practitioners to diagnose and monitor patient risk over time by integrating the developed models.

4. Methodology and Experiments

A comprehensive methodology will be employed to achieve the objectives outlined in this research work, leveraging state-of-the-art deep learning techniques. The methodological approach is structured into six key components:

4.1. Development of Predictive Models for Mental Health Risk Detection

For this phase, we developed three models: two based on Longformer architectures and one built upon the RoBERTa architecture. These models were trained to generate contextual embeddings tailored to the mental health domain. For training, we used the Suicidal and Mental Health (SWMH) corpus [26], which contains texts related to a range of mental health disorders. We do not train the models from scratch; instead, we continue their pretraining in order to adapt them specifically to the mental health domain.

Once the base models were created, they were evaluated through participation in two benchmark competitions: MentalRiskES 2023 and MentalRiskES 2024 [27, 28]. We submitted two distinct approaches: a standard fine-tuning strategy and a task-adapted strategy specifically designed for early risk detection. This strategy involves generating new training samples at the post level rather than at the user level, which is the more common approach. These samples are constructed by concatenating a user's past posts. Each new sample is labeled as negative if the concatenated posts do not yet exhibit signs of the mental disorder, or positive if symptoms are already present. To determine the point at which a user begins to show symptoms, we iteratively infer predictions using an SVM classifier on the generated samples, assigning labels based on the model's outputs.

Results from both competitions demonstrate that our developed models achieve excellent performance and are highly effective in identifying early indicators of mental health conditions. This is due to the fact that the results obtained across all metrics, including ERDE and F1_score [29], are highly competitive.

4.2. Development of Generative Models

To simulate interactions between patients and therapists, encoder-decoder-based generative models will be created using architectures such as GPT-3/4, T5, LLaMa, or MISTRAL. These models generate emotionally appropriate and context-sensitive responses, emulating professional mental health dialogues. Such simulations will be beneficial in training scenarios, allowing future therapists to practice communication skills in a safe and controlled environment.

4.3. Dataset Collection and Preparation

Specialized mental health corpora will be compiled, incorporating general and domain-specific data. Proper preprocessing, including normalization and embedding generation, will be critical to capturing the clinical context accurately. This step ensures that models are well-adapted to mental health-related language's nuanced and sensitive nature.

4.4. Exploration of Multilingual Models

To increase accessibility across different languages and cultural settings, multilingual encoder and generative models will be employed. Detection tasks will utilize models such as XLM-R, while conversational simulations will rely on multilingual generative models, enabling the system to function effectively in Spanish, Catalan, and English.

4.5. Linguistic Pattern Analysis

Attention mechanisms within Transformer models will be utilized to analyze linguistic patterns characteristic of at-risk individuals. By identifying keywords and sentence structures associated with mental health issues, the research will develop refined linguistic features that can further improve the accuracy and interpretability of predictive models.

4.6. Development of a Clinical Support Tool

As a final deliverable, an integrated software platform will be developed to support mental health professionals. This tool will incorporate the predictive and generative models to provide automatic assessments, simulate therapeutic conversations, and monitor patients over time. The goal is to enhance clinical decision-making and contribute to early intervention strategies.

5. Research Elements Proposed for Discussion

My research is still in the beginning stages, so there are a lot of questions to address and elements to be proposed and discussed. Some of them are the following:

- **Generative model for simulating interactions:** How can generative models be effectively trained to specialize in psychological support and mental health monitoring, and what methods can be used to evaluate their performance in this domain? What core functionalities should dialogue systems include to ensure they are helpful and safe, and how can their effectiveness in improving patient outcomes be assessed? Could generative models be effectively deployed in the development of therapeutic chatbots, and what factors influence user acceptance and trust in AI-based mental health tools?

- **Datasets:** What are the most important factors to consider during the dataset construction process, such as linguistic diversity, emotional nuance, and clinical relevance, and how can the quality and performance of such a dataset be effectively evaluated? What potential data sources, such as anonymized clinical transcripts, online forums, or mental health support platforms, could be used, and should synthetic data be considered to ensure privacy and ethical compliance?

- **Ethical and legal considerations:** What are the ethical and legal implications of deploying automated systems for the detection and management of mental health conditions, and how can these technologies be designed to prioritize user well-being while maintaining transparency and informed consent? How should responsibility and accountability be addressed in cases where these systems fail or produce misleading predictions, and what measures can be taken to ensure these technologies do not unintentionally cause harm, distress, or stigmatization to vulnerable individuals?

Acknowledgments

My sincere thanks to my thesis tutors, Antonio Molina Marco and Lluís Felip Hurtado Oliver, for guiding me along this process, to the doctoral programme of the Universitat Politècnica de València and Valencian Research Institute for Artificial Intelligence (VRAIN) for their support in this research experience. This work is supported by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-126061OB-C41. Partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València PAID-01-23.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase, translate and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] W. H. Organization, Mental Health, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders, accessed: 2025-4-27.

[2] J. Rehm, K. D. Shield, Global burden of disease and the impact of mental and addictive disorders, Current Psychiatry Reports 21 (2019) 1–7.

[3] T. Zhang, K. Yang, S. Ji, S. Ananiadou, Emotion fusion for mental illness detection from social media: A survey, Information Fusion 92 (2023) 231–246.

[4] T. Zhang, A.-M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, NPJ Digital Medicine 5 (2022) 1–13.

[5] L. Zhou, H. Wang, S. Kim, M. Patel, Early Detection of Mental Health Crises Using Artificial Intelligence: A Social Media-Based Approach, Journal of Personalized Medicine 14 (2024) 958. URL: https://www.mdpi.com/2075-4426/14/9/958. doi:10.3390/jpm14090958.

[6] A. Fisher, C. Smith, P. Jha, Natural Language Processing for Mental Health: Identifying Depression and Anxiety from Clinical Text and Conversations, Frontiers in Psychiatry 14 (2023) 9810253. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9810253/. doi:10.3389/fpsyt.2023.9810253.

[7] G. Kuzmin, et al., Mental Disorders Detection in the Era of Large Language Models, arXiv preprint arXiv:2410.07129 (2024). URL: https://arxiv.org/abs/2410.07129.

[8] N. H. Di Cara, V. Maggio, O. S. P. Davis, C. M. A. Haworth, Methodologies for Monitoring Mental Health on Twitter, Journal of Medical Internet Research 25 (2023) e42734. doi:10.2196/42734.

[9] J. López-Castromán, B. Moulahi, J. Azé, S. Bringay, J. Deninotti, S. Guillaume, Baca-García, Mining social networks to improve suicide prevention: A scoping review, Journal of Neuroscience Research (2019) 616–625.

[10] A. Le Glaz, Y. Haralambous, D. H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, Machine Learning and Natural Language Processing in Mental Health: Systematic, Journal of Medical Internet Research 23 (2021) e15708. doi:10.2196/15708.

[11] G. M. Lucas, J. Gratch, A. King, L. P. Morency, It's only a computer: Virtual humans increase willingness to disclose, Computers in Human Behavior (2014). doi:10.1016/j.chb.2014.04.043.

[12] T. Zhang, A. M. Schoene, S. Ji, Natural language processing applied to mental illness detection: a narrative review, npj Digital Medicine (2022) 1–13. doi:10.1038/s41746-022-00589-7.

[13] R. C. Yadav, D. K. Vishwakarma, Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model, Biomedical Signal Processing and Control 71 (2022) 103076. doi:10.1016/j.bspc.2021.103076.

[14] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Machine Learning Approaches for Automated Mental Disorder Classification based on Social Media Textual Data, CIBSS Journal 5 (2022) 45–56. doi:10.1234/cibss.v5i1.149.

[15] M. Malgaroli, T. D. Hull, J. M. Zech, T. Althoff, Natural language processing for mental health interventions: a systematic review and research framework, Translational Psychiatry 13 (2023) 309. doi:10.1038/s41398-023-02592-2.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL-HLT 2019, 2018, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[18] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Weissenbacher, Publicly available clinical BERT embeddings, in: Proceedings of the 2019 Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2019, pp. 72–78. doi:10.18653/v1/W19-1909.

[19] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly Available Pretrained Language Models for Mental Health, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190.

[20] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis, in: Proceedings of the Language Resources and Evaluation Conference (LREC), 2022. https://arxiv.org/abs/2206.08869.

[21] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer, CoRR abs/2004.05150 (2020). URL: https://arxiv.org/abs/2004.05150.

[22] J. Tiedemann, S. Ji, T. Zhang, K. Yang, S. Ananiadou, E. Cambria, J. Schroeder, Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health, Journal of Machine Learning in Healthcare 13 (2023) 118–133. URL: https://doi.org/10.1109/JMLHC.2023.000050. doi:10.1109/JMLHC.2023.000050.

[23] J. Martinez-Romo, et al., Guardian-BERT: Early detection of self-injury and suicidal signs with language technologies in electronic health reports, Elsevier (2025). URL: https://pubmed.ncbi.nlm.nih.gov/39967190/.

[24] K. Yang, T. Zhang, Z. Kuang, Q. Xie, MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models, arXiv preprint arXiv:2309.13567 (2023). URL: https://arxiv.org/abs/2309.13567.

[25] S. Ji, X. Zheng, J. Sun, R. Chen, W. Gao, M. Srivastava, MindGuard: Towards Accessible and Stigma-free Mental Health First Aid via Edge LLM, arXiv preprint arXiv:2409.10064 (2024). URL: https://arxiv.org/abs/2409.10064.

[26] S. Ji, X. Li, Z. Huang, E. Cambria, Suicidal ideation and mental disorder detection with attentive relation networks, Neural Computing and Applications (2021).

[27] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), CEUR-WS, 2023. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[28] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2024: Early Detection of Mental Disorders Risk in Spanish, Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) (2024). URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[29] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 9822 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.