INTRODUCTION

Mental health disorders, such as depression and anxiety, represent a growing global crisis, affecting millions annually. Early detection and intervention are essential to improving outcomes, but traditional approaches—such as clinical interviews and self-reported questionnaires—are often resource-intensive, inaccessible to many, and unable to provide real-time solutions. Meanwhile, social media platforms like Twitter and Reddit have become digital spaces where some individuals share thoughts and emotions, offering valuable, though not universal, insights into mental health trends.

Sentiment analysis, a natural language processing (NLP) technique, has emerged as a promising tool for analyzing such digital footprints to detect emotional patterns. By classifying text into positive, negative, or neutral sentiments, it holds potential for identifying early signs of mental health concerns. However, this application faces significant challenges, including a lack of annotated datasets specific to mental health (data sparsity) and the complex, often opaque decision-making processes of advanced models like BERT and RoBERTa (interpretability). These limitations hinder the adoption of sentiment analysis tools for clinical and real-world use.

This study investigates the use of sentiment analysis and transformer-based models for mental health monitoring through social media. By addressing issues like data sparsity and interpretability, this research aims to contribute to developing reliable and scalable tools for early detection and intervention in mental health care.