1. INTRODUCTION

Mental health is a crucial global topic that requires immediate attention, as the World Health Organization (WHO) estimates that one in every four people may have a mental health disorder at some point in their lives [1]. Mental health problems that haven't been treated, particularly depression, can lead to severe consequences such as suicide and self-harm, making early detection critical in mitigating these outcomes [2]. In this situation, social media sites like Twitter (now called X) have become useful places to get real-time, user-generated content that can offer valuable insights into public sentiment about mental health issues. These platforms generate vast textual datasets that, when analyzed, can provide transformative potential in identifying emotional patterns, detecting at-risk populations, and informing public health interventions [3]. Social media platforms like Twitter serve as modern-day journals, capturing the day-to-day sentiments and emotions of individuals from diverse demographic backgrounds. By analyzing the data, public health professionals and researchers can gain insights into societal attitudes toward mental health, thus enabling timely and targeted interventions.

Sentiment analysis is a systematic method used in natural language processing (NLP) to automatically sort through textual data and find subjective feelings, views, and attitudes [4]. In the case of Twitter, sentiment analysis offers the possibility of detecting subtle emotional cues embedded in short messages, making it particularly useful for identifying individuals expressing mental health concerns. The swift proliferation of social media platforms has generated substantial volumes of unstructured data, posing both opportunities and concerns. Traditional text analysis techniques often struggle to handle such large and varied datasets, requiring more sophisticated approaches such as machine learning to effectively process and interpret the information. Over the last ten years, traditional machine learning algorithms like Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM) have been utilized extensively to tweet-level sentiment classification. These algorithms have shown considerable success in identifying the sentiment in social media posts, but they often encounter limitations when dealing with the complexities and nuances of human language.

Initial investigations in the domain of mental health detection predominantly concentrated on the extraction of textual characteristics from social media posts, employing machine learning methodologies like Support Vector Machine (SVM) in the year 2020 [5] and Random Forest in year 2019 [6] for sentiment classification. The research in year 2023 [7], applied the SVM algorithm to classify depression from approximately 10,000 comments on Facebook and YouTube, where SVM yielded the highest accuracy among tested algorithms, reinforcing its robustness in sentiment classification across different social media platforms. Similarly, the research in year 2022 [8] employed Random Forest alongside Word2Vec feature representations to identify depression symptoms on Twitter achieving 68.75% accuracy, indicating the approach's efficacy in supporting early mental health interventions via tweet analysis. While these early methods demonstrated the potential of machine learning in mental health detection, they were often limited by the inherent challenges of human language, especially in informal settings like social media. Sentences may be fragmented, contain abbreviations or slang, and may not follow the typical grammatical structures expected by many traditional machine learning models. These factors reduce the effectiveness of conventional algorithms in interpreting the sentiments expressed by users, especially those related to complex mental health conditions.

Sentiment analysis involves discerning whether a written work conveys a positive, negative, or neutral tone. It proves particularly beneficial for monitoring mental well-being in the context of social media. Within the realm of mental health, the application of sentiment analysis serves to uncover adverse emotional states, including depression and anxiety, thereby facilitating the prompt recognition of individuals who may be at risk. However, these tasks are not without challenges. Social media platforms, contain a wealth of slang, emojis, hashtags, and misspellings, making it difficult for traditional models to identify emotional undercurrents. Moreover, a single post may convey a complex mixture of emotions, which can be hard to categorize using that simple positive-negative classification scheme. This complexity necessitates the use of more advanced techniques, particularly ensemble learning methods, to achieve more accurate and reliable results in sentiment classification.

While traditional machine learning methods like SVM and Random Forest have been widely applied to tweet classification, these models face challenges due to the informal, often slang-filled language found on social media platforms. Furthermore, these models may struggle to understand the complexities and subtleties inherent in human communication, which can make sentiment analysis on platforms like Twitter particularly difficult.

In response to these challenges, ensemble learning has surfaced as a compelling solution. Ensemble learning constitutes a sophisticated approach wherein a collection of individual models, referred to as base models, are amalgamated to forge a more robust and precise predictive model [9]. By combining the strengths of several models, ensemble learning can overcome the limitations of single-model classifiers, particularly when applied to complex tasks like sentiment analysis. Popular ensemble techniques include bagging, boosting, and voting, which have shown to improve the performance of sentiment classification tasks [10]. Ensemble learning methods leverage the diversity of individual based models to achieve better generalization and higher accuracy.

The main advantage of ensemble methods is their ability to improve prediction accuracy by mitigating biases that may exist in individual models and reducing the risk of overfitting. By combining multiple models, ensemble techniques reduce the possibility of a single model making an error in the prediction, as the diversity of models leads to more reliable outcomes. In the context of mental health sentiment analysis on social media, ensemble learning is particularly advantageous, as it integrates the diverse strengths of classifiers like SVM, RF, and NB. This integration helps to better capture subtle emotional cues and provides a more reliable solution for detecting mental health-related sentiments.

This research aims to present an ensemble learning framework that thoughtfully integrates RF, NB, and SVM classifiers via a refined ensemble methodology [11]. By employing systematic hyperparameter optimization and dynamic weight adjustment based on validation performance, the proposed framework aims to enhance the detection of subtle emotional expressions, improve overall classification accuracy, and offer a more robust solution for mental health sentiment analysis on Twitter data. This research seeks to tackle the complexities of detecting mental health-related sentiments in social media posts, improving upon previous single-model methods by creating a more accurate and generalizable ensemble model.

Furthermore, the research aims to contribute to the development of scalable, data-driven solutions for detecting mental health risks in real-time, which can be vital for timely interventions in public health. The outcomes of this study could potentially help public health agencies and mental health professionals identify at-risk individuals based on their online behavior, providing opportunities for early intervention and support.