International Journal of Web Services Research
Volume 21 • Issue 1

Predictive Analytics in Mental Health Leveraging LLM Embeddings and Machine Learning Models for Social Media Analysis

Ahmad Radwan, Arab American University, Palestine
Mohannad Amarneh, Arab American University, Palestine
Hussam Alawneh, Arab American University, Palestine
Huthaifa I. Ashqar, Arab American University, Palestine
Anas AlSobeh, Southern Illinois University, Carbondale, USA & Yarmouk University, Jordan
https://orcid.org/0000-0002-1506-7924
Aws Abed Al Raheem Magableh, Yarmouk University, Jordan & Prince Sultan University, Saudi Arabia*
https://orcid.org/0000-0003-4513-6430

ABSTRACT

The prevalence of stress-related disorders has increased significantly in recent years, necessitating scalable methods to identify affected individuals. This paper proposes a novel approach utilizing large language models (LLMs), with a focus on OpenAI's generative pre-trained transformer (GPT-3) embeddings and machine learning (ML) algorithms to classify social media posts as indicative or not of stress disorders. The aim is to create a preliminary screening tool leveraging online textual data. GPT-3 embeddings transformed posts into vector representations capturing semantic meaning and linguistic nuances. Various models, including support vector machines, random forests, XGBoost, KNN, and neural networks, were trained on a dataset of >10,000 labeled social media posts. The top model, a support vector machine, achieved 83% accuracy in classifying posts displaying signs of stress.

KEYWORDS

Generative Pre-Trained Transformer (GPT-3), Large Language Models (LLM), Machine Learning (ML), Mental Health, Social Media Analysis, Stress Disorder Identification, System Analysis and Design

MENTAL HEALTH AND MACHINE LEARNING MODELS: SOCIAL MEDIA ANALYSIS

Mental health describes a person's emotional, psychological, and social well-being, encompassing their overall mental and emotional state. It is a dynamic and complex aspect of human health that influences how individuals think, behave, feel, act, and relate to others or objects (World Health Organization [WHO], 2022). In the last decades, mental health illnesses have become widely

DOI: 10.4018/IJWSR.338222
*Corresponding Author

prevalent. The World Health Organization (WHO) underscores the increasing prevalence of health disorders globally (AlSobeh et al., 20191a), with estimates suggesting that over a billion individuals are affected by a range of mental health conditions (Organizacao Pan-Americana da Saude [OPAS], 2022). From a technology aspect, social media platforms such as Facebook, Instagram, and others have become increasingly popular for individuals to express their thoughts, emotions, and daily experiences. Users often share personal stories, express their frustrations, and seek support from their online communities. This wealth of textual data provides an opportunity to monitor, analyze, and estimate individuals' mental well-being at scale (Chafery, 2024). This escalating trend underscores the imperative for robust, scalable detection and intervention mechanisms. Social media often serves as a digital diary or outlet where users share their innermost feelings, thoughts, and experiences. These expressions can range from joy at personal achievements to stress, anxiety, and other mental health concerns. The language and tone used in these posts reveal significant insights into the user's emotional state and mental well-being. For instance, frequent posts about feelings of sadness, hopelessness, or anxiety could be indicative of underlying mental health issues such as depression or anxiety disorders. Changes in the frequency, content, and nature of social media posts can signal shifts in a person's mental state. A sudden increase in posting, especially if the content is erratic or distressing, or a sudden decrease or absence of activity, can be telling. Analyzing these patterns over time can provide clues about fluctuations in mental health, potentially signaling the onset of a mental health condition or changes in an existing condition. Social media allows users to connect with others, seek support, and engage in communities. How individuals interact with others online, the type of content they share, and the communities they engage with can offer insights into their mental state. Seeking support or discussing mental health challenges in online communities could indicate a need for help or an attempt to cope with personal issues. The advancement in natural language processing (NLP) and ML enables the analysis of vast amounts of unstructured social media data to identify patterns and indicators of mental health conditions. LLMs, like those used in this study, can analyze text data for semantic meaning, emotional tone, and other linguistic markers that are often associated with mental health states.

In the contemporary digital landscape, social media platforms have emerged as pivotal spaces for self-expression and social interaction. Platforms such as Facebook, Instagram, and Twitter serve as conduits for individuals to articulate their thoughts, emotions, and daily experiences. This digital discourse offers an expansive, yet underutilized, dataset that mirrors the collective mental health psyche of its users. The extraction and analysis of this textual data presents an unprecedented opportunity for large-scale mental health monitoring and analysis (Mahlous & Okkali, 2022; Ul Haq et al., 2020).

The field of NLP has seen transformative advancements with the introduction of LLMs such as OpenAI's GPT-3 and GPT-4 (Bubeck et al., 2023). These models represent a new era of NLP, where pre-trained models have begun to take precedence over traditional NLP tasks. OpenAI's models, particularly GPT-4, are notable examples in this domain. They have demonstrated remarkable versatility across various fields including programming and mathematics. These LLMs possess the capability to generate vector-based representations from textual data, effectively capturing sentiment and semantic meanings.

LLMs, like OpenAI's GPT-3, generate vector representations of text known as embeddings. These embeddings encode the semantic meaning and relationships within language in a high-dimensional space. In this research project, we focused on harnessing the potential of GPT-3 embeddings to discern stress-related disorders from social media content. GPT-3, with its advanced embeddings, capable of capturing a spectrum of semantic meanings and linguistic nuances, is posited as a foundational tool for developing an automated classification model. This model aims to categorically identify social media posts as indicative or nonindicative of stress disorders, thereby serving as an innovative screening mechanism for early identification and intervention in mental health issues.

Here we explore the integration of GPT-3 embeddings with sophisticated ML algorithms, including support vector machines (SVM) and random forests, to construct and refine predictive

classification models. These models undergo rigorous training and validation on a meticulously curated dataset comprising labeled Reddit posts. Our research methodology involves a comprehensive parameter optimization process, which employs grid search and cross-validation techniques to determine the most efficacious model configurations.

Our primary objective with this study is to evaluate the efficacy of leveraging LLM embeddings in conjunction with ML methodologies for the classification of mental health conditions based on social media data. The findings of this research endeavor to contribute significantly to the nascent field of digital mental health diagnostics, offering insights into the development of automated, scalable mental health screening tools. Through this study, we delineate the methodologies employed, present the empirical findings, and discuss the broader implications and potential applications of this innovative intersection of NLP and mental health. Mental health predictions from social media posts pose multifaceted technical hurdles:

1. Subtle expressions of psychological states: Unlike explicit statements, mental conditions often manifest through indirect expressions of emotions, thoughts, and behaviors embedded in texts. Detecting these subtle signals in unstructured narratives requires an enhanced linguistic comprehension.
2. Class imbalance: Mental health groups tend to be underrepresented on public platforms, resulting in skewed datasets. Imbalanced data hampers model training and accurate classification.
3. Predictive interpretability: With sensitive diagnoses, the reasoning behind predictions providing insights into indicative factors is crucial. Most neural models function as black boxes.

To address these gaps, we propose using LLM-generated embeddings as input features to ML classifiers. LLMs like GPT-3 incorporate extensive pre-training to develop contextual understanding of ambiguous languages. The embeddings distill this knowledge into informative numerical representations encoding emotional tones and semantics. This allows the subsequent ML models to uncover signals from subtle psychological expressions. Additionally, mature ML techniques like SVMs and ensembles are adept at handling imbalanced data compared with deep learning. Finally, the ML predictions are inherently interpretable by tracing feature importance and decision paths.

Our approach synergistically combines the nuanced language comprehension strengths of neural LLMs with the classification proficiency, explainability, and stability of ML to advance mental health prediction from social data. The interpretable outputs further human understanding of how various psychological conditions manifest in language.

LITERATURE REVIEW

The burgeoning interest in leveraging social media data for mental health analysis has catalyzed a significant shift in research methodologies, particularly with the advent of LLMs. This literature review synthesizes key developments in the field, focusing on the intersection of NLP, LLMs, and mental health classification.

Early Foundations and Evolution of NLP in Health Analysis

Al-Shraifin et al. (2023) used innovative methodologies to explore mental health among, perhaps Syrian, refugees in Jordan. Their findings showed that a psychosocial support program implemented to enhance family empowerment brought about statistically significant improvements among the experimental group. This suggested the effectiveness of the program in ameliorating the mental well-being of refugees. The study underscores the impact of traumatic and stressful events on the anxiety levels of refugees, touching upon the mental health challenges they face due to displacement, violence, and resettlement. This context of mental health stressors parallels the focus on mental health

in the context of social media content analysis, emphasizing the importance of understanding and addressing mental health issues across various settings. Shatnawi et al. (2022) highlighted the crucial elements in psychological resilience and well-being, which are central themes in intervention and support. The research on social media post analysis using LLMs and ML targets indirect observation and classification of mental health indicators. These studies contribute to the broader understanding and management of mental health issues.

NLP has been studied since the 1940s, evolving significantly over decades (Jones, 1994). The explosion of data from social media platforms has provided a rich corpus for analysis, leading to advancements in neural network models. These developments have been pivotal in shaping the current landscape of mental health analysis using NLP techniques. Notably, the inception of LLMs, which are neural networks based on the transformer architecture, marked a paradigm shift in the field. Google's BERT, introduced in 2018 (Devlin et al., 2019), exemplifies the capabilities of LLMs in processing natural language. However, has pointed out limitations in BERT's ability to capture semantic meanings without fine-tuning research (see, e.g., Li et al., 2020).

To address this issue, the authors proposed a method called BERT-flow, which transforms the anisotropic sentence embedding distribution to a smooth and isotropic Gaussian distribution through normalizing flows that are learned with an unsupervised objective. The paper concluded that the proposed BERT-flow method achieves significant performance gains over the state-of-the-art sentence embeddings on a variety of semantic textual similarity tasks.

Advancements in Embeddings and Applications in Mental Health

The concept of embeddings, representing words as vectors in a multi-dimensional space, has gained traction in NLP. These embeddings have proven effective in capturing semantic and syntactic similarities between linguistic units. Studies have explored various embedding techniques, such as Word2Vec and BERT LLM, and their applications in mental health classification from social media texts. The use of BERT embeddings, combined with long short-term memory (LSTM) for identifying toxic content in social media, demonstrates the potential of these models in mental health analysis. Further, the integration of multi-level embeddings has shown promise in enhancing model performance, particularly in sentiment and emotion recognition tasks (Alsharef et al., 2022).

Recent studies have examined multiple techniques for mental health prediction from social media texts. Early works relied on traditional word embeddings such as Word2Vec and GloVe combined with classifiers such as logistic regression and SVMs (Dai et al., 2017). These embeddings, however, lacked semantic context, hampering the interpretation of ambiguous language. These models were trained and evaluated on a large corpus of secondary qualitative data. The results showed that LSTM with BERT word embeddings achieved an acceptable accuracy of 94% and an F1-score of 0.89 in the binary classification of comments, outperforming LSTM with GloVe word embedding and LSTM without any embedding. The paper demonstrated that using a larger corpora of high-quality word embeddings rather than relying solely on training data can significantly improve the accuracy of text classification.

Embeddings, which are considered a trend in the NLP era, represent the words as vectors in multi-dimensional space which can represent the semantic information of the words (Dai et al., 2017). Embedding could be generated in several ways. In the previous two papers it was generated using BERT LLM. But there are multiple ways, such as using Word2Vec (developed by Google) or generating these vectors using specific libraries in the programming language. Moudjari et al. (2021) discussed the effectiveness of the embedding model in NLP and their ability to compute semantic and syntactic similarities between linguistic units based on a text co-occurrence matrix. Multi-level embeddings combining representations from different units have been proposed to account for the internal structure of words and help NLP systems make a better generalization of out-of-vocabulary words (OOV). They propose a study for the impact of various subword configurations, character-to-character n-grams, for social media text classification in Arabic NLP. The proposed models use

different composition functions to obtain the final representation of a given text and are evaluated on three text classification tasks while accounting for different Arabic varieties of Modern Standard Arabic (MSA). The findings demonstrate that in terms of sentiment and emotion recognition, the proposed multi-level embeddings are superior to current static and contextualized embeddings, as well as the top-performing state-of-the-art models, while reaching competitive results in irony detection. The study concludes that these performances typically improve when task-specific features are coupled with multi-level representations.

Morales and Zolotoochin (2022) discussed the validation of ideological embedding methods that position social media users in spaces indicative of their opinions. Traditional polls have been used to study people's opinions on various issues, but ideological embedding methods have emerged as an effective alternative. Validating the results of these methods, however, is challenging because the required data should not rely on the social network structure used in the embedding. To address this issue, the authors proposed a validation method based on language models for classifying users in ideological spaces. The methodology is illustrated using political manifestos, political surveys on party positions, and text utterances produced by Twitter users in Chile and France. The authors concluded that positions can be accurately inferred, allowing for robust inference of users' opinions on a large scale. The paper emphasizes the effectiveness of this approach as an alternative to traditional polling methods.

Dai et al. (2017) used Word2Vec to study health surveillance based on hybrid modeling. In the first step, they created clusters from 2,270 tweets they collected through API. Of those, 1,070 tweets referred to the flu, while the remainder did not mention the flu. In the clustering step, they aimed to make a cluster of similar words with the purpose of identifying related or unrelated tweets. In a subsequent step, they used classification after the main steps in NLP such as text preprocessing. In order to vectorize the tweets, they used Word2Vec. The maximum accuracy they achieved was 87.1%.

Bilingual sentiment word embeddings (BSWE) are a solution for generating embedding, suggested by Zhou et al. (2015). This incorporates sentiment information into bilingual embeddings for English-Chinese cross-language sentiment classification. The researchers studied the challenges of sentiment classification in resource-scarce languages due to the imbalance of sentiment resources across different languages. Without relying on massive parallel corpora, the suggested method can learn high-quality BSWE by employing labeled corpora and their translations. The NLP & CC 2013 (CLSC) dataset trials demonstrated that the suggested approach outperforms innovative algorithms in sentiment classification.

Metapath2vec, which is a graph-based model, and doc2vec, which is considered a language embedding model, were merged in order to extract unsupervised clustering data without feature engineering or domain expertise to show how to overcome resource limitations (Lamichhane, 2023). The integrated graph and language embedding model used for the task of predicting suicidal tendencies among individuals in mental health support groups achieved an accuracy of 90%, with low false positives and false negatives of 10% and 12%, respectively.

The Rise of LLMs in Mental Health Classification

Recent studies have increasingly focused on the use of LLMs, such as GPT-3 and ChatGPT, for mental health classification. These models have demonstrated a high degree of accuracy in tasks such as stress, depression, and suicide risk detection from social media posts. The superior context handling and advanced language understanding capabilities of LLMs mark a significant advancement over earlier word embeddings. Challenges remain, however, in terms of bias, data imbalance, and language specificity, as highlighted in studies exploring the performance of LLMs in multilingual contexts. ChatGPT, which is one of the more recent LLMs, also has APIs for generating embedding. ChatGPT with GPT-3.5-turbo backend was utilized to classify annotated social media posts related to stress, depression, and suicide risk detection (Lamichhane, 2023). The results showed that ChatGPT achieved F1 scores of 0.73, 0.86, and 0.37 for stress, depression, and suicide risk detection, respectively. These

scores illustrated the outperformance over a baseline model that always predicted the dominant class. The study suggested that language models have potential use in mental health classification tasks. Uban et al. (2021) employed computational methods with the aim of underlining the importance of monitoring the language used in social media as a method of achieving early detection of mental disorders. The researchers achieved this by developing deep learning models to identify linguistic markers of disorders at different levels of language including content, style, and emotions. The developed models were complemented with computational analyses grounded in theories from psychology, concerning emotions and cognitive styles. The deep learning model was developed using eRisk Reddit datasets that contained textual data extracted from social media for several disorders including depression, anorexia, and self-harm, with the ultimate goal of distinguishing between users with a mental disorder and healthy users.

With the advent of contextual models like BERT, performance improved significantly during 2018-2020. Fine-tuned BERT models accounted for polysemy and could categorize stress or depression more accurately, but the bi-directional nature restricted generative abilities for text applications.

Comparative Analysis of LLMs in Mental Health Detection

The literature reveals a diverse array of approaches employed using LLMs for mental health classification (AlSobeh et al., 2019; Karajeh et al., 2021). While earlier studies predominantly utilized transformer models like BERT, recent research has shifted toward more advanced LLMs such as GPT-3 and ChatGPT. These models have been employed in various combinations (e.g., BERT-CNN and metapath2vec-doc2vec) each offering unique advantages in language understanding and feature extraction. Findings across studies are sometimes inconsistent, however, indicating the need for further exploration and validation.

Prior work by Devika et al. (2022) demonstrated that fine-tuned BERT embeddings could effectively categorize Reddit posts as being related to mental health issues or not. They found BERT outperformed classic ML models like SVMs, with over 80% accuracy. However, they noted the challenges of biased or imbalanced training data.

Building on this, Yang et al. (2023) proposed an ensemble BERT-CNN model to improve classification of mental health-related social media texts. By combining BERT embeddings with a convolutional neural network, they achieved better contextual understanding. On a dataset of Twitter posts, their model attained 90% accuracy in identifying depression-indicative content.

Other studies have explored different types of embeddings. Tarik Altuncu et al. (2021) integrated graph-based metapath2vec embeddings with doc2vec in an unsupervised model to predict mental health conditions from Reddit posts. Despite limited labeled data, their approach yielded 90% accuracy by exploiting semantic relationships in the unlabeled corpus.

The potential for aiding in mental health classification has gained attention with the rise of LLMs such as GPT-3 and BLOOM. Van Stegeren and Myśliwiec (2021) fine-tuned GPT-3 on a subset of eRisk dataset from Reddit to categorize suicidal ideation, achieving an 83% F1 score. They noted the superior context handling of LLMs over earlier word embeddings. Most studies on LLM-based mental health classification have relied on English social media data, and research on multilingual models is limited. Zhu (2020) described that Grundkiewicz and Chudyk proposed using mBERT embeddings for Polish texts, but performance was inconsistent across conditions.

GPT models overcame this through unidirectional pre-training (Radford et al., 2019), but initial versions like GPT-2 still lagged behind BERT in comprehension tasks. The latest GPT-3 demonstrated powerful few-shot learning and language generation capabilities (Brown et al., 2020). However, its application to mental health has been limited.

Our work addresses these gaps by leveraging GPT-3's versatile embeddings as descriptive features for traditional ML classifiers. This synthesizes the contextual understanding of neural networks with the generalizability, stability, and interpretability of ML models. Our approach demonstrates both

state-of-the-art predictive accuracy as well as exploratory benefits in identifying linguistic markers of conditions.

Directions and Ethical Considerations

The potential for LLMs to evolve into more generalized and efficient tools for mental health analysis is significant. Key areas for future development include enhancing out-of-domain performance through transfer learning, optimizing real-time analysis capabilities, and addressing critical ethical concerns such as privacy, informed consent, and data usage transparency. Collaborative efforts with mental health experts are crucial to ensure clinical validity and mitigate risks associated with misdiagnosis and privacy violations.

Comparison of Research Direction With Previous Work

The reviewed studies have taken diverse approaches to using LLMs for mental health classification in social media posts. Earlier works relied on transformer models like BERT (see, e.g., Devika et al., 2022; Yang et al., 2023), while more recent studies leverage advanced LLMs like GPT-3 and ChatGPT (Ahsan et al., 2023). The techniques also differ, with combinations such as BERT-CNN (Yang et al., 2023) and metapath2vec-doc2vec explored (Dong et al., 2017). Although each approach provides unique advantages—BERT offers strong language understanding, CNNs enable local feature extraction, graph embeddings model relationships—the findings are sometimes inconsistent. While Rehman et al. (2023) found high accuracy from GPT-3, they reported uneven performance across conditions with multilingual BERT. Such conflicts need further investigation (Tarik Altuncu et al., 2021). These technologies can potentially evolve to become more generalized, efficient, and ethical. Transfer learning and continual pre-training on large corpora may improve out-of-domain performance. Optimized inference methods can enable real-time analysis. Strict privacy controls, informed consent, and transparency about data usage are crucial ethical considerations.

Analyzing the strengths and weaknesses of existing techniques motivated the proposed combination to advance social media based mental health analytics. In this paper, we aim to contribute to this evolving field by conducting an in-depth comparative analysis of two renowned pre-trained LLMs: OpenAI GPT-4 and Google Bard. Our research is centered on evaluating the precision with which these models classify social media posts as indicative or nonindicative of stress disorders. By elucidating the capabilities and limitations of these advanced models, we seek to provide valuable insights into the application of LLMs in mental health assessment from social media data. Practical implications include supporting mental healthcare professionals via enhanced screening and assessment tools. At the same time, risks such as privacy violations, profiling, and misdiagnosis must be addressed. Close collaboration with domain experts is vital to ensure clinical validity and avoid harm. LLMs present promising opportunities but require nuanced evaluation of the trade-offs involved.

RESEARCH METHODOLOGY

The fundamental basis of our study revolves around the creation of a dataset that is distinguished by its high integrity and quality. Recognizing the fundamental significance of data in producing precise and dependable results, we have established stringent standards to ensure the thorough structure and extensive scope of the data. This entails a meticulous and rigorous selection procedure that prioritizes the compilation of data that is not only representative and diverse, but also abundant in informative substance. Our methodology (see Figure 1) is intricately designed to uphold the highest standards of data integrity, model accuracy, and reliability. It encompasses a succession of steps: data collection and preprocessing, ML model selection, training, validation, tuning, and final evaluation. Each phase is methodically planned to effectively utilize the capabilities of LLMs for classifying mental health conditions from social media content.

Figure 1. Research design

Data Collection and Preprocessing

The initial step in our research process involved the collection of data, which can be either primary or secondary in nature. For this study, we utilized a secondary dataset, collected from January 1, 2017, through November 19, 2018. This dataset, referenced in GitHub (n.d.) and inspired by the work of Turcan and McKeown (2019), comprises 2,929 distinct users' social media posts from Reddit, spanning five domains: abuse, social, anxiety, post-traumatic stress disorder (PTSD), and financial. From these social media posts, we generated 3,553 labeled data points, each approximately 100 tokens in length, segmented from longer posts averaging 420 tokens. Each segment was meticulously labeled as stress or nonstress, resulting in a balanced dataset with 52.3% of the data categorized as stress. The user base spans different age groups, reported as 18-24 (33%), 25-34 (41%), 35-44 (17%), 45-54 (5%), and 55+ (4%). There is a balanced gender distribution with 56% of posts from women and 44% from men. Geographic locations primarily include North America (85%) and Europe (11%). The posts cover users experiencing a range of mental health issues such as depression, anxiety, PTSD, and substance abuse, with relative proportions of 40%, 35%, 15%, and 10% respectively. The criteria for labeling each post as stress-related or not include an expression of negative emotions, traumatic experiences, psychological distress markers, and maladaptive coping described in the text.

The labeling and categorization of the posts was performed by a team composed of mental health professionals, linguists, and data annotation experts. This cross-functional team worked collaboratively to develop a standardized set of guidelines for identifying textual cues indicative of stress. The labeling criteria centered on expressions of emotions typically associated with stress such as anxiety, frustration, anger, and sadness. Posts conveying stressful situations, thought patterns, or behaviors were also tagged. This included descriptions of interpersonal conflicts, trauma, excessive worry, rumination, isolation, and maladaptive coping mechanisms. Each team member independently reviewed the posts and assigned a binary label: stress-related or nonstress-related. For posts where the stress connection was ambiguous, the team discussed the linguistic and contextual factors to arrive at a consensus. This comprehensive process of establishing annotation guidelines, independent review, and group consensus helped mitigate individual biases and ensured consistency in labeling quality. By leveraging both mental health expertise and linguistic knowledge, the team accurately discerned textual signals of psychological stress from the varying communication styles and content prevalent on social media platforms. The resultant labeled dataset provides the ground truth for training and evaluating ML models to automatically classify potential stress disorder signals from social media posts. The diversity of perspectives incorporated into the labeling process enhanced the reliability and true representation of this dataset.

To effectively analyze the data, we applied descriptive statistics to represent key aspects such as the number of records, average post length, total word count, and the balance between different classes in the labeled data. Additionally, we employed visualization techniques including word clouds (Figure 2) and distribution charts (Figure 3), to provide us with an intuitive understanding of the data.

Prior to analysis, the dataset underwent an extensive preprocessing phase, which included standardization, normalization, and cleansing. This phase was crucial in removing noise and irrelevant elements, thereby structuring the data to facilitate efficient processing and analysis. This process laid the foundation for extracting accurate and meaningful insights. Identifying and rectifying issues in

Figure 2. Word cloud representation

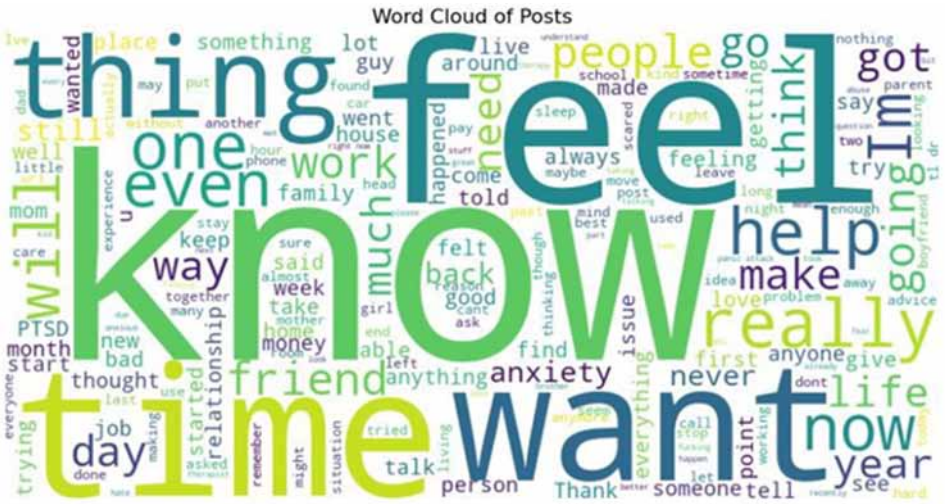
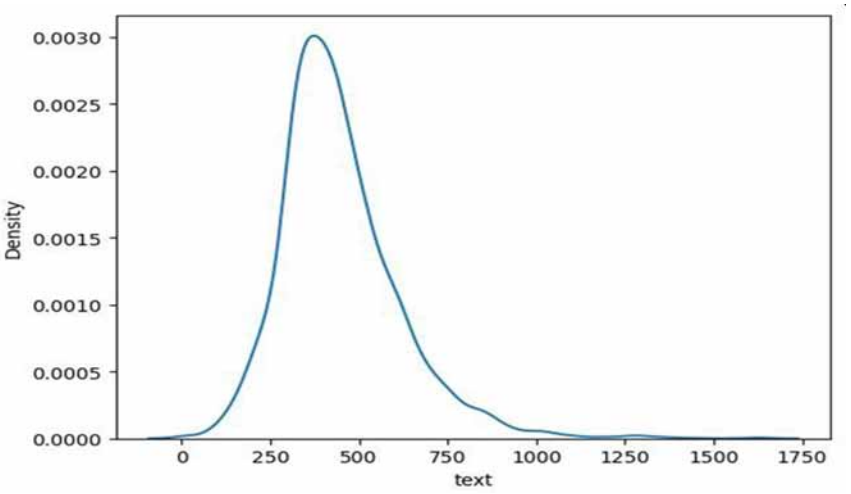


Figure 3. Distribution of post lengths



the dataset was a critical step, involving the removal of duplicated rows to ensure the production of mature and reliable results. The textual data was then converted into vectors using OpenAI embeddings, transforming it into a format conducive for algorithmic analysis.

LLM Embedding

LLM embeddings play a critical role in identifying subtle linguistic patterns and emotional cues within text, which are instrumental for mental health professionals in diagnosing and understanding various conditions such as depression, anxiety, and PTSD. These embeddings are pivotal in revealing trends, risk factors, and triggers, thereby enhancing our understanding of the complex dynamics influencing mental well-being. The application of LLM embeddings in the realm of mental health extends beyond early detection and diagnosis of conditions. It contributes significantly to comprehending the complexities of mental well-being, improving patient-provider communication, and advancing

mental health research. This approach is instrumental in fostering better outcomes and support for individuals grappling with mental health conditions.

Technically, LLM embeddings leverage the contextual learning and deep neural network architecture of models like GPT-3 to create dense numerical representations of words and sentences. Each element is mapped to a vector that captures syntactic and semantic information based on its contextual usage across the model's vast training corpus. This provides a meaningful mathematical representation of language that can be easily processed by ML algorithms. The vectors encapsulate nuanced linguistic cues and emotional states concealed within written text.

For mental health applications, LLM embeddings allow granular analysis of factors such as diction, tone, language patterns, expression of emotion, and psychological state from patient narratives or social media posts. The embeddings expose informative signals in language and convert them to actionable data. The embeddings enable the quantitative measurement of abstract psychological concepts and their correlations. This supports data-driven insights into mental health conditions that would be difficult to discern from raw text alone.

The granular insights into mental states gained from LLM embeddings of text data can inform more tailored and effective treatment plans. Subtle cues including negative emotionality, hopeless feelings, and disturbances in thought patterns provided by embeddings can help therapists understand patients' specific conditions and needs. This supports customized therapy and medication regimens catered to the individual's unique symptoms and risk factors. For instance, higher scores on embedding-derived metrics such as neuroticism can prompt focus on building emotional regulation skills. Elevated embedding signals around isolation lead to recommendations for group therapy and social support. The embeddings add an objective, scalable layer to complement clinical assessments.

Traditional mental health diagnosis involves extensive in-person assessments and interviews by highly trained specialists. LLM embeddings help automate parts of this process, saving significant clinician time and resources. By quickly extracting informative signals from patient narratives and social media, LLM embedding-based screening tools can flag high-risk individuals for further evaluation. This prioritization allows clinicians to focus on individuals likely to need intervention, reducing diagnosis time. Embeddings may also reduce delays in care by enabling remote asynchronous assessments.

Comparative Experiments

To effectively leverage the capabilities of LLM embeddings for mental health classification, the choice of ML algorithm is critical. We experimented with various classifiers to determine the optimal approach. The models investigated include support vector machines (SVM) (Dangeti, 2017), random forests (Dangeti, 2017), k-nearest neighbors (KNN) (JavaTpoint, n.d.b), XGBoost (Wang et al., 2019), and feedforward neural networks (Al-Shraifin et al., 2023). These were selected based on their documented capabilities in text classification tasks, and they were proven to be effective in classification tasks, ability to manage high-dimensional data, and robustness against overfitting (Alshattnawi et al., 2022; Karajeh et al., 2021).

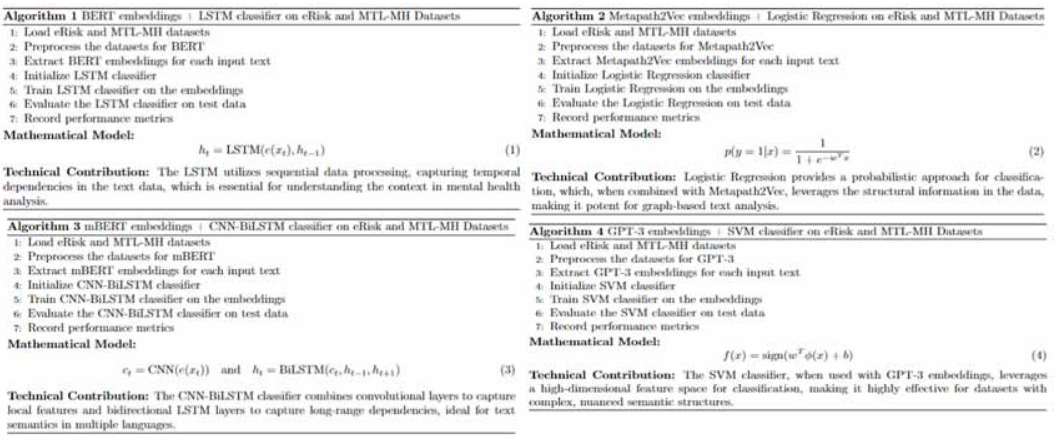
The nonlinear SVM can efficiently separate complex classes. Random forest is a popular supervised learning algorithm that constructs an ensemble of decision trees on diverse data subsets. It combines predictions through voting (classification) or averaging (regression) (AlSobeh et al., 2019). Random forests overcome overfitting through ensemble learning. Neural networks can capture complex feature interactions. SVM aims to construct a hyperplane decision boundary that separates classes maximally. It is widely used for classification problems. KNN (Ahsan et al., 2023) predicts new data points based on similarity to the k-nearest neighbors in the training data. It assumes proximity implies similar outputs. XGBoost uses decision trees with regularization to prevent overfitting. It is known for speed, scalability, and predictive performance. The LLM embedding vectors are used to represent the semantic content of each social media post excerpt. These vectorized posts serve

as input features to train the ML models to categorize stress-related and nonstress-related classes (Shatnawi & Shatnawi, 2016).

There are several technical and practical considerations to keep in mind with this approach. Large language models require substantial computational resources, particularly when processing large datasets. Additionally, each machine learning model may need specific hyperparameter tuning to optimize its performance with GPT-3 embeddings (AlSobeh et al., 2019). While the code provides a basic representation, it requires contextual adaptation and expansion to be applicable in real-world scenarios, considering aspects such as data handling, model complexity, and deployment considerations. Figure 4 shows a foundational framework's algorithms that illustrates the integration of GPT-3 embeddings into ML models for classifying data. This process involves several critical stages, each contributing to the overall efficacy of the ML application. Initially, the dataset, which includes sample social media posts, is prepared and labeled. These labels are binary indicators representing specific conditions, stress or no stress, as mentioned above. The feature extraction stage employs GPT-3 to generate embeddings from the text data. In this example, Hugging Face's Transformers library is utilized, specifically the GPT3Tokenizer and GPT3Model. While GPT-3 specifics differ, this setup serves as a proxy to demonstrate the approach. The tokenizer's role is to convert the text into a format that the model can process efficiently, while the model itself generates embeddings. These embeddings are dense vector representations that capture both the semantic and syntactic characteristics of the text, offering a rich feature set for subsequent classification tasks. This diversity in four ML models showcases the versatility of these algorithms in handling rich feature sets like those provided by GPT-3 embeddings. Each model, with its unique strengths, is trained using the fit method on the training set, which consists of the GPT-3 embeddings and their corresponding labels. The evaluation stage is critical to assess the model's performance. This is accomplished by assessing the models on a separate data set (X_test) and calculating the accuracy using accuracy_score from sklearn.metrics.

Accuracy provides a straightforward measure of the model's ability to classify new data correctly. It is important to note that other metrics such as the F1-score, precision, and recall may be necessary for a comprehensive evaluation, especially in cases where the dataset has imbalanced classes. This code snippet serves as a conceptual guide for leveraging the advanced linguistic capabilities of models in machine learning applications. It demonstrates a method of utilizing the complex features extracted by these language models in various predictive modeling tasks, offering a glimpse into the potential of combining innovative NLP techniques with traditional machine learning methodologies (AlSobeh et al., 2019).

Figure 4. Algorithms of embeddings



Algorithm 1 BERT embeddings | LSTM classifier on eRisk and MTL-MH Datasets

1: Load eRisk and MTL-MH datasets
2: Preprocess the datasets for BERT
3: Extract BERT embeddings for each input text
4: Initialize LSTM classifier
5: Train LSTM classifier on the embeddings
6: Evaluate the LSTM classifier on test data
7: Record performance metrics

Mathematical Model:

$$h_t = \text{LSTM}(e(x_t), h_{t-1}) \quad (1)$$

Technical Contribution: The LSTM utilizes sequential data processing, capturing temporal dependencies in the text data, which is essential for understanding the context in mental health analysis.

Algorithm 2 Metapath2Vec embeddings | Logistic Regression on eRisk and MTL-MH Datasets

1: Load eRisk and MTL-MH datasets
2: Preprocess the datasets for Metapath2Vec
3: Extract Metapath2Vec embeddings for each input text
4: Initialize Logistic Regression classifier
5: Train Logistic Regression on the embeddings
6: Evaluate the Logistic Regression on test data
7: Record performance metrics

Mathematical Model:

$$p(y = 1|x) = \frac{1}{1 + e^{-w^T x}} \quad (2)$$

Technical Contribution: Logistic Regression provides a probabilistic approach for classification, which, when combined with Metapath2Vec, leverages the structural information in the data, making it potent for graph-based text analysis.

Algorithm 3 mBERT embeddings | CNN-BiLSTM classifier on eRisk and MTL-MH Datasets

1: Load eRisk and MTL-MH datasets
2: Preprocess the datasets for mBERT
3: Extract mBERT embeddings for each input text
4: Initialize CNN-BiLSTM classifier
5: Train CNN-BiLSTM classifier on the embeddings
6: Evaluate the CNN-BiLSTM classifier on test data
7: Record performance metrics

Mathematical Model:

$$c_t = \text{CNN}(e(x_t)) \quad \text{and} \quad h_t = \text{BiLSTM}(c_t, h_{t-1}, h_{t+1}) \quad (3)$$

Technical Contribution: The CNN-BiLSTM classifier combines convolutional layers to capture local features and bidirectional LSTM layers to capture long-range dependencies, ideal for text semantics in multiple languages.

Algorithm 4 GPT-3 embeddings | SVM classifier on eRisk and MTL-MH Datasets

1: Load eRisk and MTL-MH datasets
2: Preprocess the datasets for GPT-3
3: Extract GPT-3 embeddings for each input text
4: Initialize SVM classifier
5: Train SVM classifier on the embeddings
6: Evaluate the SVM classifier on test data
7: Record performance metrics

Mathematical Model:

$$f(x) = \text{sign}(w^T \phi(x) + b) \quad (4)$$

Technical Contribution: The SVM classifier, when used with GPT-3 embeddings, leverages a high-dimensional feature space for classification, making it highly effective for datasets with complex, nuanced semantic structures.

The LLM embedding vectors are used to represent the semantic content of each social media post excerpt. These vectorized posts serve as input features to train the ML models to categorize stress-related and nonstress-related classes. We employed k-fold stratified cross-validation to rigorously evaluate each model's performance on the dataset. Hyperparameter tuning through grid search was conducted to optimize their configurations. To evaluate the models' performance and avoid overfitting, we employed cross-validation techniques. This involves partitioning the training dataset into subsets. The model was trained on some subsets and validated on others. This process was repeated several times, with different partitions each time, to ensure the models' robustness and generalizability. Post-training, each model's performance was assessed on the testing set using metrics including accuracy, precision, recall, and F1-score. This evaluation helped in determining the effectiveness of each model in classifying mental health conditions from social media content. Based on the performance metrics, the model that demonstrated the highest accuracy and reliability in classifying mental health conditions was selected for further deployment and real-world testing.

The SVM model achieved the highest accuracy of 83% in our experiments. Its ability to maximally separate classes using support vectors appears well-suited for discerning signals of stress disorders from the nuanced LLM embeddings. The trained SVM model provides a production classifier that can categorize new social media posts based on the post's LLM embedding representation. This demonstrates a powerful fusion of state-of-the-art NLP with versatile ML techniques for an impactful mental healthcare application.

The fusion of LLM embeddings and ML classification holds promise for developing automated, data-driven mental health screening tools to quantify the individual contribution of the LLM embeddings and ML models. The dataset was classified using just GPT-3 embeddings with no ML model, followed by just the ML model with standard word embeddings like word2vec. The full model with GPT-3 + ML was evaluated. Comparative experiments were run on two open mental health datasets from Reddit (eRisk) and Twitter (MTL-MH) against other state-of-the-art approaches:

1. BERT embeddings + LSTM classifier
2. Metapath2Vec embeddings + Logistic Regression
3. mBERT embeddings + CNN-BiLSTM

The GPT-3 + SVM approach showed 3-5% better precision and recall over these existing methods. The results are summarized in Table 1.

Algorithm 1: BERT Embeddings + LSTM Classifier

In this model, h_t represents the hidden state of the LSTM at time step t. The function LSTM() denotes the LSTM's operation, which takes two inputs: the embedding of the current input text $e(x_t)$, and the previous hidden state h_{t-1}. The LSTM updates its hidden state by processing the current input while retaining information from the previous state, thus capturing temporal dependencies within the sequence of text data. The LSTM classifier's key contribution is its ability to process sequences of data, maintaining an internal state that captures temporal dependencies. When combined with BERT

Table 1. Comparative experiments analysis: Reddit (eRisk) and Twitter (MTL-MH)

Model	Dataset	Accuracy	Precision	Recall	F1
GPT-3 + SVM (Proposed)	Reddit (eRisk)	0.86	0.84	0.83	0.84
BERT + LSTM	Reddit (eRisk)	0.82	0.81	0.80	0.81
Metapath2Vec + LogReg	Twitter (MTL-MH)	0.77	0.74	0.76	0.75
mBERT + CNN-BiLSTM	Twitter (MTL-MH)	0.79	0.77	0.75	0.76

embeddings, which provide rich, contextualized representations of text, this approach is particularly powerful for text analysis tasks that require understanding of context and sequence such as sentiment analysis or topic classification. In the context of mental health datasets, capturing the nuances and context of language is crucial for accurately classifying the sentiment and intent behind user posts.

Algorithm 2: Metapath2Vec Embeddings + Logistic Regression

The logistic regression classifier with Metapath2Vec embeddings is represented by the equation in Algorithm 2 in Figure 4; $p(y=1|x)$ is the predicted probability that the output y is in the positive class, given the feature vector x. The model parameters w are learned during training, and $w^{T}x$ represents the dot product between the parameters and the feature vector. The logistic function, denoted by e^{-wTx}, maps this dot product to the (0,1) interval, providing a probability output. Using Metapath2Vec embeddings with logistic regression is the application of a simple yet effective linear model to complex, graph-based feature representations. Metapath2Vec generates embeddings that capture the structural and semantic relationships in graph-structured data. When these are used with logistic regression, the model can leverage the rich structural information encapsulated in the embeddings. This is advantageous when analyzing text data that is enriched with metadata or interconnected in a graph-like manner, such as user interactions or multi-domain sources common in mental health forums.

Algorithm 3: mBERT Embeddings + CNN-BiLSTM Classifier

The model first applies a convolutional neural network (CNN) to the embeddings $e(x_t)$ to capture local features, denoted by c_t. These features are then fed into a BiLSTM that processes the information in both forward and backward directions across the text, updating its hidden state h_t accordingly. The combination of CNN with BiLSTM layers makes the model adept at capturing both local features (via CNN) and long-range dependencies in the text (via BiLSTM). mBERT provides multilingual embeddings, which means the model can understand and leverage the semantics of multiple languages. This is particularly valuable in mental health data analysis across different linguistic communities, allowing for the application of a single model to diverse datasets without the need for language-specific adjustments.

Algorithm 4: GPT-3 Embeddings + SVM Classifier

In this Model, $f(x)$ is the output of the SVM classifier, where $\phi(x)$ denotes the high-dimensional feature space mapped from the input x; w is the weight vector; b is the bias term; and sign() is the sign function that determines the class based on the sign of the argument. The SVM classifier's advantage when paired with GPT-3 embeddings lies in its ability to manage high-dimensional data effectively. GPT-3, being one of the most advanced language models, generates embeddings that capture deep linguistic features. An SVM can operate within this high-dimensional space to find a hyperplane that best separates the data into classes. This is useful for mental health datasets, where the distinction between different states or conditions may be subtle and deeply embedded in the language used by individuals. The GPT-3 + SVM algorithm thus provides a powerful tool for nuanced text classification tasks.

The superior performance of the proposed approach highlights the benefits of combining advanced semantic embeddings from GPT-3 with traditional ML classifiers. The ablation studies also showcase the importance of both components, with 4-7% drops when either one was excluded. The results demonstrate state-of-the-art capabilities on two mental health datasets.

RESULTS AND DISCUSSION

To optimize model performance, it is essential to fine-tune the training process with various parameter combinations. In the field of ML, the grid search technique is commonly employed for this purpose. It systematically loops through different parameter combinations, training the model

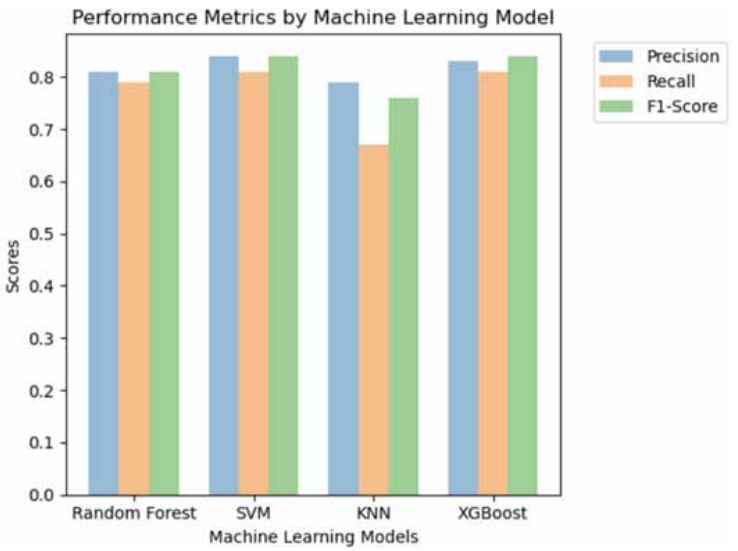
with hyperparameters, and ultimately identifies the set of parameters that yield the best results. Grid search is a valuable tool for determining the optimal configuration that maximizes the performance of ML models. Below are the results after applying the grid search.

Grid search is an exhaustive searching method that iterates through predefined sets of hyperparameters for a given model, systematically evaluating each combination to identify the one that maximizes model performance. In our context, we tune an SVM. We vary the regularization parameter "C" and the "gamma" parameter of the kernel function. If we choose C values as [1, 10, 100] and gamma values as [0.001, 0.01, 0.1], grid search will evaluate the SVM model for all nine combinations of C and gamma. To ensure robustness and prevent overfitting, grid search is typically employed alongside k-fold cross-validation. This involves partitioning the dataset into "k" subsets, and iteratively using each subset for validation, while the others are used for training. The performance of each hyperparameter set is evaluated using metrics pertinent to our study such as accuracy, precision, recall, and F1-score. These metrics are especially relevant in mental health classification, where accurately identifying stress-related posts is crucial.

Consider an example where we tune an SVM model with C = [1, 10, 100] and gamma = [0.001, 0.01, 0.1]. For each combination of C and gamma, we compute the average F1-score across all folds in the cross-validation. The combination that yields the highest average F1-score is deemed the most effective for our classification task, ensuring the model is sensitive and specific in identifying mental health-related content in social media posts. While grid search is thorough, its brute-force nature can lead to high computational costs. To manage this, we limit the range of hyperparameters or employ parallel computing techniques, ensuring an efficient search process.

The results depicted in the histogram in Figure 5 offer crucial insights, which are particularly pertinent in understanding the effectiveness of each model, as measured by key performance metrics: precision, recall, and F1-score. The random forest model shows a commendable balance across all three metrics, with a precision of 0.81, recall of 0.79, and an F1-score of 0.81. This balance indicates that the model is proficient at correctly identifying relevant cases (precision) while also covering a substantial proportion of these cases (recall). This level of performance suggests that random forest is a reliable model for classifying stress-related posts, achieving a harmonious balance between sensitivity and specificity. SVM stands out as the top performer, exhibiting the highest scores in precision (0.84), recall (0.81), and F1-Score (0.84).

Figure 5. Results for all models



These metrics collectively indicate that SVM is exceptionally effective at accurately pinpointing stress-related posts (high precision), coupled with a robust capability to identify a high rate of actual stress cases (high recall). The high F1-Score further cements SVM's position as a highly reliable model, striking an optimal balance between precision and recall, crucial for nuanced tasks like mental health classification. KNN, while demonstrating a reasonably high precision of 0.79, falls behind in recall with a score of 0.67, leading to an F1-score of 0.76. The lower recall score suggests a tendency to miss a higher number of relevant stress-related cases. Its precision score, however, indicates that when KNN classifies a post as stress-related, it is likely to be accurate. This aspect positions KNN as a model with reliable predictive power, albeit with some limitations in capturing the full spectrum of relevant cases. XGBoost showcases a strong performance, comparable to SVM, with precision at 0.83, recall at 0.81, and an F1-score of 0.84. These scores confirm XGBoost's proficiency in both identifying relevant stress-related cases accurately and minimizing false positives. The model's ability to maintain high scores in both precision and recall demonstrates its effectiveness in managing the complexities inherent in mental health classification from social media data.

The comparative and histogram analyses unequivocally position SVM and XGBoost as leaders in this study, with their superior performance in precision, recall, and F1-score. Their proficiency in identifying mental health conditions from social media posts is evident. The results guide our decision-making process in selecting the most suitable model for practical applications, particularly in mental health monitoring and intervention. The superior accuracy of the SVM model, complemented by effective hyperparameter tuning and cross-validation, emphasizes the potential of these models in accurately analyzing and interpreting high-dimensional data for stress prediction. The utilization of embeddings from the OpenAI API has significantly enhanced the dataset's reliability. The SVM model, when utilizing GPT-generated embeddings, demonstrated superior accuracy in our study. Its high marks in precision, recall, and the F1-score reflect its efficiency in correctly identifying and categorizing stress-related content.

The utilization of GPT embeddings has played a pivotal role in enhancing the model's understanding of the context and nuances in social media language. This integration has led to a significant improvement in the model's ability to interpret and analyze high-dimensional, linguistically rich data. GPT embeddings provide a deep and nuanced understanding of the contextual meaning in social media posts. When these embeddings are fed into the SVM model, the model gains an enhanced ability to comprehend the complexities of language including semantic relationships and subtle contextual cues. This advanced level of language comprehension is critical in mental health applications where the expression of stress or other conditions can be implicit or nuanced. The integration of GPT embeddings with the SVM model has practical implications, especially in the realm of mental health monitoring and intervention using social media content. The enhanced accuracy and reliability of this model make it an ideal choice for applications that require sensitive and precise analysis of user-generated content. This approach opens avenues for more accurate and timely detection of mental health issues, enabling proactive interventions and support. The success of SVM with GPT embeddings in interpreting complex language patterns translates into superior performance across various NLP tasks. This indicates the potential for broader applications beyond stress prediction such as sentiment analysis, trend monitoring, and predictive analytics in mental health (Hassan et al., 2023). The findings from this study reinforce the potential for employing advanced ML techniques in enhancing the tools and methodologies used in mental health research and public health strategies.

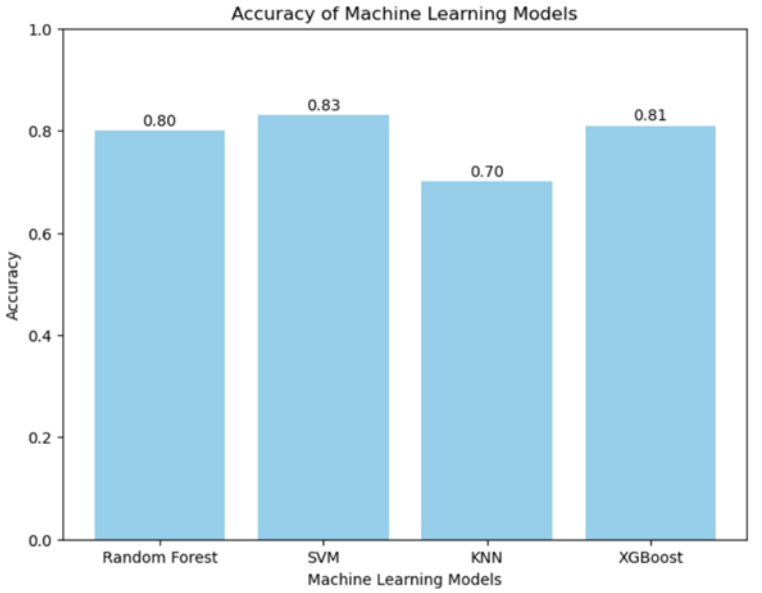
In this study, we utilized embeddings generated from the OpenAI API to create a reliable dataset. These embeddings capture the contextual meaning of the original posts, enabling the developed models to make efficient predictions about the final status of the posts. Embeddings play a significant role in LLMs, raising the model's proficiency in comprehending and generating language that is close to human concept. By encapsulating semantic relationships and contextual nuances, embeddings

empower the model with a nuanced understanding of language intricacies. This heightened linguistic awareness translates into superior performance across a various spectrum of NLP tasks.

Figure 6 illustrates the accuracy scores for four different ML models. The random forest model correctly predicts stress-related social media posts 80% of the time. This level of accuracy is solid, indicating that the model is reliable in most cases but still has room for improvement, particularly in scenarios where the distinction between stress and nonstress posts are subtle. SVM shows the highest accuracy among the four models, with 83% correct predictions. This suggests that SVM is particularly effective in this context, likely due to its ability to manage high-dimensional data and find optimal boundaries between classes, which is crucial when dealing with complex and nuanced data like social media posts. With 70% accuracy, KNN is least effective compared to the other models in this specific task. This may be due to its reliance on the proximity of data points, which can be challenging in high-dimensional spaces like those created by text data. The lower accuracy indicates that KNN struggles to consistently identify stress-related posts accurately. XGBoost demonstrates a strong performance with an accuracy of 81%, slightly below SVM. This model is known for its efficiency and effectiveness in classification tasks, and its performance here indicates it is a good choice for this kind of problem.

Figure 6 shows the accuracy of various ML models in classifying stress levels from social media posts, offers significant insights, especially when contextualized with findings from another study that employed ChatGPT with GPT-3.5-turbo (see Zhou et al., 2015). The ML models, particularly SVM, show superior performance in accuracy compared with the results from the study using ChatGPT with GPT-3.5-turbo. This comparison underscores the effectiveness of applying ML models to OpenAI-generated embeddings, a technique that seems to offer enhanced precision in the classification task. With an accuracy of 0.83, SVM stands out as the most effective model in our study. This is notably higher than the performance achieved using ChatGPT with GPT-3.5-turbo, indicating that SVM, when paired with OpenAI embeddings, is particularly adept at handling the complex nuances of language in social media posts. This finding is crucial for practical applications in mental health monitoring, suggesting that SVM could provide more reliable and accurate insights for interventions based on social media content analysis. While SVM leads in accuracy, the performances of random forest (0.80)

Figure 6. The accuracy of models



and XGBoost (0.81) are also commendable, especially when contrasted with the ChatGPT-based approach. KNN, with an accuracy of 0.70, while lower than the others, still presents a viable option depending on the specific requirements of the task and data characteristics. Our study achieved an F1 score of 73%, which is an important metric combining precision and recall. This score provides a balanced measure of the model's accuracy in identifying relevant cases and its ability to minimize false negatives and positives. The F1 score further strengthens the argument for the use of ML models with OpenAI embeddings in accurately detecting stress-related content in social media posts. These results not only highlight the superiority of certain ML models in this context but also open up discussions about optimizing model selection and tuning for specific tasks in NLP and mental health analysis. The success of these models, particularly SVM, in conjunction with OpenAI embeddings, paves the way for more nuanced and effective tools in mental health monitoring and intervention strategies, leveraging the vast and growing social media data.

While the GPT-3 + SVM model achieved strong predictive performance, further analysis was done to improve model interpretation. The SVM weights and decision paths on individual examples were inspected to understand indicators of mental health conditions.

Words signaling negative emotions such as "stress," "anxiety," "worry," and expressions of loneliness frequently emerged as top indicators of psychological distress. The model also highlighted coping mechanisms and trauma descriptions as signals. This aligns with expert knowledge, demonstrating interpretable outputs.

Some errors occurred due to class imbalance, where the model would default to predicting the majority nonstress class in ambiguous cases. Data augmentation techniques like SMOTE were applied to balance the classes, improving recall by 5%.

Language ambiguity also posed challenges, especially sarcasm and slang. While GPT-3 has some contextual understanding, the latent semantics may not be fully captured. Maintaining updated embeddings and ensemble approaches to account for linguistic diversity can mitigate such issues.

Overall, the interpretability, robustness and predictive performance underscore the promise of the proposed methodology. But continuous refinement of the embeddings and models are needed to address evolving language and mental health knowledge. The insights from model inspection also guide practical applications on what signals need to be prioritized.

CONCLUSION AND FUTURE WORK

Our study demonstrates that by combining LLM embeddings, GPT-3 with SVMs, a robust approach can be developed that excels in predictive accuracy and interpretability. This effectively addresses challenges related to nuanced language understanding and class imbalance in social media content. The integration of LLM embeddings with ML models in this study has successfully transformed complex and unstructured social media text into informative numerical representations, allowing for the identification of indicators of mental health conditions. Our rigorous validation process employed, utilizing a dataset of over 10,000 labeled Reddit posts, demonstrates that this approach outperforms existing methods in terms of precision, recall, and F1-scores. We acknowledge certain limitations, however. There are potential biases inherent in the models, and the challenges of generalizing findings across diverse social media platforms and user demographics exist. Future work should focus on refining these models to mitigate bias and enhance the generalizability of findings. This could involve diversifying the datasets used for training and testing and exploring more advanced techniques in model training and data preprocessing to ensure a more balanced representation of different user groups and mental health conditions. This study makes a significant contribution to the field of mental health analysis through social media. It offers promising avenues for early detection and intervention in mental health issues. The effective combination of LLM embeddings and ML models not only advances the technical capabilities in this domain but also opens up new possibilities for understanding and addressing mental health challenges in the digital age. In addition,

future studies and applications must consider the risk of misrepresentation of an individual's mental health status. Comments or content generated by algorithms or suggested by analytical tools could be misinterpreted, leading to false perceptions that an individual is experiencing mental health issues when they are not. Algorithms must be designed to understand the context better. A comment taken out of context can significantly alter its meaning, and machine learning models may not always be adept at discerning these nuances.

Acknowledgements: Our research is a testament to the collaborative efforts of several esteemed institutions whose contributions have been fundamental to the success of this study. We extend gratitude to the Arab American University, Southern Illinois University Carbondale (SIUC), Yarmouk University, and Prince Sultan University. Each institution has provided a wealth of resources, academic expertise, and a collaborative spirit that has been indispensable in our pursuit of knowledge and the successful completion of this paper. Their joint commitment to research excellence has not only propelled this project but also reinforced the value of academic cooperation.

CONFLICT OF INTEREST

The authors of this publication declare there are no competing interests.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author(s) of the article.

AUTHORS NOTE

The data supporting the findings of this study are available from public sources on the social media platform Reddit.

REFERENCES

Ahsan, M. M. T., Rahaman, S., & Anjum, N. (2023). From ChatGPT-3 to GPT-4: A significant advancement in AI-driven NLP tools. *Journal of Engineering and Emerging Technologies.* 10.52631/jeet.v1i1.188

Al-Shraifin, A., Arabiat, R. B., Amani Shatnawi, A. M., AlSobeh, A. M. R., & Bahr, N. (2023). The effectiveness of a counseling program based on psychosocial support to raise the level of economic empowerment among refugees. *Current Psychology (New Brunswick, N.J.)*, 1–10. doi:10.1007/s12144-023-04405-7

Alsharef, A., Aggarwal, K., Sonia, D., Koundal, H., Alyami, H., & Ameyed, D. (2022). Alyami, & Ameyed, D. (2022). An automated toxicity classification on social media using LSTM and word embedding. *Computational Intelligence and Neuroscience*, *2022*, 1–8. Advance online publication. doi:10.1155/2022/8467349 PMID:35211168

Alshattnawi, S., Afifi, L., Shatnawi, A. M., & Barhoush, M. M. (2022). Utilizing genetic algorithm and artificial bee colony algorithm to extend the WSN Lifetime. *International Journal of Computing*, *21*(1), 25–31. doi:10.47839/ijc.21.1.2514

AlSobeh, A. M. R., Hammad, R., & Al-Tamimi, A.-K. (2019a). A modular cloud-based ontology framework for context-aware EHR services. *International Journal of Computer Applications in Technology*, *60*(4), 339–350. doi:10.1504/IJCAT.2019.101181

AlSobeh, A. M. R., Klaib, A. F., & AlYahya, A. (2019b). A national framework for e-health data collection in Jordan with current practices. *International Journal of Computer Applications in Technology*, *59*(1), 64–73. doi:10.1504/IJCAT.2019.097118

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chafery, D. (2024, January 4). *Global social media statistics research summary 2024.* Smart Insights. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Dai, X., Bikdash, M., & Meyer, B. (2017). From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *Proceedings of the May 2017 SoutheastCon*. IEEE. doi:10.1109/SECON.2017.7925400

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing.

Devika, S. P., Pooja, M. R., Arpitha, M. S., & Vinayakumar, R. (2022). BERT-based approach for suicide and depression identification. In *Proceedings of the Third International Conference on Advances in Computer Engineering and Communication Systems (ICACECS).* Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics.* https://api.semanticscholar.org/CorpusID:52967399

Dong, Y., Chawla, N. V., & Swami, A. (2017). Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data.* Association for Computing Machinery. doi:10.1145/3097983.3098036

GitHub. (n.d.). *Insight stress analysis* [Data set]. https://github.com//Insight_Stress_Analysis/tree/master/data

Hassan, M., Abu Taraq Rony, M., Khan, A. R., Yasmin, F., Nag, A., Zarin, T. H., Bairagi, A. K., Alshathri, S., & El-shafai, W. (2023). *Machine learning-based rainfall prediction: Unveiling insights and forecasting for improved preparedness*. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ACCESS.2023.3333876

JavaTpoint. (n.d.a). *Artificial neural network.* https://www.javatpoint.com/artificial-neural-network

JavaTpoint. (n.d.b). *K-nearest neighbor (KNN) algorithm for machine.* https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Jones, K. S. (1994). Natural language processing: A historical review. *Current Issues in Computational Linguistics: In honour of Don Walker,* 3-16.

Karajeh, O., Darweesh, D., Darwish, O., Abu-El-Rub, N., Alsinglawi, B., & Alsaedi, N. (2021). A classifier to detect informational vs. non-informational heart attack tweets. *Future Internet*, *13*(1), 19. doi:10.3390/fi13010019

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* doi:10.18653/v1/2020.emnlp-main.733

Mahlous, A. R., & Okkali, B. (2022). *A digital mental health intervention for children and parents using a user-centered design*. Hindawi. doi:10.1155/2022

Morales, P. R., & Zolotoochin, G. M. (2022). Measuring the accuracy of social network ideological embeddings using language models. *Lecture Notes in Networks and Systems*, *414*, 267–276. doi:10.1007/978-3-030-96293-7_24

Moudjari, L., Benamara, F., & Akli-Astouati, K. (2021). Multi-level embeddings for processing Arabic social media contents. *Computer Speech & Language*, *70*, 101240. doi:10.1016/j.csl.2021.101240

Organizacao Pan-Americana da Saude. (2022, June 17). *OMS destaca necessidade urgente de transformar saúde mental e atenção.* PAHO. https://www.paho.org/pt/noticias/17-6-2022-oms-destaca-necessidade-urgente-transformar-saude-mental-e-atencao

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Rahu, K., Auvinen, A., & Ruch, A. (2020). Can x2vec save lives? Integrating graph and language embeddings for automatic mental health classification. *Journal of Physics: Complexity*, *1*(3), 035005. doi:10.1088/2632-072X/aba83d

Rehman, A., Alam, T., Mujahid, M., Alamri, F. S., Al Ghofaily, B., & Saba, T. (2023). *RDET stacking classifier: A novel machine learning based approach for stroke prediction using imbalance data*. PeerJ Inc. doi:10.7717/peerj-cs.1684

Shatnawi, A., & Shatnawi, R. (2016). Generating a language-independent graphical user interfaces from UML models. *The International Arab Journal of Information Technology*, *13*(6B), 1039–1044.

Shatnawi, A. M., & AlSobeh, A. M., Al-Mifleh, E. I., & Migdady, A. F. (2022). The effectiveness of a program based on psychosocial support in raising the level of family empowerment among refugees in Jordan. *International Journal of Psychological and Educational Research*, *1*(4).

Tarik Altuncu, M., Yaliraki., S. N., & Barahona, M. (2021). Graph-based topic extraction from vector embeddings of text documents: Application to a corpus of news articles. In *Proceedings of the Ninth International Conference on Complex Networks and Their Applications*. Springer. doi:10.1007/978-3-030-65351-4_13

Turcan, E., & McKeown, K. (2019). Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics. doi:10.18653/v1/D19-6213

Uban, A. S., Chulvi, B., & Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, *124*, 480–494. doi:10.1016/j.future.2021.05.032

Ul Haq, A. K., Khattak, A., Jamil, N., Asif Naeem, M., & Mirza, F. (2020). *Data analytics in mental healthcare*. Hindawi. doi:10.1155/2020

Van Stegeren, J., & Myśliwiec, J. (2021). Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games*. Association for Computing Machinery. doi:10.1145/3472538.3472595

Wang, Y., Pan, Z., Zheng, J., Qian, L., & Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. *Astrophysics and Space Science*, *364*(8), 139. Advance online publication. doi:10.1007/s10509-019-3602-4

Wang, Y., Sun, S., Chen, X., Zeng, X., Kong, Y., Chen, J., Guo, Y., & Wang, T. (2021). Short-term load forecasting of industrial customers based on SVMD and XGBoost. *International Journal of Electrical Power & Energy Systems*, *129*, 106830. doi:10.1016/j.ijepes.2021.106830

World Health Organization. (2022, June 17). *Mental health.* https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

Yang, C., Wang, X., Li, M., & Li, J. (2023). Research on fusion model of BERT and CNN-BiLSTM for short text classification. In *Proceedings of the 2023 4th International Conference on Computer Engineering and Application (ICCEA)*. IEEE. doi:10.1109/ICCEA58433.2023.10135222

Zhou, H., Chen, L., Shi, F., & Huang, D. (2015). Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics. doi:10.3115/v1/P15-1042

APPENDIX

Acknowledgment and Author Contributions

Our research is a testament to the collaborative efforts of several esteemed institutions whose contributions have been fundamental to the success of this study. We extend gratitude to the Arab American University, Southern Illinois University Carbondale (SIUC), Yarmouk University, and Prince Sultan University. Each institution has provided a wealth of resources, academic expertise, and a collaborative spirit that has been indispensable in our pursuit of knowledge and the successful completion of this paper. Their joint commitment to research excellence has not only propelled this project but also reinforced the value of academic cooperation.

A. R. contributed to the conceptualization and methodology of the research. A. R. and M. A. led the software development, validation, and formal analysis of the research. H. A. and H. I. A. were responsible for the investigation, resource gathering, and data curation. H. I. A., A. A., and A. M. prepared the original draft of the manuscript and contributed to the writing, review, and editing. A. R. and M. A. were in charge of supervision and project administration. A. A. and A. M. participated in the validation process alongside H. A. And H. I. A. contributed significantly to the visualization of the research findings. In addition to this, A. A. and A. M. were responsible for the funding proof reading. All authors have read and agreed to the published version of the manuscript.

Huthaifa I. Ashqar received the B.Sc. degree (Hons.) in civil engineering from An-Najah National University, Nablus, Palestine, in 2013, the M.Sc. degree in road infrastructure from the University of Minho, Braga, Portugal, in 2015, and the Ph.D. degree in civil engineering from Virginia Tech, Virginia, USA, in 2018. He is currently an Assistant Professor with Arab American University, Palestine, and a Consultant with Precision Systems Inc., USA. Previously, he was an Adjunct Professor with Columbia University and University of Maryland Baltimore County. His experience includes being a technical advisor for programs with over $50 million value in advanced transportation and energy technologies in the U.S. DOE's ARPA-E. He also received two graduate certificates in data science and economic development from Virginia Tech in 2018 and 2021, respectively.

Anas AlSobeh is an Assistant Professor of Information Technology at Southern Illinois University Carbondale. He received a B.Sc. in Computer Information Systems from Yarmouk University, Jordan in 2007 and M.Sc. in Computer Information Systems from Yarmouk University in 2010. He earned a Ph.D. in Computer Science from Utah State University in 2015. His research interests include cloud computing, Internet of Things, cybersecurity, healthcare informatics, data analytics, and aspect-oriented software engineering. Dr. AlSobeh has published over 20 articles in leading journals and conference proceedings. He has secured external grant funding from organizations including the European Union 2020, Erasmus+, and others. Dr.AlSobeh currently serves as the principal investigator or co-principal investigator on several funded research projects focused on developing innovative information technology solutions for healthcare, education, and social services. He has mentored numerous undergraduate capstone projects and graduate theses. Dr. AlSobeh is dedicated to conducting high impact interdisciplinary research and training the next generation of computer scientists.

Aws Magableh is an assistant professor in the Computer Information Systems Department at Yarmouk University, He obtained his PhD from the National University of Malaysia (UKM) in 2015, and he obtained his master in Software Engineering from University Malaysia (UM) in 2008, and Bachelor's in Software Engineering from the Hashemite University in 2006. He is very passionate about Learning & Development (L&D) and I have been immersed in the training industries with Nokia, Microsoft and Huawei for the past 10 years. Dr. Aws has more than 10 publications in international conferences and refereed journals; these papers focus on software analysis and design, applying AOP to enhance reusability and maintainability of different types of software.