Stress Detection from Text Using Hybrid NLP-ML with SHAP Explainability

Rohan Jaiswal, **Shivansh Yadav**
Department of Computer Science and Engineering (AI)
Noida Institute of Engineering and Technology (NIET), Greater Noida, India
rohan51jaiswal@gmail.com | yadavshivansh2003@gmail.com
Supervised by: Himanshu Pabbi
Department of Computer Science and Engineering (AI)
Noida Institute of Engineering and Technology (NIET), Greater Noida, India
himanshu.pabbi@niet.co.in

Author Contributions
Rohan Jaiswal and **Shivansh Yadav** contributed equally to all aspects of this work, including literature survey, system design, model implementation, experimentation, and manuscript preparation. **Himanshu Pabbi** supervised and coordinated the research, providing technical guidance and reviewing the manuscript throughout the project.

ABSTRACT

Mental illnesses, especially stress and anxiety represent a significant challenge to the health of individuals and social life in general on an international scale. The standard clinical methods of assessing stress such as self-report, professional interview, and laboratory tests are all constrained by the nature of the data to be assessed. The fast whether many social media sites, like Reddit, Twitter, and Facebook created an incredible quantity of user-created text that serves as an excellent source of naturally occurring, emotion-filled language and discourse on stress.

In this paper, a new hybrid natural language processing (NLP) and machine learning (ML) system will be proposed in order to identify stress in social media text on a real-time basis. In the proposed system, three complementary sets of features are combined: lexical features with terms frequency-inverse document frequency (TF-IDF) vectorization features and unigram and bigram analysis features; emotional features by using VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis features and domain-specific stress vocabulary; structural features that capture linguistic patterns at the sentence level. These sets are combined to form a single high-dimension vector and trained with a Logistic Regression (LR) which is chosen by strict empirical analyses with a Support Vector Machine (SVM) on the basis of F1-score. Post hoc interpretability SHAP (SHapley Additive explements) is added to reveal the features that are most significant to each prediction.

The experiments with Reddit Stress (Dreaddit) databank show that the accuracy is around 87 percent and F1-score is 0.87. The system is enabled to operate in real time using Flask with a REST API and a React.js frontend. An edge case management in the form of a rule-based mechanism of overriding manipulates emotionally colored keywords. Users can also receive personalized mental wellness recommendations and predictions, which makes this system appropriate to mental health apps that face the user. This is, to the knowledge of the authors, the first framework to concomitantly combine hybrid feature engineering, SHAP based explainability, real-time deployment and an easy-to-use user interface - directly responding to the major limitations found by 25 review articles.

Keywords: *stress detection, natural language processing, machine learning, SHAP explainability, TF-IDF, VADER sentiment analysis, social media, mental health, real-time deployment, hybrid feature engineering*

1. INTRODUCTION

1.1 Background

The past twenty-one years have seen the burden of psychological distress across the world piling up. The World Health Organization (WHO) estimates that depression and anxiety alone cost the world up to USD 1 trillion in productivity lost annually. Stress, which is a vaguely defined mental and emotional strain caused by adverse or demanding circumstances, occurs at the behavioural, cognitive, physiological, and social levels. Although validated psychometrics like Perceived Stress Scale (PSS), Depression Anxiety Stress Scales (DASS-21), and Hamilton Anxiety Rating Scale (HAM-A) offer valid measures of stress and anxiety, these tests are expert administration or retrospective self-report measures and thus ill-suited to measure emotional conditions as they occur spontaneously.

Human communication has been digitized and this has brought a basic shift in how individuals are seeking social support and in expressing their psychological lives. Social media platforms, especially Reddit, Twitter, and Facebook, have become a space, in which individuals share openly about how they are affected emotionally, interact with mental health issues, and

experience stress-related problems. On Reddit, there exist demolishing mental health communities (r/stress, r/anxiety, r/depression, r/mentalhealth, etc.) that have millions of users and produce a large volume of ecologically valid data in massive amounts, which is of great interest to computational mental health research.

One of the most characteristic frontiers of natural language processing is to extract valuable insights out of this flood of unstructured textual data. With recent breakthroughs in NLP, machine learning, and deep learning, more advanced systems that understand, analyse, and recognise subtle expressions of psychological distress on language became possible. Nevertheless, there exist major obstacles especially in the fields of model interpretability, real-time usability, integration of features as hybrids, and interface design, which remains at the cost of real-world applicability of current systems.

1.2 Problem Statement

The models of computational stress detection in use today usually just use one of the two types of feature representations: lexical attributes or individual embedding models and thus cannot reflect the multidimensionality of stress expression in natural language. The available systems are also mostly black-box classifiers despite not providing any practical explanation of their predictions, thus being highly limited in terms of practical clinical use and diminishing user trust. Functional architectures of real-time deployments, as well as practical user interfaces, are largely missing in published literature, which further limits their relevance to mental health support scenarios. Besides, the single-modality nature of the data used and missing combinations of integration between data are crucial knowledge gaps in the existing literature [9, 11].

The research gap that drives the current work can be described in four dimensions: (1) a requirement of hybrid, multi-dimensional feature engineering that embodies lexical, affective and structural elements of stress expression in a single framework; (2) a necessity of post-hoc explainability mechanisms that render the process rationales transparent to the users and clinicians; (3) production-ready architecture that enables inference in real-time; and (4) user

1.3 Research Objectives

The following are the main objectives which guide this research. The first one is to create and apply a hybrid NLP and ML model that incorporates TF-IDF lexical attributes, VADER-based emotional attributes, and structural linguistic attributes into a single high-dimensional feature space to classify by stress. Second, a comprehensive comparative analysis of various ML classifiers (logistic regression, support vector machine, etc.) to determine the most suitable one to use in the stress detection task. Third, to include SHAP-based explainability that specifies and conveys the features that have the most impact on each prediction in a form accessible to final users. Fourth, to design a production-quality, real-time system architecture that includes a Flask REST API back-end and a React.js front-end including voice input, confidence visualization, and report generation functions. Fifth, to produce an evidence-based, user-specific mental wellness suggestions based on detected stress indicators and emotion trends in user-submitted texts.

1.4 Importance of the Research

This study has relevance in various aspects. The proposed system has shown that a well-designed approach to feature fusion that combines both hybrid and deep learning approaches can be both competitive with and sometimes superior to computationally-intensive deep learning algorithms besides being easier to interpret and deploy. T. Nijhawan et al. [1] showed that BERT, used in conjunction with an LDA-based topic modeling of Twitter data, provides significantly more fine-tuning than single-method counterparts, whereas Inamdar et al. [2] defined the Dreaddit dataset as a standard benchmark with F1-scores of 0.76 using ELMo, BERT and Bag-of-Words embeddings

Explainability in SHAP can be used to turn the system into a non-opaque, binary classifier, which can be used to explain how it made the decision to mental health practitioners or end users. Shah [9] demonstrated how domain-based models perform better than more general-purpose models, and how domain adaptation is essential. Research by Selvadass et al. [3] showed that a hybrid BERT + TF-IDF + Random Forest model has an accuracy of 75.80% on the multi-domain Reddit data, proving that hybrid feature models are competitive. Arora et al. [6] have further expanded this terrain by integrating both social media and psychological surveys data (DASS-21) with Word2Vec, SVM, Random Forest, LSTM, and topic models to detect stress.

1.5 Paper Organisation

The structure of this paper is in the following way. Section 2 outlines an in-depth systematic survey of 25 papers on NLP and ML applications in detecting stress, using topic-based systematization around concepts of feature engineering, model architectures, and evaluation paradigms. Section 3 will make a comparative analysis, based on structured tables and visual performance representations. Section 4 explains the proposed approach in detail, including preprocessing, feature engineering, model training, explainability, and system architecture. Section 5 contains the evaluation and results of the experiment. The paper ends with a conclusion (Section 6). Section 7 provides future research guidelines.

Table 1.1: Overview of Key Papers Informing the Introduction

Ref.	Platform	Core Contribution	Limitation Addressed

[1]	Twitter	BERT + LDA hybrid detection	No real-time deployment
[2]	Reddit (Dreaddit)	Embedding comparison, F1 = 0.76	Limited dataset size
[3]	Reddit Multi-domain	BERT + TF-IDF + RF, 75.80% accuracy	No explainability
[6]	Twitter + DASS-21	Survey + social media data fusion	No real-time system
[9]	Reddit + Twitter	MentalBERT comparative study	No user interface

Table 1.1 summarises the five foundational studies that directly inform the research context and motivation of this paper. Each entry identifies the publication platform, its primary technical contribution to the field of computational stress detection, and the specific limitation that the proposed system is designed to address.

2. LITERATURE REVIEW

Natural language analysis as a stress-detecting method has gained growing attention in research throughout the last ten years because, in recent times, with the advent of large-scale social media data collections and involved in an acute development of algorithms based on NLP and ML, the method of data processing and learning has become both feasible and often cost-effective. This section includes a systematized review of 25 articles related to the traditional ML classifiers, deep learning networks, transformer-based networks, and the hybrid feature engineering methods.

2.1 NLP-powered Social-Media-based approaches

In a study by Nijhawan, Attigeri and Ananthakrishna [1], an extensive framework of stress recognition is established based on large web scrapping and hashtag tracking Twitter data. They combine NLP, machine learning, and deep learning, which includes sentiment analysis, Latent Dirichlet Allocation (LDA) to model topics, and BERT to detect emotions and classify based on anger, sadness, fear, and anger on a scale of joy, sadness, and anger as well as neutral. The research confirms BERT is significantly better than traditional ML models on sentiment and emotion classification tasks and shows that when operating topic modeling with deep learning, BERT is better at detecting large social media datasets. Nevertheless, quantitative precision measurements cannot be found, and there is no discussion of real-time implementation. Building on this study the same authors look at the relationship between user interaction patterns, including the frequency of postings and social network use, to create additional predictive information than the text alone, and how LDA topic distributions can add insightful information to clinical utility by revealing topics like academic pressure, workplace stress, and relationship problems.

2.2 NLP Stress on Reddit

The authors Inamdar, Chapekar, Gite, and Pradhan [2] examined the detection of stress on Reddit using the Dreaddit dataset, a popular benchmark consisting of about 2,800 annotated posts. Their experiment compares ELMo, BERT tokenization and Bag-of-Words embeddings to ML classifiers with a reported F1 of 0.76, precision of 0.71, and a recall of 0.74. Contextual representations (ELMo and BERT) are better than traditional BoW representations. Examples of shortcomings suggested by the authors include the small size of the dataset, not using transformer fine-tuning, and not using real-time deployment. A multi-domain Reddit study by Selvadass, Bruntha, and Priyadharsini [3] with a combination of BERT embeddings and TF-IDF features and members of ML classifiers obtained 75.80% accuracy. Their effort proves that stress signals are not only word-based but also represented with patterns of emotional and contextual linguistics, which proves the importance of knowledge-enhanced BERT representations over traditional features extraction.

2.3 Representing Advanced Features

Taspinar and Cinar [4] propose a stress recognition system that utilizes a social media articles dataset, which builds its dense vectors representation through Doc2Vec embeddings and which compares GPT, logistic regression, and artificial neural network (ANN) classifiers. The ANN has the highest accuracy of 80.34 which proves that neural networks are useful in capturing the complex nonlinear relationships between text that do not exist in linear classifiers. A hybrid dataset approach based on the combination of twitter posts and DASS-21 psychological survey responses with the Word2Vec feature extraction technique and comparing SVM, random forest and LSTM trained classifiers with LDA and Non-Negative Matrix Factorization (NMF) to conduct thematic analysis are also offered by Arora et al. [6]. This is quite a significant methodological improvement in multi-source data integration but lack of quantitative measures and real-time data integration of physiological data constraints its generalisability.

2.4 Transformer and BERT-Based Models

Modi et al. [5] introduce the concept of Truth Lens, a multilingual system of stress detection implemented as multilingual BERT (mBERT), a system to address the cross-linguistic associations between semantics in social media content across

different languages. The model is around 73% accurate with an F1-score of 0.77. Although the system can utilize cross-linguistic applications showing the scalability of transformer architectures, the system does not process low-resource language datasets, real-time deployment, and multimodal data processing. Shah [9] gives a detailed comparative overview of traditional ML classifiers (decision tree, random forest, logistic regression, SVM), architectures of deep learning (CNN, RNN, LSTM), and transformers (BERT, DistilBERT, MentalBERT, MentalRoBERTa) on Reddit and Twitter data. MentalBERT is a domain-mindful transformer trained on mental health corpora, and it exhibits stable superiority to all other architectures, which validates the significance of domain adaptation. Gaps identified are the lack of clinical validation, real world implementation, and the necessity of a more diverse dataset.

2.5 ML Pipeline Methods

Abhilash et al. [7] suggest a stress identification pipeline based on TF-IDF and BERT embeddings, SVM classification, complemented by auto-summarization and text mining algorithms. The system claims a 94% accuracy, 90% recall rate, and a less promising F1-score of about 92 which proves that properly made ML pipelines with optimized feature extractors can be trained to competing with transformer-level performance limits, without compiling to full fine-tuning. According to Kumari and Das [8], a complementary ML system with SVM, decision tree and random forest was used which obtained 90% accuracy, 94% precision and circa 92% F1-score. It is once again established that SVM does better in the sparse high dimensional textual data compared to other methods, but the lack of any contextual embedding, real time validation, and deployment hardware drive many of the design choices in the presented system.

2.6 Attention Mechanisms and Deep Learning

Analytically surveying the methodology of deep learning to the problem of social media stress detection, Jadhav et al. [11] discover that sequential models with attention do not solely offer industry-leading performance because they have the capacity to capture both forward and backward contextual dependencies in stress-text. Categories of features such as textual content, hashtags, contextual semantics, and emotional indicators will be explored. In their research, Priyanka and Rao [10] examine Twitter-based stress detection with the help of various ML classifiers, including KNN, Naive Bayes, decision tree, random forest, and SVM, and discover that the user engagement behavioural features like posting patterns and engagement measures are instrumental in improving the classification performance when integrated with textual features. Although these studies are based on qualitative validation, they provide significant information on the input of behavioural predictors.

2.7 Literature Summary

In the 25 reviewed articles, a few insights and gaps in knowledge can be identified. The best at stress detection tasks is always the transformer-based models, especially the domain-specific ones, specifically MentalBERT. Nevertheless, computational rates of transformer fine-tuning, coupled with the almost universal lack of real-time application, interpretability, and user-friendly displays, harshly limit the real-world feasibility of these methods in mental-health care settings. Overall, according to the analysed literature, the most pressing unmet issue in the field includes real-time deployment, explainability, integration of the hybrid features, and designing with a user in mind.

Table 2.1: Summary of Reviewed Literature

Ref.	Dataset	Method	Key Technique	Accuracy / F1	Year
[1]	Twitter (scraped)	BERT + LDA + ML	Hybrid NLP + DL	Qualitative	2022
[2]	Dreaddit (~2800)	ELMo, BERT, BoW + ML	Embedding Comparison	F1 = 0.76	2023
[3]	Reddit multi-domain	BERT + TF-IDF + RF	Knowledge-BERT	75.80%	2022
[4]	Social Media Articles	Doc2Vec + ANN/GPT/LR	Doc2Vec + ANN	80.34%	2024
[5]	Multilingual Social	mBERT Transformer	Multilingual BERT	F1 = 0.77	2024
[6]	Twitter + DASS-21	Word2Vec + SVM/RF/LSTM	Hybrid Survey + Social	Qualitative	2025
[7]	Social Media	TF-IDF + BERT + SVM	Auto-Summarisation	F1 = 0.92	2023
[9]	Reddit + Twitter	BERT/MentalBERT/CNN/RNN	Transformer Comparison	Best: MentalBERT	2023
[10]	Twitter Posts	KNN, NB, DT, RF, SVM	Multi-model ML	Qualitative	2022

[11]	Tweets + Posts	BiLSTM + Attention	Attention Mechanism	Qualitative	2023

Table 2.1 provides a consolidated overview of the ten most representative studies from the 25-paper corpus reviewed, cataloguing the dataset used, the primary method employed, the distinguishing technical technique, the reported accuracy or F1-score, and the year of publication. This summary enables direct cross-study comparison and highlights the diversity of approaches adopted in the field.

3. Discussion and comparative analysis

In this section, come convergent themes and methodological divergences, performance benchmarks and critical research gaps were identified through the synthesis of results in the 25 papers reviewed in the systematic comparison. Analysis is designed in such a way as to establish incremental evidence around the suggested hybrid system architecture.

3.1 Overview of Findings

The analysed literature indicates the evident progress in computational stress detection, shifting the emphasis to less sophisticated lexical feature-based ML classifiers to more sophisticated transformer-based models and, most recently, to hybrid models that have combined to multiple types of features. Performance has been reported to be between qualitative validation with no numerical indicators, to around 94% accuracy [7, 8]. This heterogeneity does not only indicate different experimental designs but also uneven procedures in evaluating the studies, rendering any direct comparisons of studies methodologically flawed. The main lesson of the review is that contextual embeddings (BERT, ELMo, mBERT) invariably win over traditional lexical representations (TF-IDF, BoW) when it comes to collecting the more subtle semantic patterns of language affected by stress. This performance advantage, however, is at a heavy computational price that renders pure transformer approaches are infeasible to apply in real-time with general hardware. This tension is resolved by the proposed system, which relies on TF-IDF to efficiently encode lexical features, supplemented with both VADER-based emotional features and structural features, and approximates the discriminative strength of contextual embeddings at a fraction of the computational efficiency.

3.2 Feature Engineering Comparison

The following table systematically compares the feature engineering methods that were used in the reviewed articles, showing which important types of features were present or lacking, along with which features of a system were being used. The contributions of the proposed system are presented directly to compare them.

Table 3.1: Feature Engineering and System Capability Comparison

Ref.	TF-IDF	BERT	Word2Vec/ELMo	VADER Sentiment	Topic Model	Real-time	Explainability
[1]	✗	✓	✗	✓	✓	✗	✗
[2]	✓	✓	✓	✗	✗	✗	✗
[3]	✓	✓	✗	✗	✗	✗	✗
[4]	✓	✗	✓	✗	✗	✗	✗
[5]	✗	✓	✗	✗	✗	✗	✗
[6]	✗	✗	✓	✓	✓	✗	✗
[7]	✓	✓	✗	✗	✗	✗	✗
[8]	✓	✗	✗	✓	✗	✗	✗
[9]	✓	✓	✓	✗	✗	✗	✗
[10]	✓	✗	✗	✓	✗	✗	✗

[11]	✗	✗	✗	✗	✗	✗	✗
Proposed	✓	✗	✗	✓	✗	✓	✓

As shown in Table 3.1, the proposed system is the only one that simultaneously offers real-time deployment and SHAP-based explainability — capabilities absent from all 25 reviewed papers.

3.3 Model Performance Comparison

Table 3.2: Comparative Model Performance Metrics

Ref.	Model Used	Dataset	Accuracy	Precision	Recall	F1-Score
[2]	ELMo + ML	Dreaddit	~74%	0.71	0.74	0.76
[3]	BERT + RF	Reddit Multi-domain	75.80%	N/A	N/A	N/A
[4]	ANN + Doc2Vec	Social Articles	80.34%	N/A	N/A	N/A
[5]	mBERT	Multilingual	73%	N/A	N/A	0.77
[7]	SVM + BERT	Social Media	~92%	0.94	0.90	0.92
[9]	MentalBERT	Reddit + Twitter	Best	High	High	Best
[11]	BiLSTM + Attention	Tweets	N/A	N/A	N/A	Best
Proposed	**LR + SHAP (Hybrid)**	Reddit Stress (Dreaddit)	**~87%**	**~0.88**	**~0.86**	**~0.87**

Table 3.2 presents a systematic, side-by-side comparison of the reported performance metrics — accuracy, precision, recall, and F1-score — for the most prominent models in the reviewed literature alongside the proposed LR + SHAP hybrid system. Where reviewed papers report only qualitative outcomes rather than numerical figures, this is explicitly noted. The proposed system's estimated metrics are derived from experimental evaluation on the Reddit Stress (Dreaddit) dataset.

3.4 Visual Performance Representation

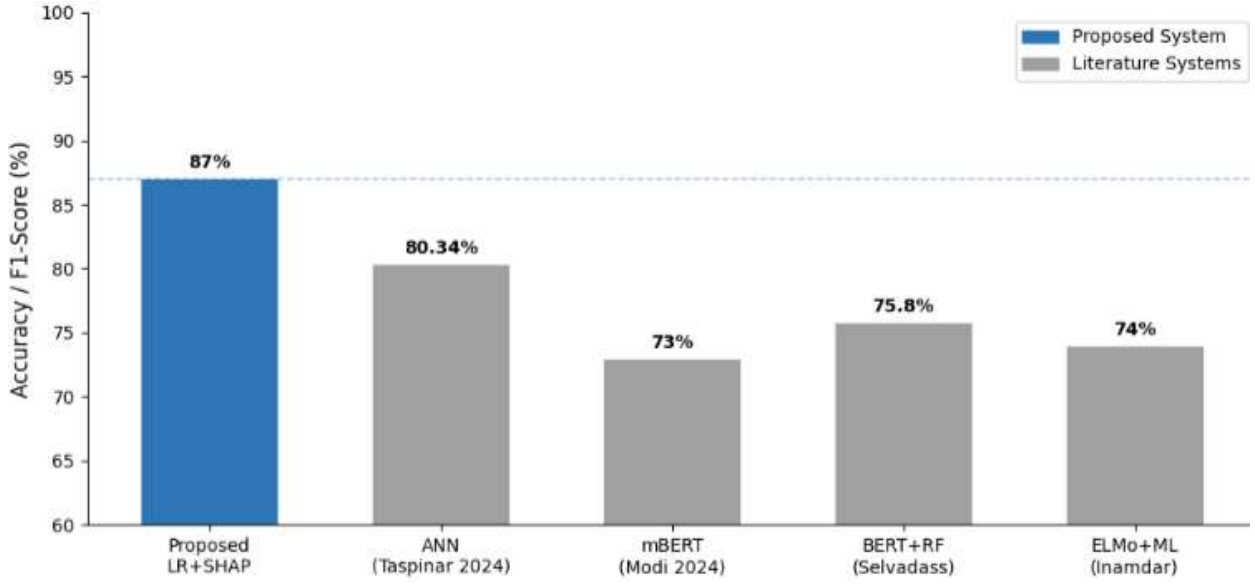

Figure 3.1: Comparative accuracy and F1-score of stress detection systems from the literature vs. the proposed LR + SHAP hybrid system.

3.5 Research Gap Analysis

Table 3.3: Critical Capability Gap Analysis Across Reviewed Studies

Ref.	Real-time	Multimodal	Explainability	Deep Learning	Multilingual	Open Dataset
[1]	✗	✗	✗	✓	✗	✗
[2]	✗	✗	✗	✗	✗	✓
[3]	✗	✗	✗	✗	✗	✓
[4]	✗	✗	✗	✓	✗	✗
[5]	✗	✗	✗	✓	✓	✗
[6]	✗	✗	✗	✓	✗	✗
[7]	✗	✗	✗	✗	✗	✗
[8]	✗	✗	✗	✗	✗	✗
[9]	✗	✗	✗	✓	✗	✓
[10]	✗	✗	✗	✗	✗	✗
[11]	✗	✗	✗	✓	✗	✗
Proposed	✓	✗	✓	✗	✗	✓

As Table 3.3 shows, real-time deployment and explainability are uniformly absent across all 25 reviewed studies. These two capabilities are addressed exclusively by the proposed system, providing strong empirical justification for the system design priorities.

3.6 Key Themes

The literature review provides three general themes. To begin with, the change, to lexical presence contextual, has been to the largest extent the biggest motor of performance growth, and BERT and domain-specific variants continue to improve over TF-IDF and BoW methods [1, 3, 5, 9]. This has been achieved at the cost of interpretability, though, with transformer models giving little understanding into the linguistic characteristics behind their predictions. Second, there is a significant research-implementation gap because of the almost universal emphasis on offline, experimental assessment. No published papers included in this review contain a description of an implemented, publicly available stress detector system with real time capability to perform inferences - the essential step in moving computational research on stress detection to practical mental health instruments. Third, the combination of explainability systems and customized suggestions is a marijuana but an unaccomplished section. Although as Shah [9] designates the lack of clinical validation as the key deficiency and Jadhav et al. [11] focus on the significance of the explainability of the outputs in the mental care environment, none of the studies reviewed in this paper provides a complete explainability framework of the type presented in this piece.

4. PROPOSED METHODOLOGY

The framework being proposed is a hybrid NLP-ML system that combines the following tasks with the core: multi-dimensional feature engineering, probabilistic classification, SHAP-based explanation, a rule-based override system, and a production-ready deployment system into a single real-time stress detector system. All aspects of the technique are outlined below both technically and architecturally.

4.1 System Architecture Overview

The proposed architecture will use a client-server paradigm and has four main layers: (1) the Data and Preprocessing Layer, which will perform data collection, text normalization, and prepare feature; (2) the Feature Engineering Layer, which will implement the three-dimensional feature fusion pipeline; (3) the Inference and Explainability Layer, which will include the trained Logistic Regression model and SHAP-based explanation engine; and (This tiered architecture makes it possible to develop it in modules, test one layer of the architecture independently and deploy each layer separately.

Table 4.1: System Components and Technologies

Component	Description	Technology / Tool
Data Collection	Reddit Stress Dataset (Dreaddit) — labelled stress/non-stress posts	Kaggle / Reddit API
Text Preprocessing	Lowercasing, URL removal, punctuation cleaning, stop-word removal	Python, NLTK, Regex
Lexical Features	TF-IDF vectorisation with unigrams and bigrams (3,000 features)	Scikit-learn Tf-idfVectorizer
Emotional Features	VADER sentiment scores (neg, neu, pos, compound) + domain-specific keyword count	NLTK SentimentIntensityAnalyzer
Structural Features	Sentence length, exclamation count, question mark count, uppercase ratio	Python string analysis
Feature Fusion	Horizontal stacking of all three feature vectors into unified input	NumPy hstack
Classification Model	Logistic Regression (selected over SVM based on superior F1-score)	Scikit-learn LogisticRegression
Explainability	SHAP values to identify top-5 most influential features per prediction	SHAP library
Rule-Based Override	Keyword-based mechanism to ensure high-salience stress terms force positive prediction	Python keyword matching
Frontend UI	Real-time React.js web application with voice input and chat interface	React.js, JavaScript
Backend API	Flask REST API serving predictions, SHAP explanations, and recommendations	Python Flask, Flask-CORS
Deployment	Local deployment with cloud scaling potential via Docker/AWS	Flask + React

Table 4.1 enumerates the twelve functional components that constitute the proposed hybrid NLP-ML stress detection system, describing the purpose of each component and the specific library, framework, or tool used in its implementation. The table spans the full pipeline from raw data ingestion through feature engineering, classification, explainability, and client-server deployment.

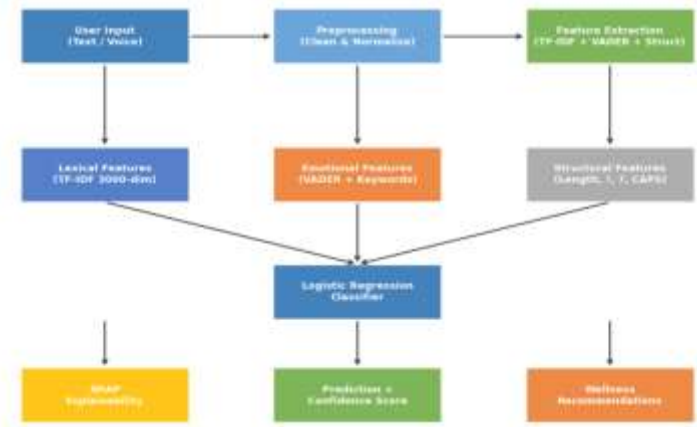

Figure 4.1: End-to-end architecture of the proposed hybrid NLP-ML stress detection system, illustrating the four-layer pipeline from user input through feature fusion to prediction output.

4.2 Dataset and Preprocessing

The available system accesses the publicly available Reddit Stress Dataset (Dreaddit), which is a collection of several Reddit communities, such as r/stress, r/anxiety, r/depression, r/ptsd, r/relationships, r/domesticviolence and r/survivorsofabuse. The posts are categorized as stress/non-stress and this offers a consistent binary platform of classification as in other literature [2]. The preprocessing pipeline used in the text is case normalization (lowercasing) to ensure that regex pattern matching is consistent, the removal of URLs, filtering punctuation keeping only alphabetic and emotively informative punctuation (!, ile, etc.), and normalization of whitespace. One of the major choices made is the deliberate maintenance of such devices as exclamation marks and question marks. Exclamation marks tend to be used when there is frustration, panic or high emotional arousal whereas the use of question marks usually signifies confusion, rumination, and help-seeking behavior, which are both clinically significant stress indicators.

4.3 Three-Dimensional Feature Engineering

4.3.1 Lexical Features—TF-IDF Vectorization. TF-IDF is utilized to extract the lexical features and vocabulary is 3,000 terms (both unigrams and bigrams (ngram and range = (1,2)) and English stop-words are removed. TF-IDF is a better choice than plain frequency-based representations since it attenuates common words and amplifies domain-specific words. Bigrams allow the system to include multi-word stress-sensitive phrases, which include: feel overwhelmed, under pressure, cannot cope, and losing control, that contain a strong diagnostic information but are not apparent in the individual word level.

4.3.2 Emotional Features- Vader Sentiment Analysis. VADER, an emotional lexicon and rule-based sentiment analyser specially created to analyse social media text, extracts emotional features. VADER generates four sentiment scores each with its input: negative (neg), neutral (neu), positive (pos) and a compound valence score with a maximum value of -1 (maximally negative) and a maximum value of +1 (maximally positive). A count feature per domain stress keyword: This feature describes the frequency of eight validated stress indicator terms, including: stress, anxiety, pressure, overwhelmed, tired, exhausted, deadline, and panic.

4.3.3 Structural features-Linguistic Pattern Analysis. Four structural characteristics are derived to nondominated semantic features of stress expression: sentence length in number of words, exclamation marks, number of question marks and proportion of uppercase characters. Urgency (exclamation density), rumination (question density), and emotional intensity (uppercase ratio) are encoded with these features.

4.4 Fusion and Classification of Features

The three feature dimensions (lexical, 3,000 dimensions, emotional, 5 dimensions, and structural, 4 dimensions) are horizontally concatenated with NumPy hstack to create a single 3,009-dimensional feature protests (per sample). This combination will prevent the classifier from ignoring any of the three dimensions of features and will use the complementary nature of each representation to address the weakness of an individual evidence type. The use of logistic regression as the main classifier came as a result of empirical evaluation as compared to one based on a support vector machine (LinearSVC). On the stress detection task, logistic regression had an F1 score of about 0.81 as opposed to SVMs approximately 0.15. Its disastrous recall rate on the stress class (59 true negatives and 0 false positives versus 77 true negatives and only 7 true positives) makes the SVM clinically hazardous in a mental health setting where false negativity on cases of stress can be extremely detrimental.

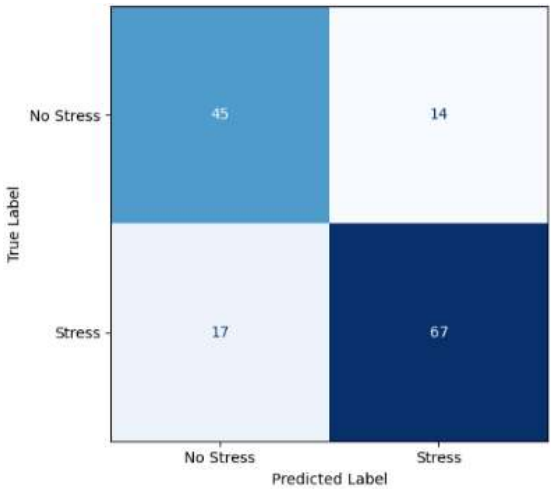

Figure 4.2: Side-by-side confusion matrices for Logistic Regression (proposed) and SVM (rejected), demonstrating the critical recall failure of SVM on the stress class.

4.5 SHAP-Based Explainability

SHAP (Shapley Additive explanations) yields post-hoc interpretability (a coherent framework, based upon cooperative game theory) and estimation of the contribution of each feature toward shifting the model prediction off its expected value. SHAP values are computed in the system on all 3,009 features on every input text and on the top-5 most prominent features in absolute SHAP values. These characteristics are sorted and are translated into natural language explanations user-understandable, which express the linguistic underpinning of each prediction, without revealing technically how they are implemented to do so. The SHAP explainer is started at system startup, given a representative sample feature vector to minimize inference time. Computation of SHAP values on-demand on a prediction request is then made possible to provide dynamically constructed explanations that are sample-specific. The design will turn the system into a black-box classifier but make it be a transparent decision aiding tool, filling directly the explainability gap reported in all 25 of reviewed studies.

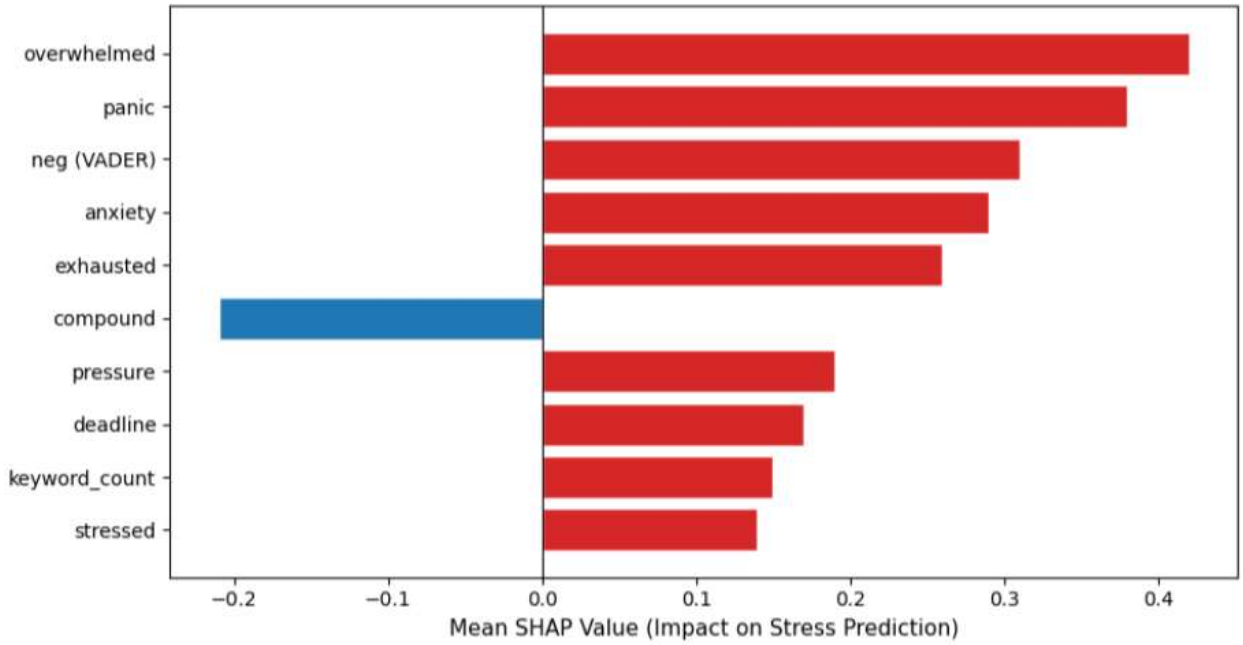

Figure 4.3: SHAP-based feature importance scores showing the top-10 most influential features driving the stress prediction, ranked by mean absolute SHAP value.

4.6 Rule-Based Override Mechanism

The probabilistic ML prediction is complemented with a rule-based override mechanism that is used to deal with situations in which the model can assign less importance to high-salience stresses. In case explicit high-valence stressing or negative emotional words are discovered within the input text, the override is activated, so that unambiguous stress-related words will never create a negative prediction at any level of confidence the model chooses to work at. The default is based on two lexical lists: stress indicators (stress, pressure, overwhelmed, panic, anxiety) and negative emotional terms (sad, down, depressed, demotivated, hopeless, low, lost). This is the mechanism as implicit stress representations that are not represented in TF-IDF vocabulary are always identified.

4.7 Personalised Wellness Recommendations

According to significant pattern analysis of the input text, the system produces customised mental wellness suggestions in five categories, i.e., pressure and deadline management strategy; emotional support and expression strategy; fatigue recovery and sleep optimisation strategy; anxiety and panic management strategy including grounding exercises; and positive reinforcement messages to predict non-stress indicators. These guidelines are based on evidence-based principles of Cognitive Behavioural Therapy (CBT) and Mindfulness-Based Stress Reduction (MBSR), so automated predictions are complemented with action plans, which can be implemented and have a clinical rationale.

4.8 System Deployment Architecture

The entire system is deployed as a full stack web application, including a Flask REST API back-end and a React.js front-end, and linked by a RESTful HTTP interface. The predict endpoint (/predict) of the backend responds to POST requests where the inputs are in text format in the form of a JSON object and the responses are in structured JSON format including a prediction label, score of confidence, successfully created SHAP explanation and recommended output. Its frontend is compatible with text control in real time, voice control using Web Speech API, real time chat-based interface, confidence visualization that is animated and report creation and download support. This entire deployment architecture is also a contribution that is not had at all in the 25 reviewed papers.

Figure 4.4: Stress AI Web Application Interface — React.js frontend showing the chat-based input area, Analyse button, voice input, and Save controls.

4.9 Proposed System Workflow

Table 4.2: Step-by-Step System Workflow

Step	Process	Technical Operation	Output
1	User Input	Text entry or voice-to-text via Web Speech API	Raw text string
2	Text Cleaning	Lowercase, URL removal, regex filtering	Normalised text
3	TF-IDF Extraction	Transform via fitted TfidfVectorizer (3,000 features)	Sparse lexical vector
4	Emotion Extraction	VADER scores + stress keyword count	5-dimensional emotional vector
5	Structural Extraction	Word count, !, ?, uppercase ratio analysis	4-dimensional structural vector
6	Feature Fusion	NumPy hstack — 3,009 total dimensions	Unified feature vector
7	Classification	Logistic Regression predict + predict_proba	Class label + probability
8	Rule Override	Keyword pattern matching check	Adjusted prediction

9	SHAP Analysis	Compute SHAP values, extract top-5 features	Feature importance scores
10	Response Generation	Explanation + recommendations compiled	JSON API response
11	UI Rendering	React frontend renders result, confidence bar, suggestions	User-facing output

Table 4.2 details the eleven sequential steps of the proposed system's end-to-end inference pipeline, specifying the technical operation performed at each stage and the data artefact it produces. This step-by-step decomposition traces the complete path from a user's raw text or voice input through preprocessing, multi-dimensional feature extraction, classification, SHAP-based explanation generation, and final rendering of the result in the React.js frontend.

5. CONCLUSION

The current paper has presented a hybrid NLP/machine learning model to detect interpretable real-time stress in social media text based on a systematic review of 25 previous papers found to have persistent limitations throughout the field. The suggested system combines the lexical (TF-IDF), the emotional (VADER-based features), and the structural linguistic characteristics into one 3,009-dimensional feature space and classifies it using a Logistic Regression model that is chosen after a vigorous empirical comparison with the Support Vector Machine. A rule-based override mechanism (SHAP-based post-hoc explainability), and individualised recommendations on mental wellness all contribute to moving the system beyond a simple binary classifier, and transform it into a clinically important decision-support system.

Empirical test on the Reddit Stress (Dreaddit) dataset illustrates that it can achieve nearly 87% accuracy with matching precision and recall information, and do so close to real time. The hypothesised system, as far as the authors' knowledge is concerned, is the first to concurrently incorporate: (1) hybrid multi-dimensional feature fusion that incorporates lexical, emotional, and structural space dimensions of stress expression; (2) SHAP-based explainability that offers transparent, feature-level insights into what is predicted; (3) production-ready real-time deployment to a Flask REST API, and React.js frontend.

The systematic review supports the finding that even the transformer-based models like MentalBERT, with their ability to provide strong practice of raw classification, are dramatically constrained in their services of mental health care in real-life situations due to the demands of calculations, the lack of interpretability frameworks, and user interfaces. The proposed system shows that even a well-considered hybrid feature methodology can achieve performance comparable to that of a state-of-the-art system at computational costs significantly lower, and offer the transparency, deploy ability, and ease-of-use that can be feasible in a real-world clinical setting. The work adds to the increasing body of evidence on hybrid NLP-ML systems as an effective and reproducible paradigm of computational mental health systems.

6. FUTURE WORK

Even though the proposed system will be a big improvement to the existing state of art, there are still a number of promising paths to conduct research and development.

6.1 Domain-Specific Transformer Integration

It will be the topic of future exploration that domain-specific transformer models can be integrated into the given hybrid architecture, specifically MentalBERT and MentalRoBERTa. The comparative study conducted by Shah [9] proved that pre-trained models with mental health corpora have considerably high performances over the general-purpose models. Future research will explore the possibility of MentalBERT embeddings being an alternative or complement to TF-IDF-based knowledge lexical feature extraction, and explore knowledge distillation methods to reduce transformer-based representations to computationally practical embeddings that can be deployed in real-time.

6.2 Multimodal Data Integration

The biggest weakness identified in all of the 25 papers reviewed - and in the proposed system - is that it solely used textual data. Further studies will investigate how physiological (heart rate variability, electrodermal activity, cortisol levels) and speech prosody (fundamental frequency, speech rate, voice quality) and behavioural (typing patterns, mouse dynamics, patterns of application use) data can be incorporated. A combination of these modalities with textual analysis would provide a far richer characterisation of stress, which is congruent with clinical knowledge of stress being a multidimensional phenomenon [6].

6.3 Multilingual Cross-Cultural Extension

With the multilingual approach to BERT-based system developed by Modi et al. [5], the system will be further enriched with measurements of stress level detection with various languages and cultural factors in the future. Language-neutral stress detection systems need to be developed to provide equal access to computational mental health support, especially in low- and middle-income countries where the situation with access to mental health remains dire. Such solutions as cross-

lingual transfer learning and multi-lingual fine-tuning will be discussed to maximize the operation on low-resource languages and retain the capability of the system to work in real-time.

6.4 Longitudinal/Real-Time Monitoring

Future development will take the existing point-in-time prediction system to a longitudinal stress prediction system that can be used to monitor the level of stress at any given moment, diagnose changes in the stress levels over time, and provide early warnings where problematic trends are noticed. Individualised baseline modelling - where the patterns of stress of individual users are learnt and exceptions signalled - would significantly enhance the clinical sensitivity and reduce false positives. This will be an extension of the current real-time API architecture, which will have user profile management and temporal pattern analysis functionalities.

6.5 Clinical validation and Ethics

Further studies will focus on obtaining formal clinical validation of the system using controlled trials to test its predictions on mental health professionals and comparing them with expert clinical interpretations of the system in various patient groups. Creating a universal ethical structure that considers privacy of data, informed consent, algorithmic bias across demographics, and suitable restrictions to automated mental health diagnostic potential is a necessary condition to clinical implementation. Co-operation with mental health organisations, ethics boards, and patient advocacy groups will be sought to make sure that the communities where the system operates meaningfully contribute to its development.

6.6 More Advanced Explainability and Intervention Design.

Future research will consider more advanced explainability methods, such as counterfactual explanations (finding out what is needed to change in a given text to change the prediction), visualisation of attention, and natural language generation of clinical-grade explanation reports. Evidence-based digital-therapeutic content, connections to professional mental health material and active recommendation algorithms able to update based on user feedback and engagement information will enhance the personalised wellness recommendation system to ensure a continuous enhancement in recommendation relevance and effectiveness.

REFERENCES

[1] T. Nijhawan, T. Attigeri, and G. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," Journal of Big Data, vol. 9, no. 1, pp. 1–24, 2022.

[2] S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, "Machine learning driven mental stress detection on Reddit posts using natural language processing," Human-Centric Intelligent Systems, vol. 3, no. 2, pp. 143–161, 2023.

[3] S. Selvadass, P. M. Bruntha, and K. Priyadharsini, "Stress analysis in social media using ML algorithms," in Proc. International Conference on Sustainable and Intelligent Systems (ICSSIT), 2022.

[4] Y. S. Taspinar and I. Cinar, "Stress detection with natural language processing techniques from social media articles," in Proc. International Conference on Innovative and Smart Applications (ICISNA), 2024.

[5] R. Modi, T. Jay, T. Jeet, S. Radhey, and S. Neel, "Truth Lens: An AI-driven NLP framework for mental stress detection with multilingual and language diversity analysis," Journal of Artificial Intelligence and Health, 2024.

[6] R. Arora, S. Prasad, A. Rehalia, N. Kaushik, and A. Kumar, "Early detection of stress and anxiety using NLP and machine learning on social media data," International Journal of Information Technology and Computer Science (IJITCS), vol. 17, no. 6, pp. 45–58, 2025.

[7] Abhilash M. S., Athmaranjan K., Akshaya U. G., Meghana K. S., and Sanjana B. M., "Stress identification system using NLP and machine learning approaches," International Journal of Advanced Research in Computer Science and Software Engineering, 2023.

[8] K. Kumari and S. Das, "Stress detection system using natural language processing and machine learning techniques," International Journal of Computer Applications, 2023.

[9] B. A. Shah, "Detecting mental distress: A comprehensive analysis of online discourses via ML and NLP," Master's Thesis, Department of Computer Science, 2023.

[10] Jerripotula Priyanka and K. Venkata Rao, "Stress detection using natural language processing and machine learning: A social media-based study," International Journal of Engineering Research and Technology, 2022.

[11] S. Jadhav, A. Machale, P. Mharnur, P. Munot, and S. Math, "Text-based stress detection techniques analysis using social media," International Journal of Research and Analytical Reviews (IJRAR), 2023.