

Detecting Stress in Social Media Posts: A Comparative Study of Machine Learning Algorithms and GPT-Based Classification

Santosh Bhandari*¹

Faculty/Lecturer, Department of Computer Science, The British College, Kathmandu, Nepal

santosh.it288@gmail.com

ARTICLE INFO

Article history:

Received 01 Oct 2026
Accepted 15 Oct 2026
Available online 24 Oct 2026

Keywords:

Stress Detection, GPT-2, Sentiment Analysis, Machine Learning, Natural Language Processing, Twitter, Social Media

Indexed in:



INDEX COPERNICUS
INTERNATIONAL



and in major libraries

ABSTRACT

Social media platforms have become a rich source of information about users' mental well-being, including stress. This study compares a fine-tuned GPT-2 classifier with four traditional machine learning (ML) algorithms—Naïve Bayes, Decision Tree, Logistic Regression and Random Forest—for detecting stress in tweets. The experiments used the mental-health-stress Twitter dataset from Kaggle, which contains 20,001 labeled posts. After pre-processing, the posts were classified as stressed or not stressed, and all models were evaluated using accuracy, precision, recall and F1-score. The GPT-2 classifier achieved the highest accuracy (88.7%) and an F1-score of 0.89, compared with 70–78% accuracy for the traditional algorithms. These results indicate that transformer-based language models capture the contextual cues of stress in short, informal text more effectively than frequency-based classifiers. Such models could support scalable, non-invasive screening tools for mental-health monitoring, although the study is limited by its single dataset, undocumented labelling procedure and single train-test split.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information,

see <https://creativecommons.org/licenses/by/4.0/>

Introduction:

Stress is a prevalent global mental health challenge, intensified by the fast pace and demands of modern life. In recent years, the widespread adoption of social media platforms such as Twitter has created new opportunities for detecting stress through digital footprints. These platforms act as informal outlets where users share their thoughts, emotions and daily experiences, including stress-related content [1].

Chronic stress has been associated with depression and suicidal ideation [2], and suicide remains a major cause of death worldwide [3]. Early identification of stress is therefore important for timely support and intervention.

Traditionally, stress has been assessed through self-reports, clinical interviews and physiological sensing, for example wearable devices that record signals such as the electrocardiogram (ECG) [4], [5]. Although effective, these approaches can be costly, intrusive and difficult to apply at scale. In contrast, social media offers a non-invasive, real-time source of information about a user's psychological state [1].

Generative Pre-trained Transformer (GPT) models are trained on large text corpora and have been applied to text generation, classification and other language-understanding

tasks because of their contextual understanding of language [6]. This study develops a stress detection system for tweets using natural language processing (NLP) and ML. A GPT-2 classifier was fine-tuned with the Hugging Face Transformers library, and its performance was compared with four commonly used classifiers implemented in scikit-learn. The motivation is to contribute towards scalable, non-invasive tools for early stress detection from real-world data.

The main contributions of this paper are: (i) a like-for-like comparison of a fine-tuned GPT-2 classifier and four traditional classifiers on a labeled Twitter dataset of 20,001 posts; (ii) an evaluation using accuracy, precision, recall and F1-score, together with an analysis of GPT-2 training behaviour; and (iii) a discussion of the limitations and ethical considerations of stress detection from social media text.

Related Work:

Machine learning and deep learning have been widely applied to stress detection using either physiological signals or text. In the physiological domain, Gedam and Paul [4] reviewed stress detection using wearable sensors and ML techniques, and Can et al. [5] surveyed smartphone- and wearable-based approaches to stress detection in daily life. Li and Liu [7] applied deep neural networks to wearable-

sensor data for stress detection, and Garg et al. [8] compared several classifiers on a public wearable-sensor dataset and reported that Random Forest outperformed the other models they tested.

In the text domain, Nijhawan et al. [1] combined ML models and a BERT classifier to detect stress and emotion in tweets and reported a very good detection rate. Inamdar et al. [9] classified Reddit posts as stressful or non-stressful using ML classifiers with several text representations, including bag-of-words, ELMo and BERT-tokenizer features. Ilias et al. [10] enhanced BERT and MentalBERT with extra-linguistic information to detect stress and depression on three public datasets. Closest to the present work, Ramteke et al. [11] compared conventional ML models with large language models, including OpenAI's GPT models, for stress detection in social media posts and found that the language models reached much higher recall (up to 99%) than the best conventional models (about 83%). The present study adds a controlled comparison in which an openly available GPT-2 model and four classical baselines are trained and evaluated on the same public Kaggle Twitter dataset and the same train-test split.

Methodology:

Fig. 1 summarises the experimental workflow. The labeled tweets were pre-processed, used to train a GPT-2 classifier and four traditional classifiers, and evaluated on held-out data, after which the models were compared.

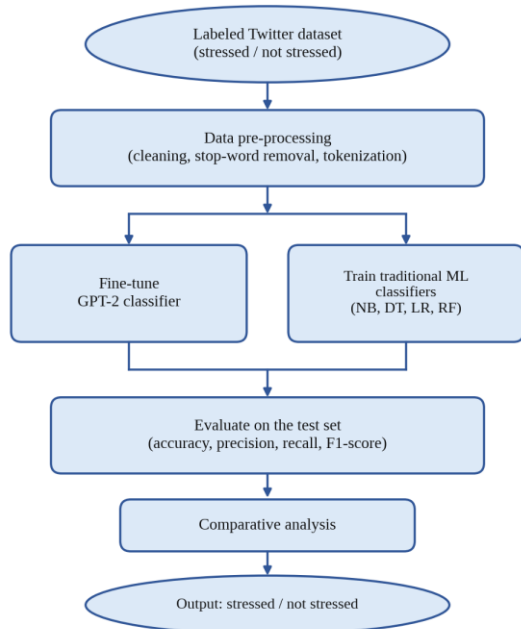


Fig 1_System Flowchart for GPT-Based Stress Detection

Dataset

The experiments used the mental-health-stress Twitter dataset obtained from Kaggle, which contains 20,001 posts. Each record includes the post identifier, creation time, tweet text, user identifier, follower, friend, favourite, status and share counts, and a binary label indicating whether the post is stressed or not stressed. Only the tweet text and the label were used in this study. The labels were used as provided with the dataset; the procedure by which they were assigned (manual or automated annotation) is not

documented in the material available to the authors, which is acknowledged as a limitation.

Data Pre-processing

The following steps were applied before model training:

- **Text cleaning:** URLs, hashtags, punctuation, numbers and special characters were removed so that only meaningful text was retained; all text was converted to lower case.
- **Stop-word removal:** common English stop words (e.g., “is”, “and”, “the”) were removed using the NLTK stop-word list.
- **Tokenization:** text was split into tokens with the WordPunct tokenizer for the traditional models and with the byte-pair encoding (BPE) tokenizer of GPT-2 for the transformer model.
- **Stemming and feature extraction:** for the traditional models, tokens were stemmed and converted into TF-IDF feature vectors.

Classification Models

Four traditional classifiers were implemented with scikit-learn. Naïve Bayes is a probabilistic classifier that applies Bayes’ theorem under a conditional-independence assumption between features. A Decision Tree recursively partitions the data using the most informative features. Logistic Regression models the probability of a class with a logistic function whose coefficients are optimised by gradient-based methods. Random Forest is an ensemble of decision trees trained on random subsets of the data and features, whose predictions are aggregated by majority vote.

GPT-2 is a transformer-based language model pre-trained on a large English corpus in a self-supervised manner to predict the next token in a sequence [6]. This allows it to learn rich contextual representations without human-labeled data. In this study, the pre-trained model was fine-tuned for binary sequence classification with the Hugging Face Transformers library. Fig. 2 illustrates the general GPT-style architecture, and the fine-tuning procedure is given below.

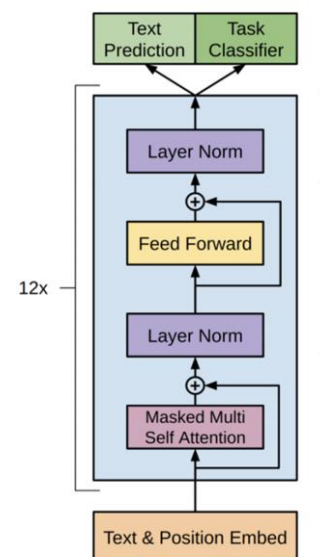


Fig 2_Overview of the GPT-Style Transformer Architecture with Text-Prediction and Task-Classifer Outputs

Algorithm 1: Fine-tuning the GPT-2 classifier

1. Load the labeled tweet dataset.
2. Pre-process the text and convert the labels to integers.
3. Split the data into training and testing sets.
4. Load the pre-trained GPT-2 model and tokenizer.
5. Tokenize the training and testing texts.
6. Fine-tune the model on the training set.
7. Predict the labels of the testing set.
8. Output the accuracy and the classification report.

Evaluation Metrics

Model performance was measured using the following metrics, where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$

Experimental Setup

All experiments were run on a single local machine with an AMD Ryzen 5 processor and 8 GB of RAM, using Python with the scikit-learn and Hugging Face Transformers libraries. The data were divided into training and testing sets using a train: test ratio of 80:20 split, and all five models were trained and evaluated on the same split to ensure a fair comparison. The GPT-2 model was fine-tuned and training and validation loss were recorded during fine-tuning.

Results:

Table 1 compares the five classifiers. The fine-tuned GPT-2 classifier achieved the highest accuracy (88.7%) and the highest precision, recall and F1-score on the test data. Among the traditional classifiers, Naïve Bayes performed best (accuracy 78%, F1-score 0.78), followed by Logistic Regression (76%), Random Forest (74%) and Decision Tree (70%). GPT-2 therefore exceeded the best traditional baseline by 10.7 percentage points in accuracy and by 0.11 in F1-score.

Algorithm	Accuracy	Precision (class 0 / 1)	Recall	F1-score
Naïve Bayes	78%	0.80 / 0.77	0.78	0.78
Decision Tree	70%	0.73 / 0.69	0.71	0.71
Logistic Regression	76%	0.77 / 0.76	0.76	0.76
Random Forest	74%	0.77 / 0.72	0.74	0.74
GPT-2 Classifier	88.7%	0.88 / 0.89	0.89	0.89

Table 1. Performance Comparison of the Classifiers on the Test Set (Recall and F1-score Are Weighted Averages)

Training Behaviour of GPT-2:

Fig. 3 shows the training and validation loss recorded during fine-tuning. Both losses decreased from the first to the last epoch (training loss from about 0.50 to 0.10 and validation loss from about 0.80 to 0.20), indicating that the model was learning from the data. Validation loss remained above training loss, and the training-validation gap fell from about 0.30 in the first epoch to about 0.10 from the second epoch onward. This persistent but small gap

suggests mild overfitting, which should be monitored in future work, for example with early stopping or stronger regularisation.

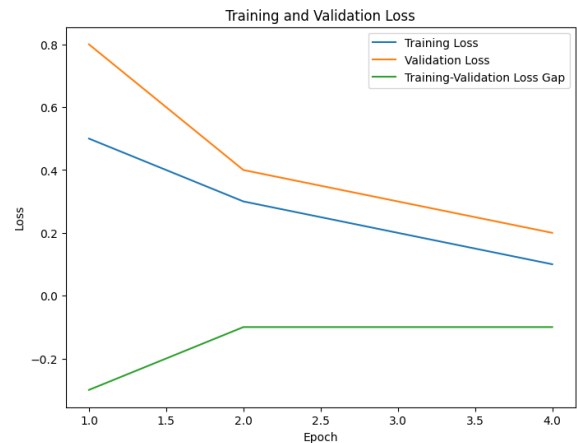


Fig 3_Training Loss, Validation Loss and Training-Validation Gap During Fine-Tuning of GPT-2

Discussion:

The fine-tuned GPT-2 classifier outperformed all four traditional classifiers on every metric. A plausible explanation is that GPT-2 builds contextual representations of whole sentences, whereas the traditional models rely on TF-IDF features that capture word frequency but not word order or context. This may help GPT-2 to:

- detect subtle or indirect expressions of stress;
- interpret words whose meaning depends on context; and
- use linguistic knowledge acquired during pre-training on a large corpus.

Among the traditional models, Naïve Bayes performed best, which may reflect its suitability for high-dimensional, sparse text features, while Logistic Regression and Random Forest gave moderate results and the Decision Tree performed worst. Because no ablation was performed, these explanations remain hypotheses rather than tested conclusions.

The results are consistent with earlier studies reporting that language models outperform conventional classifiers for stress detection in social media. Ramteke et al. [11] found that GPT-based language models achieved much higher recall than conventional ML models, and Rajput et al. [12] reported that a fine-tuned BERT model outperformed SVM, Logistic Regression, Random Forest and XGBoost on Reddit and Twitter data, with a maximum F1-score of 0.89. Nijhawan et al. [1] likewise reported strong performance for a BERT-based model on tweets. Ensemble methods have also performed well in other stress-detection settings [8], so combining several classifiers is a promising direction for improving accuracy further.

Limitations and Ethical Considerations:

Several limitations should be considered when interpreting these findings. First, only one English-language Twitter dataset was used, so the results may not generalise to other platforms, languages or time periods. Second, the procedure used to label the posts is not documented; if labels were assigned automatically, for example from keywords, the models may partly learn the labelling rule rather than stress itself. Third, a single train-test split was used, without

cross-validation, repeated runs, confidence intervals or significance tests, so the small differences among the four traditional models should be interpreted cautiously, although the gap between GPT-2 and the best baseline is large. Fourth, the traditional baselines were not extensively tuned, and stronger baselines such as SVM, XGBoost or BERT were not included.

Stress detection from social media involves personal data and sensitive health-related inferences. This study used an existing public dataset and collected no new data. Any real-world deployment would need to anonymise data, comply with platform terms and data-protection regulations, and treat model outputs as screening signals rather than clinical diagnoses, with human oversight to limit the harm from misclassification or misuse such as profiling.

Conclusion:

This study compared a fine-tuned GPT-2 classifier with Naïve Bayes, Decision Tree, Logistic Regression and Random Forest for detecting stress in 20,001 Twitter posts. GPT-2 achieved the best performance on all evaluation metrics, with an accuracy of 88.7% and an F1-score of 0.89, compared with 70–78% accuracy for the traditional classifiers. These findings support the use of transformer-based language models for text-based mental-health screening. Future work should validate the approach on multiple datasets with documented labels, use cross-validation and statistical testing, compare against stronger baselines, explore real-time and multilingual detection, and examine integration with clinical tools under appropriate ethical safeguards.

References:

1. Nijhawan, T., Attigeri, G. and Ananthakrishna, T., 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9, Article 33. <https://doi.org/10.1186/s40537-022-00575-6>
2. Vidal, C. et al., 2024. Chronic stress and lack of social support: Role in adolescent depression and suicide-related behaviors in the context of the COVID-19 pandemic. *Journal of Affective Disorders*, 365, pp.437-442.
3. World Health Organization, 2021. Suicide worldwide in 2019: Global Health Estimates. Geneva: World Health Organization.
4. Gedam, S. and Paul, S., 2021. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9, pp.84045-84066. <https://doi.org/10.1109/ACCESS.2021.3085502>
5. Can, Y.S., Arnrich, B. and Ersoy, C., 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of Biomedical Informatics*, 92, Article 103139. <https://doi.org/10.1016/j.jbi.2019.103139>
6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), p.9.
7. Li, R. and Liu, Z., 2020. Stress detection using deep neural networks. *BMC Medical Informatics and Decision Making*, 20(Suppl 11), Article 285. <https://doi.org/10.1186/s12911-020-01299-4>
8. Garg, P., Santhosh, J., Dengel, A. and Ishimaru, S., 2021. Stress detection by machine learning and wearable sensors. In: 26th International Conference on Intelligent User Interfaces – Companion (IUI '21 Companion). New York: ACM.
9. Inamdar, S., Chapekar, R., Gite, S. and Pradhan, B., 2023. Machine learning driven mental stress detection on Reddit posts using natural language processing. *Human-Centric Intelligent Systems*, 3(2), pp.80-91. <https://doi.org/10.1007/s44230-023-00020-8>
10. Ilias, L., Mouzakitis, S. and Askounis, D., 2023. Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Transactions on Computational Social Systems*. <https://doi.org/10.1109/TCSS.2023.3283009>
11. Ramteke, P. et al., 2023. Comparing conventional machine learning and large-language models for human stress detection using social media posts. In: 2023 2nd International Conference on Futuristic Technologies (INCOFT). IEEE.
12. Rajput, G. et al., 2025. Contextual stress detection in social media using transformer-based language models. In: 2025 2nd International Conference on Advanced Computing and Emerging Technologies (ACET). IEEE.