Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00378-3

Original Article

# Term Frequency-Inverse Document Frequency (TF-IDF)-Based Machine Learning and a Zero-Shot Fine-Tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) Baseline for Depression Risk Detection from Text

Parveen Kumari [1], Alpana Jijja [1]

1. *Department of Computer Science and Engineering, Sushant University, Gurugram, IND*

Received: September 18, 2026 | Review began: September 22, 2026 | Review ended: October 06, 2026 | Published: November 18, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## Abstract

Text-based depression risk detection is an important area of computational mental health because written language may contain signals associated with psychological distress and can support scalable risk-screening research. This study evaluates three term frequency-inverse document frequency (TF-IDF)-based classical machine learning classifiers - Logistic Regression, Multinomial Naive Bayes, and Linear Support Vector Machine (SVM) - against a small instruction-tuned zero-shot Fine-tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5)-base baseline on a public Reddit depression dataset. Performance was assessed using accuracy, precision, recall, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC). Linear SVM achieved the highest accuracy (0.9651), F1-score (0.9644), and ROC-AUC (0.9937), while Logistic Regression obtained a very similar precision (0.9743 versus 0.9747). Naive Bayes achieved higher recall than Linear SVM (0.9687 versus 0.9544), and the FLAN-T5 baseline produced recall of 1.0000 because it predicted every test sample as depression; its accuracy was 0.4958, and ROC-AUC was 0.5000. Therefore, the findings support the effectiveness of TF-IDF-based linear classifiers for this dataset and experimental configuration, but they should not be interpreted as a general comparison between classical machine learning and large language models (LLMs). The FLAN-T5 result represents only the tested zero-shot prompt, truncation, decoding, and score-parsing setup. Additional tuning, repeated validation, stronger transformer/LLM baselines, and cross-dataset testing are required before broader conclusions can be drawn.

**Categories:** AI applications, AI/ML-based decision support systems, Algorithm Design

**Keywords:** depression risk detection, text classification, tf-idf, logistic regression, naive bayes, linear svm, large language models, flan-t5, computational mental health

## Introduction

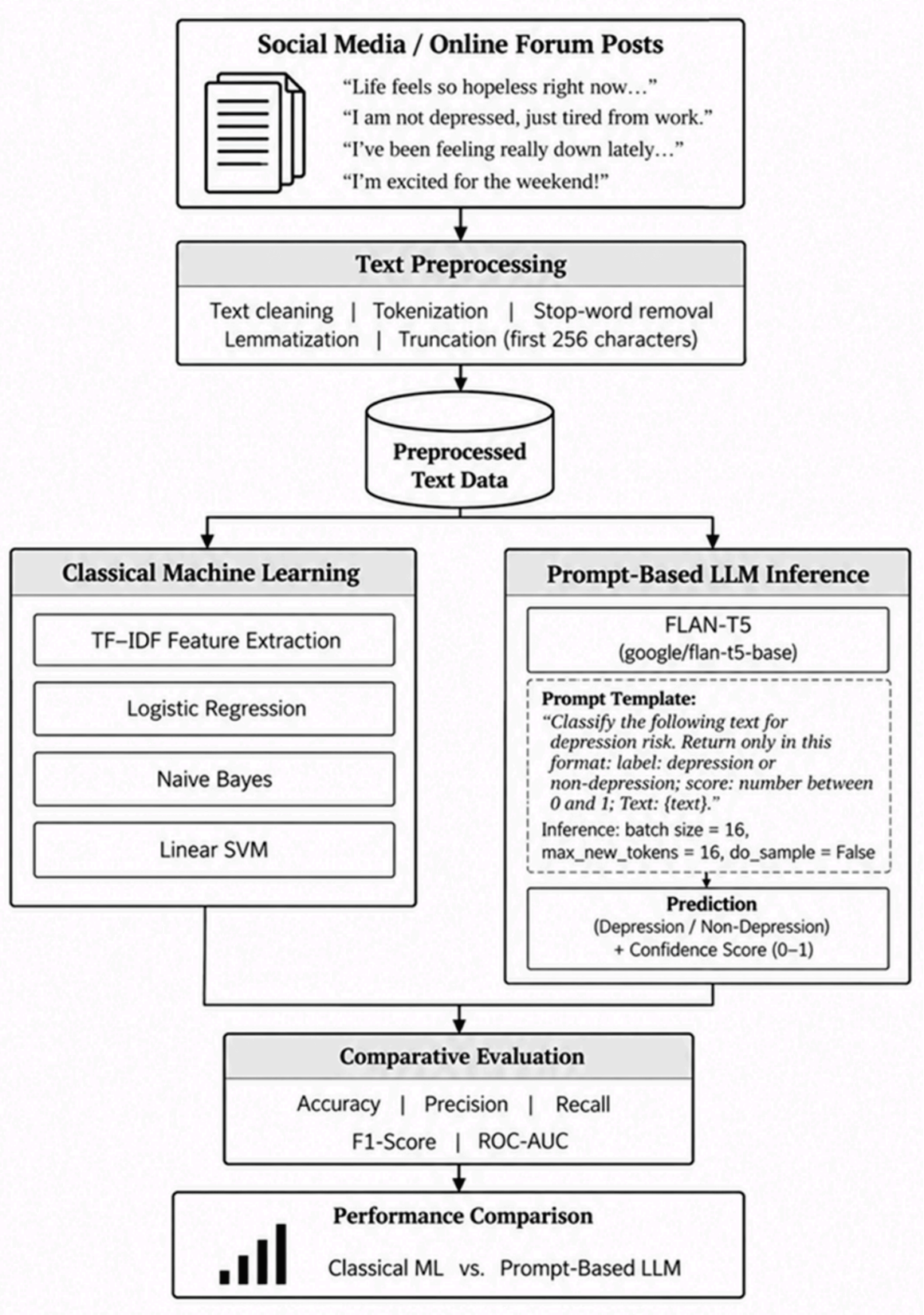
Depression is among the most common forms of mental health disorders in the whole world, as well as a significant issue in public health due to its effect on emotional stability, cognition, daily living, and quality of life [1,2]. Late detection and accessibility of mental health services usually lead to failure in timely intervention, and this has sparked the pursuit of scalable computational tools that have the potential of aiding in timely risk detection [1,3,4]. Social media and online forums text written by users have become a promising source of computational mental health research in recent years, as people often reveal their feelings, lived experiences, and psychological distress using written language [5-7]. This has been associated with increasing accessibility to such textual information, which has resulted in extensive studies on automatic detection of depression based on online language [4,5,7]. Initial experiments showed that linguistic cues,

behavioral cues, and topic patterns obtained via social media can be used to differentiate among users at risk of depression, as well as control users [8,9]. Systematic experimentation of depression-related language modeling and early risk prediction has also been made possible by eRisk test collection as well as the Reddit Self-reported Depression Diagnosis (RSDD) dataset [10-12]. These collections of data have been significant in both facilitating the reproducibility of the task and in promoting the comparison of rival methods under familiar evaluation environments [10,12,13]. This task has been based on classical machine learning techniques. Term frequency-inverse document frequency (TF-IDF)-based approaches with classifiers like Logistic Regression, Naive Bayes, and Support Vector Machines (SVMs) are still appealing as they are computationally efficient, feature-level interpretable, and attractive for sparse high-dimensional text classification [4,7,14]. In addition to basic lexical modeling, past research has also demonstrated that multimodal and feature-based systems can be used to detect depression by adding topic, emotion, and user-behavior information [9,15]. Meanwhile, deep learning models, such as convolutional neural networks (CNNs), hierarchical attention networks, and transformer-based networks, have been considered to enhance the ability to capture semantic context as well as long-range dependencies in user-generated text [13,16-18]. Such techniques have frequently enhanced learned representations but can still differ in their performance across datasets and collection procedures, particularly when there is a domain shift or label construction variance across corpora [16,19]. Most recently, with large language models (LLMs) taking over as the new frontier in mental health text analysis, they allow instruction-following inference, prompt-based reasoning, and evidence extraction in flexible ways without necessarily being designed as a full-fledged task [20,21]. The new studies indicate that LLMs have the potential to aid explainable evaluation, symptom-based prompting, and systematic evidence summarizing within mental care environments [21,22]. Even so, recent comparisons also suggest that prompt-based LLMs do not always perform better than task-specific discriminative models and can be outperformed by strong traditional or encoder-based baselines on depression detection tasks, especially with limited model size, prompt design, and calibration [20,23,24]. This is why comparative evaluation is of particular importance.

Another important issue in this area concerns validity, generalizability, and responsible use. Reviews of computational mental health research have highlighted persistent challenges related to dataset bias, construct validity, annotation quality, demographic imbalance, and the risk of overinterpreting social media-based predictions as clinical diagnoses [3,5,7]. Moreover, temporal variation in user condition, cross-domain transfer, and control-group construction can significantly affect observed performance, meaning that results should be interpreted carefully and framed as risk detection rather than diagnosis [11,16,19,25]. These concerns are especially relevant when comparing highly interpretable sparse-feature baselines with prompt-based LLM systems whose outputs may appear fluent but may not yield reliable class separation [20,21,24]. Figure *1* presents an overview of the comparative framework considered in this study, where raw text samples are processed through both TF-IDF-based classical machine learning models and a prompt-based LLM baseline for depression risk detection from text [4,14,20,21].

Motivated by these considerations, this paper compares text-based depression risk detection using TF-IDF-based classical machine learning models and a prompt-based LLM baseline. Specifically, Logistic Regression, Naive Bayes, and Linear SVM are trained on TF-IDF representations and evaluated against a prompt-based Fine-tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) baseline. The study aims to determine whether lightweight prompt-based LLM inference can match the effectiveness of established sparse-text classifiers on a benchmark depression dataset. We evaluate performance using accuracy, precision, recall, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC). The results show that, under the present experimental configuration, the TF-IDF-based models achieved higher observed performance than the tested zero-shot FLAN-T5-base baseline. This finding is specific to the evaluated dataset, hold-out split, and implementation settings and should not be interpreted as evidence of general superiority of classical machine learning over LLMs.

**How to cite this article:**
Kumari P, Jijja A (November 18, 2026) Term Frequency-Inverse Document Frequency (TF-IDF)-Based Machine Learning and a Zero-Shot Fine-Tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) Baseline for Depression Risk Detection from Text. Cureus J Comput Sci 3 : es44389-026-00378-3. DOI https://doi.org/10.7759/s44389-026-00378-3




**How to cite this article:**
Kumari P, Jijja A (November 18, 2026) Term Frequency-Inverse Document Frequency (TF-IDF)-Based Machine Learning and a Zero-Shot Fine-Tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) Baseline for Depression Risk Detection from Text. Cureus J Comput Sci 3 : es44389-026-00378-3. DOI https://doi.org/10.7759/s44389-026-00378-3

**FIGURE 1: Overview of the proposed comparative framework for text-based depression risk detection.**

TF-IDF, Term Frequency-Inverse Document Frequency; SVM, Support Vector Machine; FLAN-T5, Fine-tuned Language Net-Text-to-Text Transfer Transformer; ROC-AUC, Receiver Operating Characteristic-Area Under the Curve; ML, Machine Learning; LLM, Large Language Model

## Related work

Text-based depression risk detection research has taken place in three key directions: classical machine learning, deep learning, and more recently through LLM approaches. The initial experiments on computational mental health research have already proved that linguistic patterns in social media posts can give informative indications on detecting depression-related behavior and, thus, encourage automated screening based on user-generated text [3,4,7]. This domain was further accelerated by the introduction of resources like eRisk and RSDD, which allowed evaluating the comparative performance on shared datasets and task configurations [10-12].

### *Classical Machine Learning for Depression Detection*

Classical machine learning techniques remain widely used baselines for text-based depression risk detection. They usually use sparse lexical representations using bag-of-words or TF-IDF with other classifiers, including Logistic Regression, Naive Bayes, and SVMs [4,8,14]. Their continued use reflects several practical advantages: they are computationally efficient, can support feature-level inspection, and can be highly competitive in high-dimensional text classification tasks [4,7]. Earlier research has shown that these models can achieve high performance when depression-related cues are reflected in vocabulary use, emotional expression, and repetitive language patterns [8,9]. Moreover, these methods suit benchmark environments where reproducibility, computational efficiency, and stable optimization are important [10,14].

### *Deep Learning Approaches*

In order to address the shortcomings of handcrafted or sparse lexical features, subsequent research investigated deep learning-based richer semantic modeling. CNNs, recurrent neural networks, hierarchical attention models, and transformer-based encoders have all been used in the detection of depression in text [13,16-18]. These techniques attempt to learn contextual dependencies, latent semantics, and long-range text structure in a superior way compared to the traditional shallow classifiers [16,18]. Nevertheless, they do not always perform equally well with datasets, particularly when the target labels are contaminated, the training samples are small, or when the data collection procedure is different across platforms [3,5,19]. This leaves strong classical procedures like Linear SVM and Logistic Regression to be competitive in most text classification tasks [4,14].

### *LLMs in Mental Health Text Analysis*

Recent advances in generative artificial intelligence have introduced LLMs as a new paradigm for mental health text analysis. Prompt-based models can perform classification and evidence generation through instruction-following behavior, without task-specific architecture design [21,22]. Recent CLPsych work has shown that LLMs can be useful for explainable depression-related evidence extraction and symptom-oriented reasoning from social media data [21,22]. At the same time, broader recent evaluations indicate that LLMs are not uniformly superior to traditional machine learning or encoder-based systems for mental health classification tasks [20,23,24]. Their effectiveness depends strongly on model scale, prompt design, output formatting, and the alignment between the prompting strategy and the target task [20,21]. This suggests that LLMs should be evaluated carefully as comparative baselines rather than assumed to outperform simpler discriminative models in all scenarios [23,24].

### *Research Gap*

Previous studies have established strong classical, neural, and transformer-based approaches for depression-related text classification, while recent work has also explored prompt-based LLMs [4,14,20,21,23,24]. However, performance depends on dataset construction, model scale, prompting strategy, and evaluation design. The present study therefore addresses a

deliberately narrow question: how do three TF-IDF-based classical classifiers compare with one lightweight, zero-shot FLAN-T5-base configuration on the same dataset and hold-out split? The study is not intended to represent the broader performance of LLMs or fine-tuned transformers. Instead, it provides a reproducible baseline comparison and identifies limitations of the tested zero-shot prompting configuration.

## Materials And Methods

This paper presents a comparative study of depression risk detection on text, comparing three classical machine learning models with a prompt-based LLM baseline. It involves text representation, model training, prompt-based inference, and performance evaluation in the workflow [4,14,21].

*Text Representation*

In the case of the classical models, the input text is converted to TF-IDF features, which are commonly applied in text classification to indicate the significance of terms in a document collection [4,7,14]. Individual words and short contextual phrases are captured with the help of unigrams and bigrams.

*Classical Machine Learning Models*

Three supervised classifiers were evaluated using scikit-learn pipelines with TF-IDF features. For all three classical models, TF-IDF vectorizer was configured with max_features=10,000, ngram_range=(1,2), and stop_words="english". Logistic Regression used max_iter=1000 and random_state=42; all other Logistic Regression parameters were left at scikit-learn defaults. Multinomial Naive Bayes was used with its default parameters. Linear SVM used LinearSVC with random_state=42 and was wrapped in CalibratedClassifierCV using sigmoid calibration with 3-fold cross-validation (cv=3) to obtain probability estimates for ROC-AUC calculation. These models were trained on the TF-IDF feature vectors of the training data.

*Prompt-Based LLM Baseline*

The prompt-based LLM baseline used google/flan-t5-base. Each text sample was converted to a string, newline characters were removed, leading and trailing whitespace was stripped, and the input was truncated to the first 256 characters. The exact prompt template was: “Classify the following text for depression risk. Return only in this format: label: depression or non-depression; score: number between 0 and 1, Text: {text}.” Inference was performed in batches of 16 with max_new_tokens=16 and do_sample=False. The generated output was converted to a binary label by matching “label: depression” or “label: non-depression.” If neither label was detected, the fallback label was non-depression. A numeric value following “score:” was parsed as the confidence score and clipped to the interval (0, 1). When no valid score was present, a fallback score of 0.5 was assigned. ROC-AUC for the FLAN-T5 baseline was computed using these parsed confidence scores.

*Evaluation Protocol*

After removal of missing text/label values, the dataset was divided using a stratified 80:20 train-test split with random_state=42. This produced 6,184 training samples and 1,547 test samples while preserving class proportions. Accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate performance. For Logistic Regression and Naive Bayes, ROC-AUC was computed from predict_proba scores; for the calibrated Linear SVM, it was computed from calibrated probability estimates; and for FLAN-T5, it was computed from the parsed prompt-generated confidence scores.

*Framework Objective*

The framework compares three TF-IDF-based classical classifiers with one specific zero-shot FLAN-T5-base baseline for depression-risk classification. The objective is limited to the present dataset, hold-out split, and implementation settings; it is not a general comparison of classical machine learning with the wider class of LLM or transformer approaches.

### Dataset

This study used the publicly available "Depression: Reddit Dataset (Cleaned)" from Kaggle, containing 7,731 labeled English-language text samples. The dataset includes 3,900 non-depression and 3,831 depression samples. The labels were used as provided in the dataset. Since the original label-generation process was not independently clinically validated, this limitation was considered when interpreting the results.

*Dataset Provenance and Validity Clarification*

The Kaggle release provides the cleaned text and binary labels used in this study. However, the available dataset documentation does not provide sufficient information to reconstruct the original Reddit data-collection procedure, subreddit composition, detailed cleaning process, or clinical basis for label generation. Therefore, the labels were used as provided in the dataset and are considered research labels rather than clinically verified diagnoses. The present experimental workflow did not include systematic exact-duplicate, near-duplicate, or train-test leakage analysis. Consequently, the potential influence of lexical overlap, subreddit-specific vocabulary, or topic-related differences on classification performance cannot be completely excluded. These factors limit the interpretation and generalizability of the reported results and do not support claims of clinical validity or performance across other platforms or populations.

### Experimental setup

The 7,731 samples were divided using a stratified 80:20 split with random_state=42, resulting in 6,184 training samples and 1,547 test samples. The test set contained 780 non-depression and 767 depression samples. For the classical models, TF-IDF used up to 10,000 unigram and bigram features with English stop-word removal. Logistic Regression (max_iter=1000, random_state=42), Multinomial Naive Bayes (default parameters), and LinearSVC (random_state=42) with sigmoid probability calibration (CalibratedClassifierCV, cv=3) were trained on the training set. The FLAN-T5 baseline used google/flan-t5-base with deterministic decoding (do_sample=False), max_new_tokens=16, batch_size=16, and the prompt/scoring procedure described in the "Prompt-Based LLM Baseline" section. Performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC.

## Results And Discussion

| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.9573 | 0.9743 | 0.9387 | 0.9562 | 0.9924 |
| Naive Bayes | 0.8985 | 0.8482 | 0.9687 | 0.9044 | 0.9832 |
| Linear SVM | 0.9651 | 0.9747 | 0.9544 | 0.9644 | 0.9937 |
| FLAN-T5 (Prompt-based LLM) | 0.4958 | 0.4958 | 1.0000 | 0.6629 | 0.5000 |

**TABLE 1: Comparative performance of the evaluated models for depression risk detection from text.**

SVM, Support Vector Machine; FLAN-T5, Fine-tuned Language Net-Text-to-Text Transfer Transformer; LLM, Large Language Model; ROC-AUC, Receiver Operating Characteristic-Area Under the Curve

Table *1* summarizes performance on the held-out test set. Under the present experimental configuration, the TF-IDF-based classical models achieved substantially higher accuracy, F1-score, and ROC-AUC than the tested zero-shot FLAN-T5-base baseline. Linear SVM achieved the highest accuracy (0.9651), F1-score (0.9644), and ROC-AUC (0.9937), and the

highest precision by a very small margin (0.9747 versus 0.9743 for Logistic Regression). It did not achieve the highest recall: Naive Bayes obtained 0.9687 recall, and FLAN-T5 obtained 1.0000 recall because the latter predicted all test samples as depression. Thus, Linear SVM achieved the strongest observed overall balance of the reported metrics on this hold-out split; this descriptive result should not be interpreted as statistically established superiority over the other classifiers.

Logistic Regression also yielded highly competitive performance, achieving an accuracy of 0.9573, F1-score of 0.9562, and ROC-AUC of 0.9924. The relatively small performance gap between Logistic Regression and Linear SVM suggests that linear decision boundaries over TF-IDF representations are highly suitable for this classification problem. In contrast, Naive Bayes obtained comparatively lower overall results, with an accuracy of 0.8985 and F1-score of 0.9044, although it maintained a high recall of 0.9687. This indicates that Naive Bayes is effective in identifying depression-related samples, but at the cost of a larger number of false positive predictions.

The prompt-based FLAN-T5 baseline exhibited substantially weaker performance than the classical machine learning models. It achieved an accuracy of 0.4958, precision of 0.4958, recall of 1.0000, F1-score of 0.6629, and ROC-AUC of 0.5000. Examination of the prediction output showed that FLAN-T5 predicted the depression class for all 1,547 test samples: all 767 depression samples were predicted as depression, while all 780 non-depression samples were also predicted as depression. This explains the recall of 1.0000 for the depression class, precision of 0.4958, and overall accuracy of 0.4958. The confidence score used for ROC-AUC was the numeric value parsed after “score:” in the generated response; when a valid score was absent, the implementation assigned 0.5. ROC-AUC was then calculated using these scores with sklearn.metrics.roc_auc_score. The resulting ROC-AUC of 0.5000 indicates no ranking discrimination in this configuration. Thus, the reported FLAN-T5 result should be interpreted as a degenerate baseline outcome under the specific prompt, decoding, and score-parsing procedure used in this experiment, rather than as evidence about FLAN-T5 performance under other configurations.

**Baseline validity clarification**

We evaluated FLAN-T5-base as a zero-shot baseline under the specific experimental configuration described in the "Prompt-Based LLM Baseline" section. We truncated each input to the first 256 characters and parsed the generated responses to obtain class labels and confidence scores, with a fallback score of 0.5 when no valid score was available. Under this configuration, FLAN-T5-base predicted all test samples as depression and therefore did not achieve meaningful class discrimination. Accordingly, these results are limited to the tested prompt, input truncation, decoding, and score-parsing procedure and should not be generalized to other FLAN-T5 configurations or broader neural and LLM approaches. Stronger neural and LLM baselines were not evaluated in the present study; therefore, no broader comparison with such models is claimed.

The Linear SVM confusion matrix in Figure *2* shows 761 correctly classified non-depression samples and 732 correctly classified depression samples. Among the 54 errors, 19 were false positives (non-depression predicted as depression) and 35 were false negatives (depression predicted as non-depression). This count-based error analysis shows that false negatives were more frequent than false positives in the held-out test set. Because the current experiment did not include a qualitative review of the misclassified texts, no claim is made about the linguistic causes of these errors. In a screening context, both error types matter: false positives may lead to unnecessary concern or follow-up, whereas false negatives may fail to flag text associated with possible risk.

**How to cite this article:**
Kumari P, Jijja A (November 18, 2026) Term Frequency-Inverse Document Frequency (TF-IDF)-Based Machine Learning and a Zero-Shot Fine-Tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) Baseline for Depression Risk Detection from Text. Cureus J Comput Sci 3 : es44389-026-00378-3. DOI https://doi.org/10.7759/s44389-026-00378-3

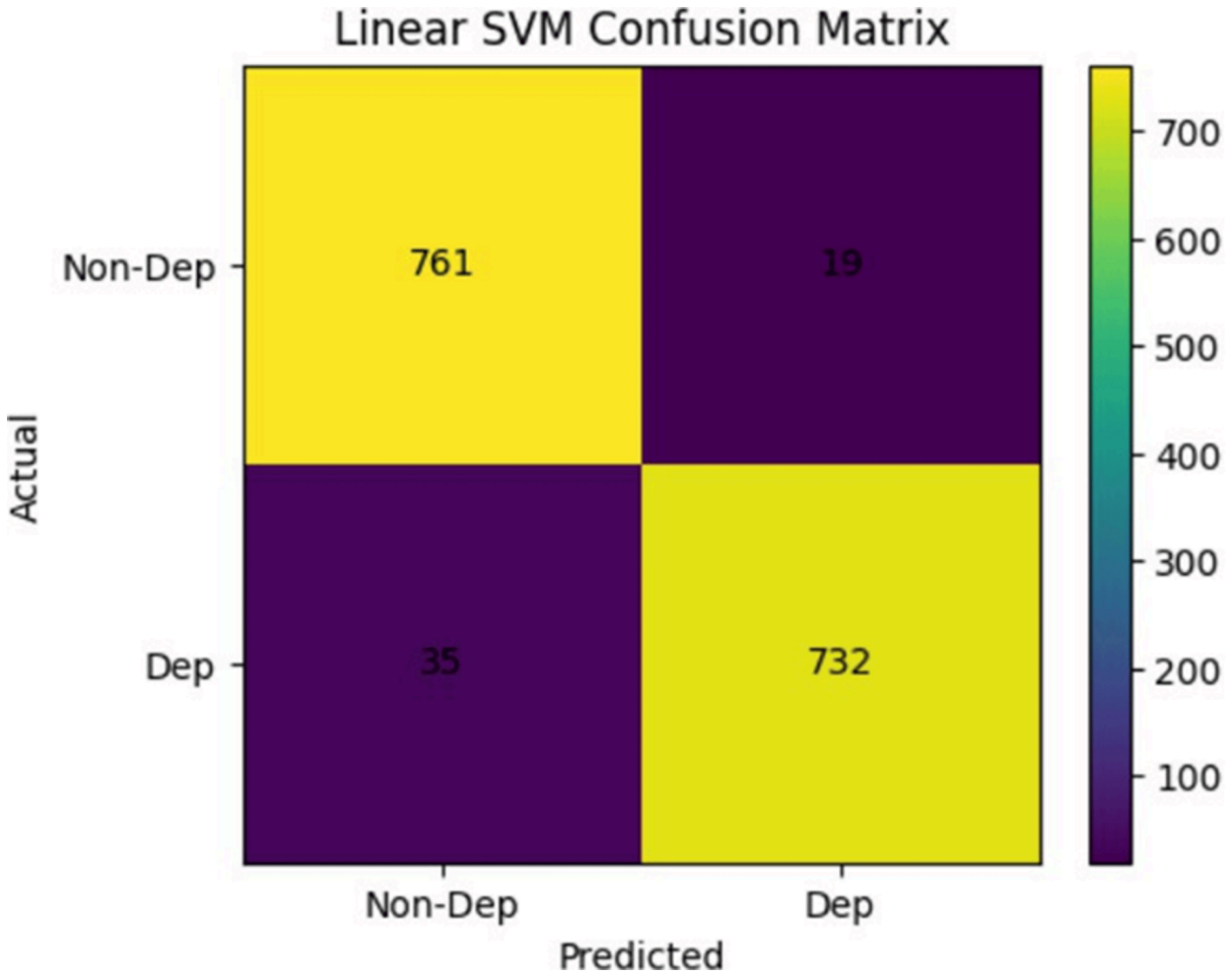



**FIGURE 2: Confusion matrix of the Linear SVM model, which achieved the highest observed accuracy and F1-score on the evaluated hold-out split.**

SVM, Support Vector Machine

The discriminative capability of Linear SVM on the held-out split is illustrated by the ROC curve in Figure *3*. The ROC-AUC of 0.9937 indicates strong ranking separation on this test set and is consistent with the reported classification metrics. Because the study uses a single dataset and a single train-test split, this result should be interpreted as performance within the present experiment rather than evidence of reliability across populations, platforms, or datasets.

**How to cite this article:**
Kumari P, Jijja A (November 18, 2026) Term Frequency-Inverse Document Frequency (TF-IDF)-Based Machine Learning and a Zero-Shot Fine-Tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) Baseline for Depression Risk Detection from Text. Cureus J Comput Sci 3 : es44389-026-00378-3. DOI https://doi.org/10.7759/s44389-026-00378-3

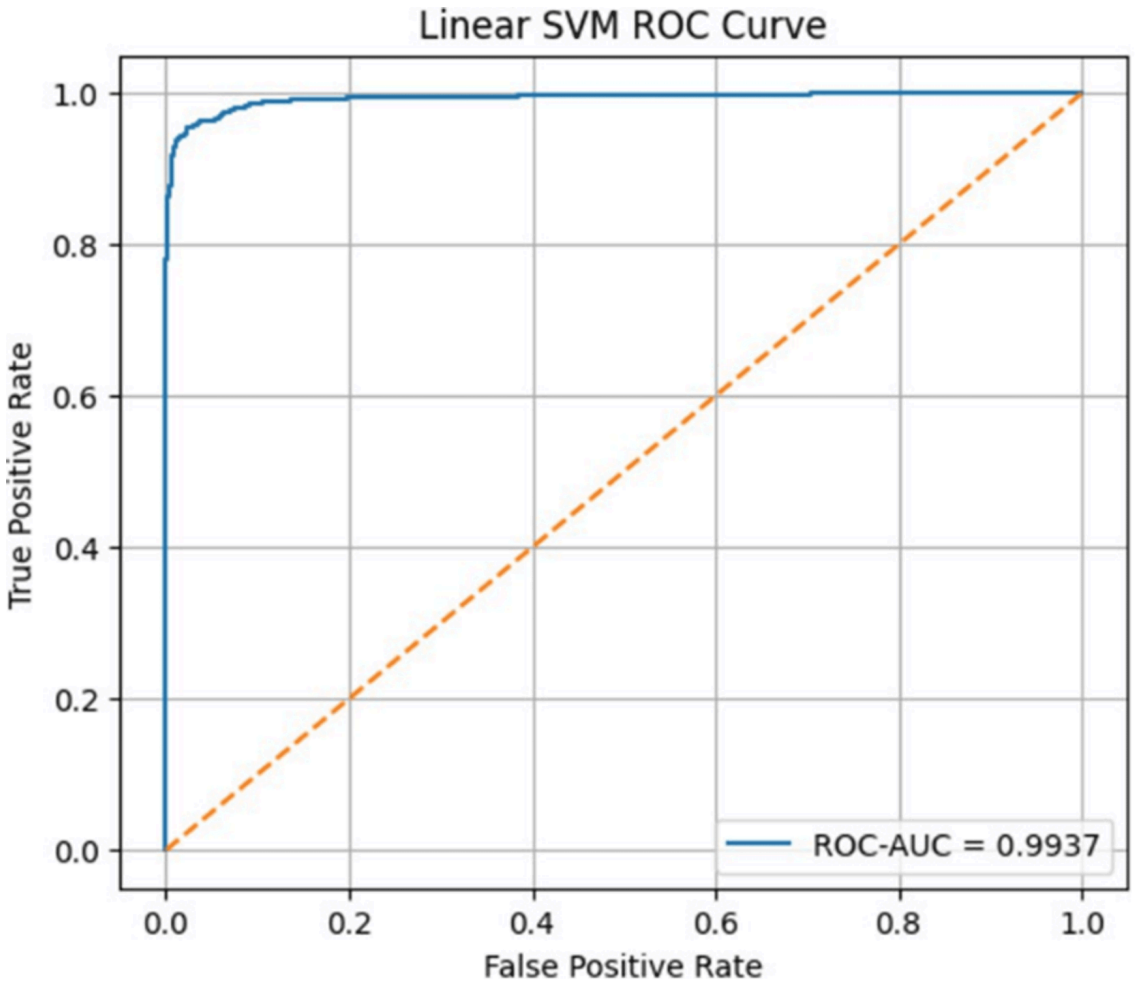



**FIGURE 3: ROC curve of the Linear SVM model with an observed ROC-AUC of 0.9937 on the evaluated hold-out split.**

ROC-AUC, Receiver Operating Characteristic-Area Under the Curve; SVM, Support Vector Machine

Overall, the findings show that TF-IDF-based classical models were highly effective on this dataset. Linear SVM provided the strongest overall balance, with the highest accuracy, F1-score, and ROC-AUC, while Naive Bayes and the degenerate FLAN-T5 output produced higher recall. The FLAN-T5 result applies only to the tested FLAN-T5-base zero-shot configuration and should not be generalized to larger LLMs, few-shot prompting, logit-based scoring, full-length inputs, fine-tuned transformers, or other datasets.

### Interpretability and model selection clarification

TF-IDF-based linear models allow direct inspection of learned feature weights, providing a basis for feature-level interpretation. However, a detailed coefficient-ranking analysis was not conducted in the present study; therefore, feature-level interpretations are not reported as empirical findings. The Linear SVM produced 54 misclassified samples, comprising 19 false positives and 35 false negatives. The present analysis reports these observed error counts without attributing the misclassifications to specific linguistic characteristics. Further examination of highly weighted features and representative false-positive and false-negative cases could provide additional insight into the model's decision patterns. The classical models were evaluated using the specified fixed settings and default parameters rather than systematic hyperparameter optimization. Consequently, the reported results represent the evaluated baseline configurations and should not be interpreted as the optimally tuned performance of these models.

## Conclusions

This study investigated the effectiveness of three traditional machine learning classifiers using TF-IDF feature extraction and a lightweight zero-shot FLAN-T5-base model for depression-risk classification from textual data. The experimental results demonstrated that Linear SVM achieved the strongest overall performance, with an accuracy of 0.9651, an F1-score of 0.9644, and a ROC-AUC of 0.9937. Logistic Regression delivered comparable precision, while Naive Bayes demonstrated higher recall. In comparison, FLAN-T5-base achieved a recall of 1.0000 but produced considerably lower accuracy (0.4958) and a ROC-AUC of 0.5000, as it classified all test samples as depression-related. These findings highlight the effectiveness of TF-IDF-based linear classifiers within the specific dataset and experimental settings considered in this study. Nevertheless, the results do not establish the general superiority of traditional machine learning over LLMs, as the LLM evaluation was limited to a single lightweight zero-shot configuration. The study also has certain limitations, including the use of a single publicly available Reddit-derived dataset, one stratified train-test split, fixed model configurations, and the absence of systematic hyperparameter tuning, cross-validation, statistical significance testing, and external validation. Additionally, potential dataset biases, label reliability, demographic imbalance, and platform-specific language characteristics may affect the generalizability of the findings.

Future research should explore more advanced transformer-based architectures, larger instruction-tuned language models, fine-tuning strategies, few-shot prompting, and alternative confidence-scoring approaches to provide a more comprehensive comparison. Further improvements could include systematic model optimization, repeated experiments, cross-validation, independent dataset evaluation, feature-importance analysis, and qualitative examination of misclassified instances. Investigating fairness, privacy, demographic bias, and the implications of false-positive and false-negative predictions would also strengthen the reliability and responsible application of such systems. Overall, this study provides an initial experimental assessment of machine learning and zero-shot language models for depression-risk classification from text. However, the proposed system has not been clinically validated and should be considered a research-oriented classification approach rather than a diagnostic or clinical screening tool. Its potential practical application would require extensive external validation, appropriate ethical safeguards, and professional human oversight.

## Additional Information

### Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

**Acquisition, analysis, or interpretation of data:** Parveen Kumari, Alpana Jijja

**Drafting of the manuscript:** Parveen Kumari

**Critical review of the manuscript for important intellectual content:** Alpana Jijja

## Disclosures

**Human subjects:** Consent was waived for all participants in this study. The study utilizes a publicly accessible secondary dataset, as specified in the Data Availability Statement, without any direct involvement or interaction with human participants. The dataset is available on Kaggle under the CC0: Public Domain license, and no efforts were undertaken to identify individual users. Furthermore, the proposed system has not undergone clinical validation and is not designed or intended for clinical screening, diagnosis, or treatment-related decision-making.

**Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue.

**Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.

- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

### Data Availability Statements

The "Depression: Reddit Dataset (Cleaned)" used in this study is publicly available on Kaggle: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/data

### Acknowledgements

AI-Assisted Writing Disclosure: AI-assisted tools were used only for limited language and grammatical refinement. All scientific content, analyses, interpretations, and conclusions were reviewed and verified by the authors.

## References

1. World Health Organization: Depressive disorder (depression). (2025). Accessed: September 18, 2026: https://www.who.int/news-room/fact-sheets/detail/depression.
2. World Health Organization: Mental disorders. (2025). Accessed: September 18, 2026: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.
3. Chancellor S, De Choudhury M: Methods in predictive techniques for mental health status on social media: a critical review. npj Digital Medicine. 2020, 3:43. 10.1038/s41746-020-0233-7
4. Liu D, Feng XL, Ahmed F, Shahid M, Guo J: Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Mental Health. 2022, 9:e27244. 10.2196/27244
5. Aldkheel A, Zhou L: Depression detection on social media: a classification framework and research challenges and opportunities. Journal of Healthcare Informatics Research. 2024, 8:88-120. 10.1007/s41666-023-00152-3
6. Salas-Zárate R, Alor-Hernández G, Salas-Zárate MDP, Paredes-Valverde MA, Bustos-López M, Sánchez-Cervantes JL: Detecting depression signs on social media: a systematic literature review. Healthcare. 2022, 10:291. 10.3390/healthcare10020291
7. Wongkoblap A, Vadillo MA, Curcin V: Researching mental health disorders in the era of social media: systematic review. Journal of Medical Internet Research. 2017, 19:e228. 10.2196/jmir.7215
8. Bokolo BG, Liu Q: Deep learning-based depression detection from social media: comparative evaluation of ML and transformer techniques. Electronics. 2023, 12:4396. 10.3390/electronics12214396
9. Shen G, Jia J, Nie L, et al.: Depression detection via harvesting social media: a multimodal dictionary learning solution. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. IJCAI, Melbourne, Australia; 2017. 3838-3844. 10.24963/ijcai.2017/536
10. Losada DE, Crestani F: A test collection for research on depression and language use. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer, Cham, Switzerland; 2016. 9822:28-39. 10.1007/978-3-319-44564-9_3
11. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, Goharian N: RSDD-Time: temporal annotation of self-reported mental health diagnoses. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. Association for Computational Linguistics, New Orleans, LA; 2018. 168-173. 10.18653/v1/W18-0618
12. Yates A, Cohan A, Goharian N: Depression and self-harm risk assessment in online forums. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark; 2017. 2968-2978. 10.18653/v1/D17-1322
13. Song H, You J, Chung JW, Park JC: Feature attention network: interpretable depression detection from social media. Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. Association for Computational Linguistics, Hong Kong; 2018. 613-622.

14. Nguyen T, Yates A, Zirikly A, Desmet B, Cohan A: Improving the generalizability of depression detection by leveraging clinical questionnaires. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland; 2022. 8446-8459. 10.18653/v1/2022.acl-long.578

15. Shen T, Jia J, Shen G, et al.: Cross-domain depression detection via harvesting social media. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI, Stockholm, Sweden; 2018. 1611-1617. 10.24963/ijcai.2018/223

16. Han S, Mao R, Cambria E: Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea; 2022. 94-104.

17. Martínez-Castaño R, Htait A, Azzopardi L, Moshfeghi Y: Early risk detection of self-harm and depression severity using BERT-based transformers: iLab at CLEF eRisk 2020. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, Thessaloniki, Greece; 2020. 2696:16.

18. Sivamanikandan S, Santhosh V, Sanjaykumar N, Jerin Mahibha C, Durairaj T: scubeMSEC@LT-EDI-ACL2022: detection of depression using transformer models. Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. Association for Computational Linguistics, Dublin, Ireland; 2022. 212-217. 10.18653/v1/2022.ltedi-1.29

19. Kulkarni H, MacAvaney S, Goharian N, Frieder O: TBD3: a thresholding-based dynamic depression detection from social media for low-resource users. Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France; 2022. 2157-2165.

20. Kermani A, Perez-Rosas V, Metsis V: A systematic evaluation of LLM strategies for mental health text analysis: fine-tuning vs. prompt engineering vs. RAG. Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025). Association for Computational Linguistics, Albuquerque, NM; 2025. 172-180. 10.18653/v1/2025.clpsych-1.14

21. Wang Y, Inkpen D, Kirinde Gamaarachchige P: Explainable depression detection using large language models on social media data. Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024). Association for Computational Linguistics, St. Julians, Malta; 2024. 108-126. 10.18653/v1/2024.clpsych-1.8

22. Jeon H, Yoo D, Lee D, Son S, Kim S, Han J: A dual-prompting for interpretable mental health language models. Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024). Association for Computational Linguistics, St. Julians, Malta; 2024. 247-255. 10.18653/v1/2024.clpsych-1.24

23. Kuzmin G, Strepetov P, Stankevich M, Chudova N, Shelmanov A, Smirnov I: Exploring large language models for detecting mental disorders. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China; 2025. 34535-34559. 10.18653/v1/2025.emnlp-main.1752

24. Lan X, Han Z, Cheng Y, Sheng L, Feng J, Gao C, Li Y: Depression detection on social media with large language models. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Suzhou, China; 2025. 2155-2171. 10.18653/v1/2025.emnlp-industry.151

25. Stamou V, Mikros G, Markopoulos G, Varlokosta S: Establishing control corpora for depression detection in Modern Greek: methodological insights. Proceedings of the Fifth Workshop on Resources and ProcessIng of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments @LREC-COLING 2024. ELRA and ICCL, Torino, Italy; 2024. 68-76.


**How to cite this article:**
Kumari P, Jijja A (November 18, 2026) Term Frequency-Inverse Document Frequency (TF-IDF)-Based Machine Learning and a Zero-Shot Fine-Tuned Language Net-Text-to-Text Transfer Transformer (FLAN-T5) Baseline for Depression Risk Detection from Text. Cureus J Comput Sci 3 : es44389-026-00378-3. DOI https://doi.org/10.7759/s44389-026-00378-3 