Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Depression Detection from Text using Machine Learning

Punam Sawale, Vihaan Khare, Shivam Vishwakarma, Sarthak Viche,
Vasdev Khajuria, Anand Wadekar, Soham Vetal
Vishwakarma Institute of technology,
Pune, India

***Abstract*—The growth of online communication creates an attractive opportunity for early and scalable detection of mental health risks, such as depression. However, the challenge of creating machine learning models that are accurate and generalizable across different text domains (e.g., clinical forum and social media), remains a challenge. This work demonstrates a thorough, reproducible framework for detection of risk of depression from online text. We used three different publicly available datasets: Suicide Watch, Mental Health Corpus, and Mental Health/Data, with a serious space used to preprocess the text, normalize the dataset by balancing the label, and rendering a poor quality dataset through stratified subsampling. Our approach included multi-dimensional feature engineering through the construction of a strong feature set, which included linguistic/readability measures (e.g., Flesch), conventional sentiment estimates (VADER, TextBlob), and advanced transformer-based sentiment (RoBERTa, DistilBERT). We conducted systematic training and evaluation of a number of supervised classifiers, including Logistic Regression, SVM, Random Forest, and Gradient Boosting. We used two-protocols for reliable model evaluation: within domain reliability through stratified 5-fold cross validation and cross-domain transfer testing for generalizability. Since our model evaluation included accuracy, precision, recall, F1-score, and ROC-AUC, we provide model interpretability through feature importance that reflects the most predictive signals.**

Keywords - Depression Detection, Machine learning, Natural Language Processing, Feature Engineering, Sentiment Analysis, Random Forest

I. INTRODUCTION

Mental health illnesses, especially major depressive disorder, are a substantial global health burden affecting millions of people and their accompanying social and financial costs. Concurrently, the digital age has radically transformed how we connect with others. Online sites, social media (such as Reddit), and specific health-related websites have become the main scenarios for individuals to articulate personal emotions, share experiences, and connect with others. This extensive and ever-growing amount of user-generated text provides an unrivaled, non-invasive opportunity for building automated, scalable health screening tools.

Although many prior studies have shown that machine learning (ML) and natural language processing (NLP) can be effective for depression identification, a major and ongoing challenge is model generalizability. Most existing models are trained on and validated against a single, homogenous dataset (e.g., either clinical-style text or social media posts). As a result, a model trained on clinical-oriented statements can typically fail to perform well when used on the more noisy, slang-laden, and "in-the-wild" language on social media and conversely. This lack of robustness and inability to generalize across text domains presents a considerable barrier to the practical, real-world applicability of these models for broad-based screening.

In order to tackle this problem, we propose a complete and reproducible machine learning framework for detecting depression risk from text. The main goal of our research is to create and experimentally evaluate models that perform well in their source domain and generalize across other text domains. We utilize three distinct publicly available datasets: Suicide Watch, the Mental Health Corpus, and the Mental Health/Data collection. Our approach is a multi-dimensional feature engineering approach, which generates a rich feature set incorporating: (1) traditional linguistic and readability measures, (2) classical sentiment analysis, via VADER and TextBlob, and (3) more contemporary transformer-based sentiment scores, via RoBERTa and DistilBERT.

We conduct a systematic training and comparison of various supervised classifiers, including baseline classifiers such as Logistic Regression and SVM, and ensemble classifiers such as Random Forest and Gradient Boosting. In particular, our evaluation methodology is two-part: we first establish the baseline classifier performance via 5-fold stratified cross-validation on each dataset (intra-domain). Then, we perform an extensive cross-domain transfer analysis and explicitly evaluate the generalizability of the models.

The main contributions of this work are:

We implemented a systematic preprocessing and harmonization pipeline for three different publicly available mental health datasets.

We developed a multi-dimensional feature-engineering strategy that combines linguistic, traditional sentiment, readability, and modern transformer-based sentiment features into a single robust input into the model.

We conducted a rigorous dual-protocol evaluation approach, with both stratified 5-fold crossvalidation and a broad cross-domain evaluation so that the generalizability of the models is a realistic measure.

We performed a comparative analysis of multiple classifiers to determine the most robust models, and an interpretability analysis to understand the most predictive features pertinent to the task.

The rest of the paper is structured in the following way: Section 2 outlines previous work in depression classification from text. Following this, Section 3 describes our methods, analyzing data collection, pre-processing, feature engineering, and experimental design. Section 4 displays results from our intra-domain and cross-domain experiments. Finally, Section 5 discusses implications, acknowledges limitations, and concludes.

II. MATERIALS AND METHODS

This section provides an overview of the datasets, preprocessingpipeline, feature engineering methods, and evaluation protocols applied to build and evaluate models for detecting depression.

A. Dataset Collection and Description

To promote robustness and test generalizability, this study draws on three publicly available and anonymized English-language datasets that each differ from one another :

Suicide Watch: This dataset is more prominently known as a random collection of posts taken from Reddit, and are sourced from subreddits related to mental health. Each post has been labeled by various levels of suicidality, and we mapped the suicidality levels to binary depression-risk classifications

Mental Health Corpus: This represents a clinically - oriented Dataset , which contains statements labelled as "Depression" or "Normal." The corpus indicates a more formal linguistic style than social media..

Mental Health/Data: Offers a general dataset of online posts (to Reddit) which are labelled as "depressed" or "not depressed," providing us with a wide-ranging diversity of informal and topical text.

All datasets were imported so that their labels and column names are all standardized in the pipeline as a "text" field and binary "target" variable (0 for not depressed, 1 for depressed.)

B. Data Preprocessing and Cleaning

Each dataset underwent a thorough preprocessing pipeline that standardized the text and prepared it for features- extraction. The steps were executed in sequence and included the following:

Normalization: All text was converted to lowercase.

Cleaning: All punctuation, numerical digits, and special characters were removed to minimize text noise

Deduplication: All identical duplicate posts contained in each dataset were removed.

Filtering: Posts were filtered to ensure informative text and to respect the input limits for transformer-models. We kept all texts with a word count of at least 5 (to remove trivial texts) and up to 512 (the approximate token limit for BERT-family models).

Handling Missing Values: All rows with empty text fields after cleaning were eliminated.

C. Class Rebalancing

Initial exploratory analysis revealed class imbalances across some of the datasets. To prevent models from developing bias toward the majority class and to allow for reasonable evaluations to be carried out, we used a stratified subsampling approach across all datasets. The approach consisted of randomly subsampling the majority class to the same number of instances as the minority class, capped at some maximum of 5000 instances. This provided perfectly balanced data outputs if the fewest amount of instances was above 5000. The final dataframes were shuffled to ensure a random order.

D. Feature Engineering

To create a full understanding of the text, we developed a multi-dimensional feature set across three main areas. All features were combined into a single feature vector for each post

Linguistic and Readability Features: These features capture the structure and stylistic characteristics of the text. They include:Word count,Character count, word length,Sentence count,Punctuation count,Flesch Reading Ease score,Flesch-Kincaid Grade Level.

Sentiment Features: These features quantify the emotional tone of the text. We used:

VADER (Valence Aware Dictionary and Sentiment Reasoner): The 'compound' score was used as a continuous measure of sentiment intensity.

TextBlob: We extracted both polarity (a continuous score from -1.0 to 1.0) and subjectivity (a continuous score from 0.0 to 1.0).

Transformer-Based Sentiment Features To measure deeper contextual sentiment, we used pre-trained transformer models as feature extractors:

RoBERTa-based Sentiment: A fine-tuned RoBERTa model was employed to classify text into 'positive', 'neutral', or 'negative' and crucially provide the raw confidence score for its prediction.

DistilBERT-based Sentiment: A fine-tuned DistilBERT model provided a secondary set of sentiment labels and confidence scores.

E. Feature Selection

To enhance the interpretability of our models, to decrease training time, and to decrease the risk of overfitting, we examined feature importances. A Random Forest was trained on each entirety of features in the datasets and used to obtain a measure of feature importance (e.g. Gini importance). We also examined correlation heatmaps to assess inter-feature relationships and feature-target correlations. The top 'N' most predictive features were then determined for final modelling training in terms of predictive power (e.g. word count, RoBERTa sentiment score, VADER compound)

F. Model Development and Evaluation

A set of supervised learning models were assessed to determine their performance in this task. The models chosen for this purpose were:Logistic Regression,Support Vector Machine (SVM) with an RBF kernel,Random Forest,Gradient Boosting,Extra Trees Classifier.

G. Evaluation Protocols

We applied a rigorous dual-protocol evaluation framework to provide both reliability and generalizability to the model

Intra-Domain Evaluation: To evaluate model performance within a single domain, we applied a stratified 5-fold cross-validation approach on each of the three balanced datasets. Each fold maintains the 50/50 class distribution. The models were trained and validated five times, and we report the mean performance.

Cross-Domain Evaluation: To test the generalizability of the models, we performed a cross-domain transfer analysis. Models were trained on the entire balanced training data of one dataset (Suicide Watch) and then tested on the entire balanced test data of a separate dataset (Mental Health Corpus).

This "train-on-A, test-on-B" method captures a model's learning patterns and transfer efficacy to an entirely different linguistic domain. For all experiments, model performance was quantified using Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Since our datasets were balanced, we emphasize the F1-Score as a robust measure of performance, while the ROC-AUC approximation provides an estimate of the model's ability to discriminate across all classification threshold.

H. Ethical Considerations

All data used in the project come from existing publicly available anonymized sources. No new data was collected from human subjects, nor was any personally identifiable information (PII) used or kept. The project is in full compliance with all ethical stipulations for public data and from the platforms' terms of service. All scripts and data splits are saved for the sake of thorough reproducibility.

III. RESULTS AND DISCUSSION

This section presents the findings from our experimental pipeline, beginning with an analysis of the data and features, followed by the intra-domain and cross-domain model performance.

A. Exploratory Data Analysis (EDA)

Exploratory Data Analysis [see Figure 1 and Appendix A] confirmed there was meaningful linguistic variation across these datasets. The Mental Health Corpus (MHC) had the shortest average posts length (42.5 words) and a stance-based, clinical character. On the other hand, the Suicide Watch (SW) dataset had the longest posts, which were the most narrative-like posts (avg. 248.7 words). The Mental Health/Data (MHD) dataset was in between at (avg. 91.3 words).

A heatmap displaying the correlation between the features and the target [see Figure 2] yielded some initial insights into the predictive signals. The RoBERTa based sentiment score (r = 0.58) and the VADER compound (r = 0.51), both based on textual features, demonstrated the strongest positive correlations with the 'depressed' target label. Flesch Reading Ease also demonstrated a notable negative correlation (r = -0.34), indicating that text corresponding with heightened risk of depression was, on average, less complex and easier to read.

Discussion: The considerable dissimilarity in post length and structure coupled with an initial strong signal from sentiment features sets up our main hypotheses. The data variability indicates that a model trained on one (e.g., short, formal MHC) will not perform well when tested on the other (e.g., long, informal SW).

B. Feature Importance

A Random Forest classifier was trained on the full feature set to rank feature importance. The results, visualized in **[Figure 3]**, were remarkably consistent across all three datasets.

The top five most predictive features, in order, were:

- RoBERTa-based Sentiment Score (Gini Importance: 0.24)
- VADER Compound Score (Gini Importance: 0.19)
- Word Count (Gini Importance: 0.11)
- Flesch Reading Ease (Gini Importance: 0.08)
- TextBlob Polarity (Gini Importance: 0.06)

Discussion: This is an important finding. It demonstrates that we're most effective using a mixture of deep contextual sentiment (RoBERTa), lexicon sentiment (VADER), and basic text structure (Word Count, Readability). This really supports our multi-dimensional feature engineering approach over a method which could use only sentiment or only embeddings.

C. Intra-Domain Model Performance

The performance of each model was first evaluated with 5-fold stratified cross-validation on each dataset independently. The results, which are summarized in Table 2, indicate the upper-bound performance of each model within the relevant domain.

TABLE I. INTRA-DOMAIN MODEL PERFORMANCE (AVERAGE F1-SCORE / ROC-AUC)

Model	**Suicide Watch (SW)**	**Mental Health Corpus (MHC)**	**Mental Health Data (MHD)**
Logistic Regression	0.81 / 0.87	0.79 / 0.85	0.82 / 0.88
SVM (RBF)	0.80 / 0.86	0.78 / 0.84	0.81 / 0.87
Random Forest	0.86 / 0.91	0.84 / 0.89	0.87 / 0.92
Gradient Boosting	**0.88 / 0.93**	**0.85 / 0.91**	**0.89 / 0.94**
Extra Trees	0.87 / 0.92	0.84 / 0.90	0.88 / 0.93

Discussion: In the context of a pristine, single-domain setting, the ensemble models (Gradient Boosting, Random Forest) consistently and favorably outperformed the baseline models. The best performing model, Gradient Boosting, achieved a high F1-score of 0.88 on Suicide Watch and 0.89 on MHD. High scores suggest all models are adept at learning predictive patterns when training data and test data are in the same linguistic domain. However, the aforementioned high performance may not be a true reflection of real-world readiness.

D. Cross-Domain Generalizability Analysis

This was the pivotal test of our objective. The winning model (Gradient Boosting) was trained on one dataset and tested on the others. The results, found in Table 3, show a significant and crucial drop in performance.

TABLE II. CROSS-DOMAIN PERFORMANCE (F1-SCORE) FOR GRADIENT BOOSTING MODEL

Training → Testing Dataset	*Suicide Watch (SW)*	*Mental Health Corpus (MHC)*	*Mental Health Data (MHD)*
Suicide Watch (SW)	***0.88***	*0.61*	*0.67*
Mental Health Corpus (MHC)	*0.57*	***0.85***	*0.60*
Mental Health Data (MHD)	*0.64*	*0.59*	***0.89***

Discussion: This table represents the main finding from this work. The drop in performance is severe and speaks to our contention regarding the domain shift issue. The model trained on MHC (F1 0.85) has lost its performance on SW posts by 33% (i.e., 0.57). Additionally, the SW model (F1 0.88) drops to 0.61 on MHC data.

This demonstrates that the models are not acquiring a global "language of depression." Instead, they overfit to the specific set of linguistic artifacts of the training domain (e.g., MHC's formality, SW's narrative style). This evidence calls into question the validity of models trained on one data source, and serves as a stark reminder of the absolute necessity of cross-domain testing.

E. Error Analysis

A qualitative analysis of the cross-domain errors (e.g., from the MHC ➔ SW test) revealed clear patterns:

False Negatives (Missed Risk): The MHC-trained model, which was trained on formal statements, did not flag risk in SW posts that were filled with slang, sarcasm, or "in-community" language (e.g., "I'm just going to yeet myself"). The model's sentiment analyzer misclassified these posts as neutral.

False Positives (Incorrectly Flagged): The SWtrained model, which learned to associate long, emotional, first person narratives with risk, erroneously flagged long, emotional, and non-depressive posts in other domains (e.g., an angry, detailed complaint regarding a product on MHD).

Discussion: The error analysis confirms *why* the models fail. They are brittle and lack the contextual understanding to differentiate between domain-specific style and genuine risk markers.

F. Limitations

While this study provides a robust framework, it has several limitations:

- **Proxy Labels:** The states of the dataset ("depressed," "suicidal," etc.) are not clinical diagnoses. They represent user-generated flairs or self-reports that are a noisy representation of a medical condition.
- **Text-Only Analysis:** Our model uses only the text of a single post; there is no context supplied that would be critical to making a prediction such as prior posts by the user, interactions by users in the community, or rate of posting, all of which are robust predictors.
- **Language and Demographics:** This study is focused on a dataset in the English language, which is skewed to North America and then Reddit users as a sample, so the results may not generalize to different language datasets or populations.
- **Static Datasets:** Online language is evolving quickly, and models trained on these static datasets from 2018-2020 will almost certainly experience performance issues as slang and other expressions change.

IV. CONCLUSION

This article has introduced a rigorous and replicable methodology for the design and evaluation of machine learning models to detect risk of depression in different online text types. The main focus of the past work has been to go beyond single dataset benchmarks and methodically assess the real-world question of model generalizability.

By utilizing a multi-angled feature engineering process that included linguistic, readability, traditional sentiment, and modern transformer-based sentiment features, we demonstrated that a hybrid feature model is highly predictive. Our intra-domain studies showed that if you train and test ensemble models like Gradient Boosting with the same domain data you can achieve high performances (with F1-scores as high as 0.89).

Nevertheless, our core and most consequential result stemmed from the cross-domain analysis. We have shown that there is a precipitous and consistent decline in performance (with decreases in F1 scoring by as much as 33%) when models are applied in not previously encountered domains. This performance point provides qualitative evidence that models are not deriving a common "language of depression" as a risk factor, but rather overfitting to the domain specific linguistic artifacts, vernacular, and unique post structures of the training data.

This finding sends a crucial warning to the domain: model performance, whether measured by correctness, F1 score, or accuracy, should not be relied on as an indicator of field usefulness when only tested on a single, homogenous dataset. Our error analysis shows that these models are still brittle and will struggle with sarcasm, jargon from the community, as well as other nuances of human expression.

This work is limited by using proxy labels and focuses only on text analysis, but it establishes an important baseline for robustness. Future research will have to focus on addressing the domain-shift problem, perhaps by using more advanced domain-adaptation methods, training on larger datasets, and adding more variation to the datasets, as well as including metadata about the user at the context level. This study, at the minimum, demonstrates that if we are to treat automated mental health screening tools as safe, ethical, and effective, then generalizability must be a design goal, not an afterthought.

REFERENCES AND FOOTNOTES

[1] Al-Lafe, O. et al. (2023). Depression Detection on Social Media: A Scoping Review. JMIR Medical Informatics, 11, p. e43888.

[2] Al-Tahan, Z.Z., Shehadeh, H.T.H. and Al-Tahan, M.A. (2021). Depression Detection From Social Media Text Using Deep Learning. 2021 International Conference on Innovations in E-Health and Information Technology (IEHIT), p. 1-5.

[3] Arora, A. and Arora, A. (2022). A review of text-based depression detection using machine and deep learning. Intelligent Systems with Applications, 16, p. 200221.

[4] Chandak, M.B. et al. Depression detection using ML. International Journal of Advanced Engineering and Management (IJAEM), 8(3).

[5] Hameed, S. et al. (2025). Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models. Frontiers in Artificial Intelligence, 8, p. 1627078.

[6] Kavitha, T. et al. (2023). Detection of Depression Using Machine Learning. International Journal of Research and Publication and Reviews, 6(2), p. 307-310.

[7] Mohammad, S.M. and Turney, P.D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence.

[8] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, p. 2825-2830.

[9] Pennebaker, J.W. (2011). The Secret Life of Pronouns: What Our Words Say About Us.

[10] Verma, R. et al. (2024). A systematic review of machine learning models for depression detection using social media data: Current trends, limitations and future directions. Addictive Behaviors Reports, 19, p. 100561.