Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 21 May 2026
Revised 26 June 2026
Accepted 18 July 2026

Natural Resources for Human Health 2026, Vol. 6(7s): 1076–1091

KEYWORDS:
Natural language processing, depression detection, machine learning, text analysis, data mining.

Machine Learning based Depression Detection with NLP Applicationon Twitter Text Data

Priyanka Srivastava[1], Dr. Mohammad Suaib[2]

[1]Department of CSE, Integral University, Lucknow, India
priyanka.lecturer.gpl@gmail.com
[2]Department of CSE, Integral University, Lucknow, India
suaib@iul.ac.in

ABSTRACT: Depression detection through social media analysis has gained increasing attention due to the accessibility of large-scale user-generated content. This study proposes a hybrid sentiment analysis framework that integrates Latent Dirichlet Allocation (LDA), Valence Aware Dictionary and Sentiment Reasoner (VADER), and Term Frequency–Inverse Document Frequency (TF-IDF) features to capture thematic, emotional, and statistical representations of text. The dataset, sourced from Twitter, was preprocessed with advanced filtering techniques to remove noise and irrelevant tokens, producing a robust feature space. ML classifiers are evaluated like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), logistic regression, and random forest, and SVM achieved the best accuracy (95.1%), precision (98.8%), recall (89.4%), and F1-score (93.9%). The results show that the combination of probabilistic, lexicon-based and statistical approaches improves classification results over the individual approaches and provides an interpretable, scalable framework for computational depression detection.

1. INTRODUCTION

Depression is one of the most serious mental health problems in the world today. Millions of people are suffering from it in their personal, social and professional lives. However, traditional methods are mainly based on clinical observation and self-reporting, which are often constrained by underreporting and stigma, thus early recognition is crucial for timely intervention and effective treatment. The recent widespread adoption of social media platforms such as Twitter has provided a way to study behavioral and linguistic cues that may be indicative of depressive tendencies, providing a new way for computational methods to be used in healthcare monitoring.

The recent development in sentiment and emotion analysis has been greatly improved using transformer-based models, graph neural networks and hybrid multimodal approaches. However, challenges remain in dealing with noisy, code-mixed and domain specific data, and developing interpretable models that can be reliably applied in sensitive healthcare contexts. To fill these gaps, this study proposes a hybrid sentiment analysis framework that combines Latent Dirichlet Allocation (LDA) for topic modelling, Valence Aware Dictionary and Sentiment Reasoner (VADER) for lexicon-based polarity scores, and TF-IDF for statistical weighting. By integrating these complementary approaches into the framework, it enhances the identification of depression-related patterns within tweets, yielding improved robustness and interpretability.

The significance of this work is to extend the potential of computational mental health monitoring by applying natural language processing to detect early signs of depression. The proposed hybrid framework offers high accuracy and interpretability as opposed to black-box deep learning systems, thus bridging the gap between clinical utility and machine learning research. It also highlights how social media can be used responsibly as a large-scale, non-invasive data source to help mental health initiatives and public health policy.

The main contributions of this paper are the design of an integrated sentiment analysis pipeline which combines statistical, probabilistic and lexicon-based features for detecting depression in tweets. The results indicate that the combination of LDA topic vectors, VADER sentiment scores and TF-IDF features yields better classification accuracy than the sole use of these techniques. The Support Vector Machine (SVM) provides an accuracy of 95.1%. Besides, the work suggests a systematic

approach of preprocessing and feature selection that alleviates the noise, allows the model to be trained efficiently, and enhances the robustness in real-world social media settings.

The proposed framework can be integrated into real-time social media monitoring systems to identify users with potential signs of depression and thus may help mental health professionals, carers, and policy makers with early intervention strategies. The methodology can be extended beyond healthcare to domains such as online public opinion tracking, customer feedback analysis and behavioral analytics where nuanced understanding of sentiment is essential. Notably, the model's focus on interpretability makes it a good candidate for application in sensitive decision-making contexts where transparency and ethical accountability are important.

2. RELATED WORK

Recent research in literature review on sentiment analysis, emotion detection, and opinion extraction has made significant strides with models such as ASTE-Transformer, PASTEL, and fused graph neural networks, leading to improved precision in extracting aspect–opinion–sentiment triplets. However, most approaches rely heavily on large annotated datasets, which are expensive and time consuming to produce. Cross-lingual and code-mixed data are still challenging as the models often do not generalize across multiple languages and domains. Accurately matching multiple aspects with the corresponding opinions in complex sentences, dealing with noisy real-world data, and efficiently integrating multimodal cues such as text, images, and video remain open problems. Moreover, many high-performing models are black-box, making them difficult to interpret and of limited practical use in sensitive domains such as healthcare and public opinion analysis.

Naglik (2024) proposed an ASTE-Transformer based on a dependency-aware transformer model for Aspect-Sentiment Triplet Extraction. It jointly learns aspect, opinion and sentiment relations instead of following a pipeline, thus reducing error propagation and achieving superior F1 scores on SemEval benchmarks [1]. Based on this notion of joint reasoning, Bodke (2025) proposed PASTEL, a polarity-aware framework that first generates candidate triplets and then uses a large language model as a judge to verify the polarity consistency, which substantially improves the precision on noisy social media datasets [2]. Peng (2024) performed triplet extraction improvement by deep relationship enhancement network, it refined aspect–opinion interactions iteratively to get a better disambiguate overlapping pairs to achieve a strong performance [3]. To complement these models a syntax–semantics graph convolutional network proposed that combines syntactic dependence and similarity graphs of semantic for propagating contextual signals. It gave better robustness for different sentence structures [4]. AWA-GCN, developed (2024) by adopting a language-specific graph features for solving triplet and quadruple extraction tasks. It highlights the benefits of tailored architectures for morphologically complicated languages [5]. In 2025, a research work extended to cross-lingual settings and presented a multi-domain and multilingual quadruple-extraction model. In this work authors proposed domain and language adapters to enable strong performance on different languages and application areas [6]. For better match aspects and opinions, Piao (2024) applied fused graph neural networks with a biaffine attention mechanism to compute accurate pairwise scores and reduce wrong matches in multi-aspect sentences [7]. In 2024, a hybrid Vision Transformer designed with Temporal Convolution (ViTCN) to advance multimodal emotion detection with the capacity of learning spatial and temporal cues for multi-emotion recognition in video streams, attaining competitive results on benchmark video datasets [8]. Rasool (2025) proposed the nBERT model, that was adapted to psychotherapy transcripts. the result shows that fine-tuning within a domain may show effectiveness to identify fine-grained clinical emotions, despite the limited annotated data [9]. Rezapour (2024) conducted a comparative study of several transformer backbones for textual emotion classification and showed that explanation stability varied widely across attention and gradient-based interpretability methods [10]. For SemEval-2024 Task 10, we have described the ISDS-NLP system (Creangă, 2024) that exploits contextual transformers and multilingual augmentation to improve the recognition of emotions in conversational data [11]. An anonymous 2025 study used a CNN front-end with transformer encoders to identify emotions in social media posts with English-Hindi code-mixing, and tackled the challenges of tokenization and language-switching in a code-mixed setting [12]. Opinion extraction itself has also progressed beyond single-opinion settings. Zhao (2025) proposed a multi-target opinion-based model for word extraction by applying graph convolution and sequence labelling. It was applied for identification of opinion expressions associated with each target that produce improvement in coverage in complex reviews [13]. Zhu (2023) proposed a sentence-guided grid tagging scheme for conducting aspect–sentiment quadruple extraction. it worked for mapping tokens into a two-dimensional grid and enabling the joint decoding of aspect, opinion, category, and sentiment [14]. In the medical domain, an anonymous 2025 effort merged large language models with interpretable rule-based classifiers to extract explainable aspect–sentiment pairs from clinical notes satisfying stringent interpretability requirements

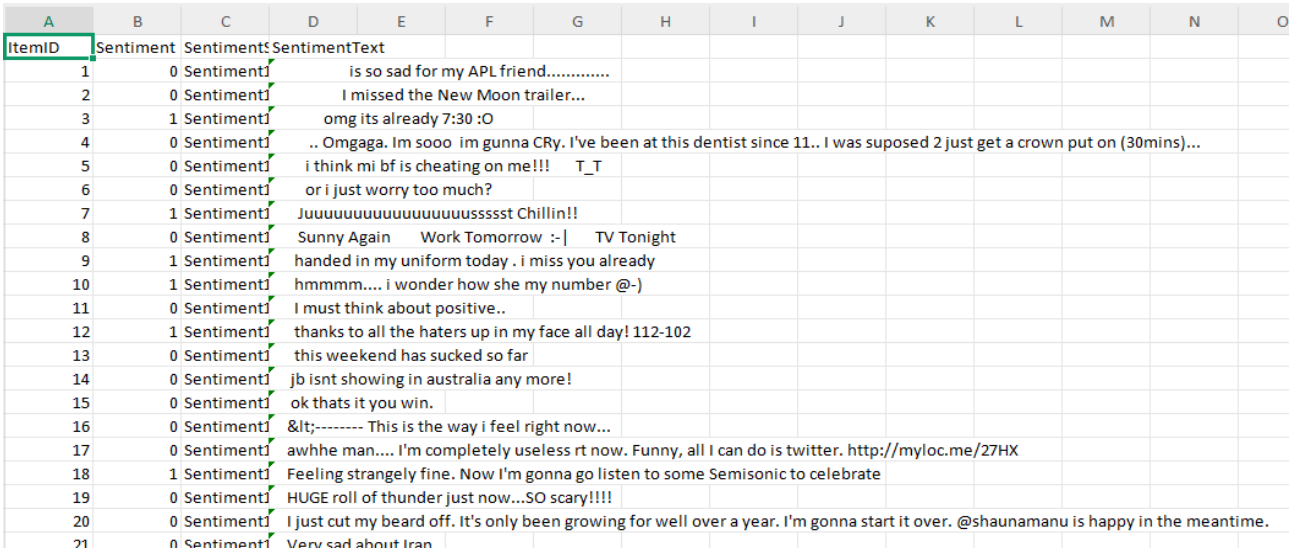
[15]. An NLPCC 2023 team investigated span-based extraction with contrastive learning, where negative sampling helped to sharpen the boundaries between correct and incorrect aspect – opinion pairs [16]. A RoBERTa–BiLSTM–Attention hybrid by unnamed authors (2025) proved to be a robust sequential model for analysing changing sentiments in social media streams for online public opinion analysis [17]. To address the shortage of Chinese training resources, Lu (2024) proposed an iterative weak-supervision approach to automatically build a large review dataset for triplet extraction and reduce the human annotation cost with competitive quality [18]. Visual emotion recognition is also still growing. Kalsum (2023) proposed a lightweight deep convolutional neural network for robust facial expression recognition under different illumination and demographic conditions. The proposed solution is a mobile friendly solution for edge deployment [19].

Table 1: Summarized literature review table

Ref. No.	Author(s) (Year)	Approach	Pros	Cons
[1]	Naglik & Lango (2024)	AspectSentiment Triplet Extraction (ASTE)	Reduces error propagation; high F1 score	Requires large datasets and high computational cost
[2]	Bodke, Zhang & Lee (2025)	Polarity-aware triplet extraction (PASTEL)	Improves precision; handles noise	Dependent on LLM availability; high inference cost
[3]	Peng & Su (2024)	Deep relationship enhancement network	Better decision of overlapping aspect–opinion pairs	Complex training and sensitive to hyperparameters
[4]	Zhang, Xu, Gao & Tang (2025)	Syntax–Semantics Graph Convolutional Network	Combines syntactic and semantic cues; robust	High graph construction cost; struggle with long sentences
[5]	Haocheng Xi, et al.(2024)	AWA-GCN: graph CNN	Language-specific design; improves accuracy	Limited portability to non-Chinese languages
[6]	Çekinel (2025)	Multilingual, multimodal explainable approach	Explores cross-lingual sentiment	Application is exploratory; lacks large-scale evaluation
[7]	Piao & Zhang (2024)	Fused GNN with biaffine attention	Accurate aspect–opinion pairing; fewer mismatches	Requires heavy graph computations; careful feature engineering needed
[8]	Zakieldin, Khattab, et al. (2024)	ViTCN: Vision Transformer + temporal convolution	Captures spatial and temporal cues	High GPU demand; needs large datasets
[9]	Rasool, Khan & Ahmad (2025)	nBERT: domain-specific BERT	Handles fine-grained emotions	Limited to psychotherapy domain only
[10]	Rezapour (2024)	Transformer with interpretability	Provides insights on explanation stability	Lacks novelty; performance varies with dataset
[11]	Creangă & Dinu (2024)	Transformer-based SemEval system (ISDS-NLP)	Effective multilingual augmentation	Dependent on quality of augmentation resources
[12]	Patankar & Phadke (2025)	CNN–Transformer	Handles code-switching; robust	Dataset-specific tokenization limits transfer to other languages
[13]	Zhao, Li & Wang	Multi-target opinion	Captures multiple opinions	Struggles with sparse/noisy

	(2025)	extraction by GCN		opinion expressions
[14]	Zhu, Chen & Liu (2023)	Sentence-guided grid tag	Joint decoding of aspect, opinion, category	Grid tag increases model complexity and memory
[15]	Zhang, Wen, Zhu, Li & Wang (2025)	LLM + interpretable model for explainable ABSA	Balances accuracy with interpretability	Dependent on medical domain knowledge; limited scalability
[16]	Yang, Dai, Li & Xue (2023)	Span-based pairwise aspect–opinion extraction with contrastive learning	Improves boundary detection; reduces ambiguous pairings	Requires careful negative sample selection; sensitive to span enumeration
[17]	Deng & Liu (2025)	RoBERTa–BiLSTM–Attention	Strong sequential modeling	Resource-intensive training; overfit to small datasets
[18]	Lu, Chen & Zhao (2024)	Iterative weak supervision	Reduces annotation cost; generates large dataset	Quality depends on seed rules; propagate labeling bias
[19]	Kalsum, Rauf & Ali (2023)	Lightweight deep CNN	Mobile-friendly; robust	Less effective on subtle/ambiguous expressions
[20]	Venkatesan, & Prasad (2023)	DeepFace + downstream classifiers	High accuracy; practical implementation	Raises privacy/ethical concerns; sensitive

Finally, Venkatesan (2023) used DeepFace embeddings with downstream classifiers to detect human emotions from facial images, showing a practical pipeline for image-based emotion detection but also raising ethical concerns in cross-cultural interpretation [20].These review indicate the need for a robust and scalable framework to extract sentiment and emotion information across domains and languages. A system that can handle noisy and code-mixed data, integrate multimodal cues without prohibitive computational cost and provide interpretable results to support decision making. Such a model would bridge the gap between the high-performance academic methods and the real-world applicability by providing accurate, efficient and explainable sentiment and emotion analysis, and thus, can be used for practical deployment.



ItemID	Sentiment	SentimentS	SentimentText
1	0	Sentiment1	is so sad for my APL friend.....
2	0	Sentiment1	I missed the New Moon trailer...
3	1	Sentiment1	omg its already 7:30 :O
4	0	Sentiment1	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
5	0	Sentiment1	i think mi bf is cheating on me!!! T_T
6	0	Sentiment1	or i just worry too much?
7	1	Sentiment1	Juuuuuuuuuuuuuuuuussssst Chillin!!
8	0	Sentiment1	Sunny Again Work Tomorrow :-| TV Tonight
9	1	Sentiment1	handed in my uniform today . i miss you already
10	1	Sentiment1	hmmmm.... i wonder how she my number @-)
11	0	Sentiment1	I must think about positive..
12	1	Sentiment1	thanks to all the haters up in my face all day! 112-102
13	0	Sentiment1	this weekend has sucked so far
14	0	Sentiment1	jb isnt showing in australia any more!
15	0	Sentiment1	ok thats it you win.
16	0	Sentiment1	<-------- This is the way i feel right now...
17	0	Sentiment1	awhhe man.... I'm completely useless rt now. Funny, all I can do is twitter. http://myloc.me/27HX
18	1	Sentiment1	Feeling strangely fine. Now I'm gonna go listen to some Semisonic to celebrate
19	0	Sentiment1	HUGE roll of thunder just now...SO scary!!!!
20	0	Sentiment1	I just cut my beard off. It's only been growing for well over a year. I'm gonna start it over. @shaunamanu is happy in the meantime.
21	0	Sentiment1	Very sad about Iran

Figure 1. Screen shot of CSV file view of twitter sentiment labelled dataset used under this article for depression detection model.

Natural Resources for Human Health
eISSN: 2583-1194

3. METHODOLOGY

The proposed algorithm for Sentiment analysis involves the combination of multiple feature extraction techniques to form a comprehensive feature set for machine learning classification. We first preprocess the raw text corpus by means of tokenization, part-of-speech tagging, lemmatization, punctuation removal, stop word removal and filtering of very short or long words. Irrelevant tokens, such as HTML entities and special characters, are also removed from the specific domain to reduce noise and improve the quality of the textual representation. To represent the distribution of term frequencies in the corpus, a bag-of-words model is built after the pre-processing. We extract features in three complementary dimensions: (i) we apply Latent Dirichlet Allocation (LDA) to extract topic distributions, which provides an insight into the latent thematic structure of the documents, (ii) we perform lexicon-based sentiment analysis with VADER to yield polarity scores, capturing the emotional content of the text, and (iii) we compute Term Frequency-Inverse Document Frequency (TF-IDF) on the top-K most frequent terms, highlighting words that are frequent in a document but rare across the corpus. Then, these features are concatenated to form a single feature matrix, which can be used to train supervised machine learning models such as Support Vector Machines (SVM), to classify sentiment or depression related labels. The algorithm provides a rich representation of textual data in multiple aspects, where syntactic, semantic and emotional information is combined to provide a robust sentiment analysis.

3.1 Data Used

The dataset used for model development and analysis is twitter text records under different sentiment labels. The data is publicly free access fromhttps://www.kaggle.com/datasets/ywang311/twitter-sentiment?resource=download. The data records are known by name "twitter_sentiment", it has size of 157MB and holds total 1048575 text (statements) each text message is labelled with respective Label: 0,1 ➔[Depressed, Not depressed]. The total number of labels of Depressed Cases i.e. 0 is 554f70 (52%). This dataset development is dedicated to the depression detection by using Tweets of the users. There are two kind of tweets that are used in this database: random tweets that do not indicate depression and tweets that shows the user may have the depression as shown in figure 1.

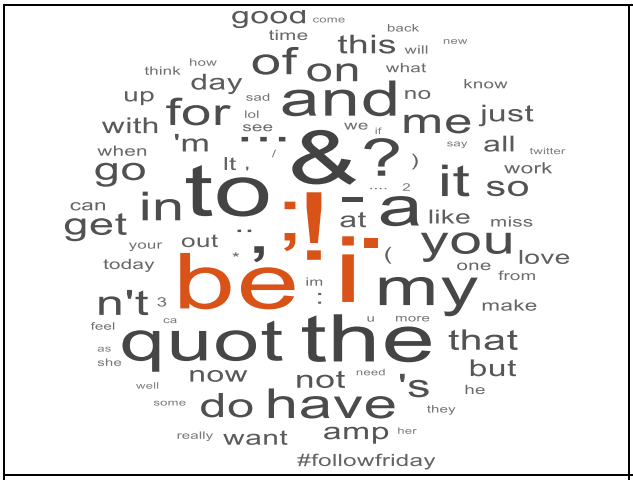


Figure 2a. Unique words in Raw data.

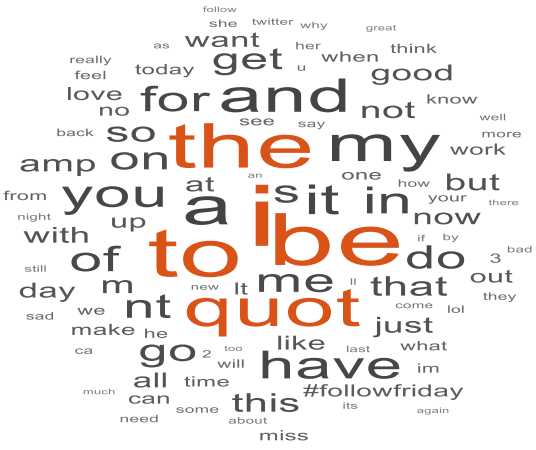


Figure 2b. Unique words after removing punctuation from raw data.

Natural Resources for Human Health
eISSN: 2583-1194

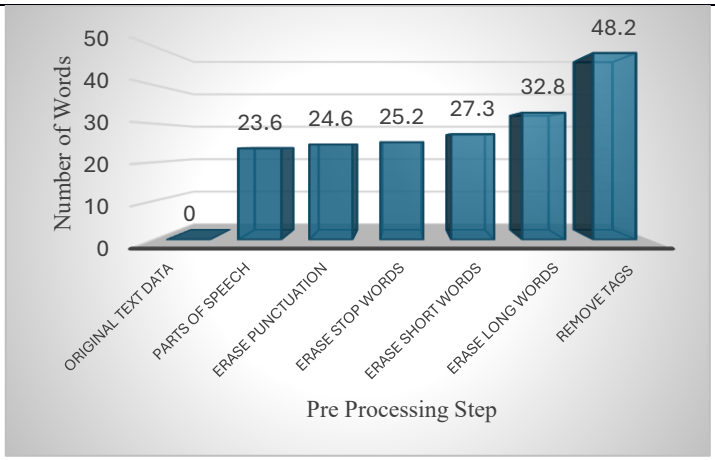
Figure 2c. Unique words after removing stop words from raw data.

Figure 2d. Unique words after removing short/long words from raw data.

3.2 Data Preprocessing

Table 2. Reduction of unwanted words during preprocessing steps

Pre Processing Step	Number of Words	Percent Reduction
Original text data	22434	0
Parts of speech	17144	23.6
Erase punctuation	16923	24.6
Erase stop words	16787	25.2
Erase short words	16303	27.3
Erase long words	15085	32.8
Remove tags	11630	48.2



In the sequence-wise representation of figure 2a to 2d it is shown that symbols like" ; . "!" are taking maximum number of text data. After removing this punctuation, the highest number of words in count is "i, to, be, quot" that are large in count but do not reflect any instances of depression. On removal of the stop words (figure 2c) it may be observed that the 'red' colored large font words are "quot, get, go, nt, s,2, 3, u". The words with <=2 letter length or words with length higher than 100 letters like http address etc. are useless hence the very short and long words are removed and the final cleaned data along with original rough data related cloud of words are shown in figure 3.

It may be observed that finally clean words are meaningful like "good, like, feel, love". These words if considered under final dataset may help in generating such feature space that is helpful for developing machine learning model. After removing/erasing unwanted words the bag of cleaned words is finally retained. The bag of words is the structure file that consist of list of unique words known as vocabulary, the count of word in a text (twitter comment). This final bag of words is used to calculate the number of repetitions of each unique word. The top frequently used 100 words are sorted out for both the under label 0/1. This reflects the collection of words that are more used by normal cases and cases suffering from depression. These data is used to extract the features for developing the ML models as described in the upcoming subsections.

Natural Resources for Human Health
eISSN: 2583-1194

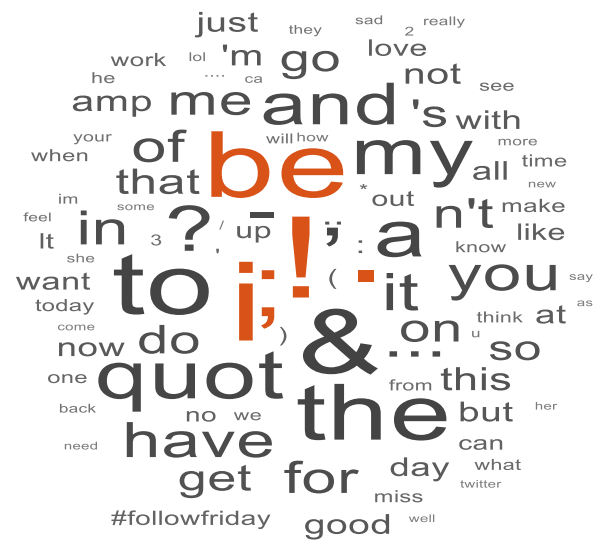


Figure 3a. Raw Data

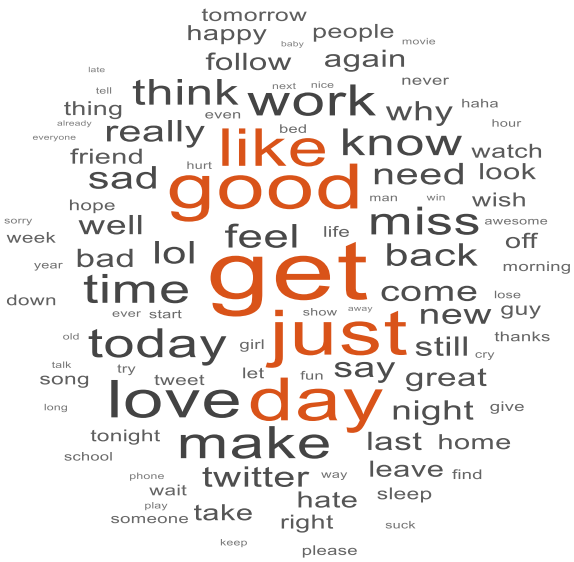


Figure 3b. Final Cleaned data

3.3 Feature extraction

The cleaned data after removal of unwanted text that has no correlation with the sentiments is used to generate three types of features known as: (a) LDA, (b) TF-IDF and (c) VADER. The details of these features and the steps to evaluate them from specified twitter sentiment cleaned dataset is explained first.

(a) **Latent Dirichlet Allocation (LDA):**LDA is a generative probabilistic model applied for the task of discovering topics from text data. Each document is represented as a mixture of topics, where each topic is a distribution over words. LDA is used to discover thematic structures in text, so that documents with similar distributions of topics can be represented as short vectors of topic probabilities.In Latent Dirichlet Allocation, Dirichlet distribution generates topic proportions in documents and word proportions in topics.For Kcategories:

$$Dir(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$$

θ_i: Probability of category i, α_i= Prior concentration parameter, $\Gamma(\cdot)$: Gamma function

The parameter αcontrols sparsity where small α i.e. $\alpha<1$ produces sparse distributions and large value of it produces uniform distribution.

Multinomial Distribution models: It shows the count of outcomes across multiple categories.For Kcategories:

$$Multinomial(x) = P(x_1, \ldots, x_K) = \frac{M!}{x_1! x_2! \ldots x_K!} \prod_{i=1}^{K} p_i^{x_i}$$

M: Total trials, x_i: Count of category i, p: Probability of category i.

In this work LDA is calculated by Gibb's sampling formulae as the probability that word $i$belongs to topic k:

$$P(z_i = k \mid z_{-i}, w_i) \propto \frac{n_{dk}^{-i} + \alpha}{\sum_k n_{dk}^{-i} + K\alpha} \times \frac{n_{kw}^{-i} + \beta}{\sum_w n_{kw}^{-i} + V\beta}$$

Where:Document-topic distribution: $\theta_d \sim Dirichlet(\alpha)$

Generate word from topic: $w_{dn} \sim Multinomial(\phi_{z_{dn}})$

Where: ϕ_k= word distribution for topic k

(b) **VADER (Valence Aware Dictionary and Sentiment Reasoner):** VVADER is a lexicon-based sentiment analysis tool that measures the emotional meaning of the text . It provides sentiment scores on multiple dimensions including positive, negative, neutral and a composite score. VADER is especially suited for text data, in particular short texts, social media posts or informal writing, and it gives numerical features that capture the emotional content of each document. For valence score v_i of i^{th} word:

S: Base valence score: Calculation total sentiment strength from all sentiment-bearing words: $S = \sum_{i=1}^{N} v_i$
B: Booster adjustment increases/decreases sentiment intensity using modifier words;$S' = S + B; B = \pm 0.293$

N: Negation adjustment reverses or weakens sentiment polarity by negation terms:$S' = S \times N\ ; N = -0.74$

C: Capitalization emphasis increases sentiment strength for uppercase words:$S' = S + C$ and$C = 0.733$

EP: Exclamation emphasis increases emotional intensity bycount of exclamation marks: $EP = \min(N_!, 4) \times 0.292$

QP: Question mark emphasis adjusts sentiment intensity according to question marks: $QP = \begin{cases} 0.18 \times N_?, & N_? \leq 3 \\ 0.96, & N_? > 3 \end{cases}$

Contrastive conjunction rule reduces/emphasizes sentiment before/after "but": $S_{before} = 0.5S$ *and* $S_{after} = 1.5S$

Total valence score combines all adjusted sentiment contributions as a raw score: $x = \sum_{i=1}^{N} v_i + B + C + EP + QP$
Compound score normalization converts raw sentiment score in between −1 and +1:$compound = \frac{x}{\sqrt{x^2+\alpha}}; \alpha = 15$

Positive/Negtive/Neutral sentiment score measures the proportion of positive sentiment in the text.

$Positive = \frac{\sum positive\ valence}{Total\ valence}$; $Negative = \frac{\sum negative\ valence}{Total\ valence}$; $Neutral = \frac{Neutral\ Tokens}{Total\ Tokens}$
Final constraint ensures that all sentiment proportions sum to unity: $Positive + Negative + Neutral = 1$

(c) **TF-IDF (Term Frequency-Inverse Document Frequency)**: TF-IDF is a statistical measure that evaluates the importance of a word in a document relative to the entire corpus. TF is used to capture that how often a word appears in a document, while IDF down weights terms that are ued very commonly across many documents. In the context of text data, TF-IDF is highlighting distinctive terms that perform discrimination in between documents, and produces weighted features suitable forML models.In text data, TF-IDF highlights distinctive terms that can discriminate between documents, producing weighted features suitable for machine learning models.

TF measures how frequently a term appears in a document: $TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$
IDF measures how important a term is across all documents:$IDF(t) = \log\left(\frac{N}{df_t}\right)$
Smoothed IDF avoids division by zero and stabilizes the weighting: $IDF(t) = \log\left(\frac{1+N}{1+df_t}\right) + 1$
TF-IDF score calculates the weighted importance of a term in a document: $TFIDF(t,d) = TF(t,d) \times IDF(t)$

3.4 Machine Learning (ML) model development:

The features vectors derived for each twitter comments are concatenated to form a feature matrix of three column and considered as the Input dataset 'X' and the respective label attached with each twitter comment is taken as the Output data 'Y'. The dataset is subdivided in training and testing data subset at the ratio of 70% training and 30% testing data taken out randomly. The training dataset {X_{train} and Y_{train}} is fed to the machine learning algorithm to develop decision models that has capability to classify the in text belong to depression state or not by using respective feature value. Four types of machine learning models are developed named as SVM, Logistic regression (LR), KNN and Random Forest (RF).

(a) Random Forest

Random Forest is a supervised machine learning algorithm for classification and regression tasks. It is based on the concept of ensemble learning technique that is employin multiple decision trees for improving prediction accuracy and reduce

overfitting. The algorithm was proposed by Leo Breiman in 2001 and it is widely used in many applications such as medical diagnosis, weather forecasting, image processing, fraud detection, and IoT-based systems.It builds many trees and aggregates them for prediction with more robust output than a single decision tree that may have high variance and lack of generalisation.

Random Forest starts with a training dataset that is represented as:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where x is input features and y represents target outputs. The algorithm builds several bootstrap samples of the original dataset by random sampling with replacement. A separate decision tree is created for each bootstrap data set. Instead of considering all features available at a node, a random subset of features is chosen during tree formation. The randomness leads to less correlation between trees and better overall robustness of the model.

In each tree the best split is determined using impurity measures such as Gini Index:

$$Gini(t) = 1 - \sum_{i=1}^{c} p_i^2$$

where $p_i$is the probability of class i. Lower impurity values mean node purity is better. Once a number of trees are built, the predictions of the trees are combined by Random Forest via majority voting for classification or by averaging for regression:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x)$$

where *B* is the total number of trees and T_b (x)is the prediction of the b[th]tree. Random Forest combines multiple decision trees to decrease variance, improve accuracy and provide strong generalization performance on complicated data.

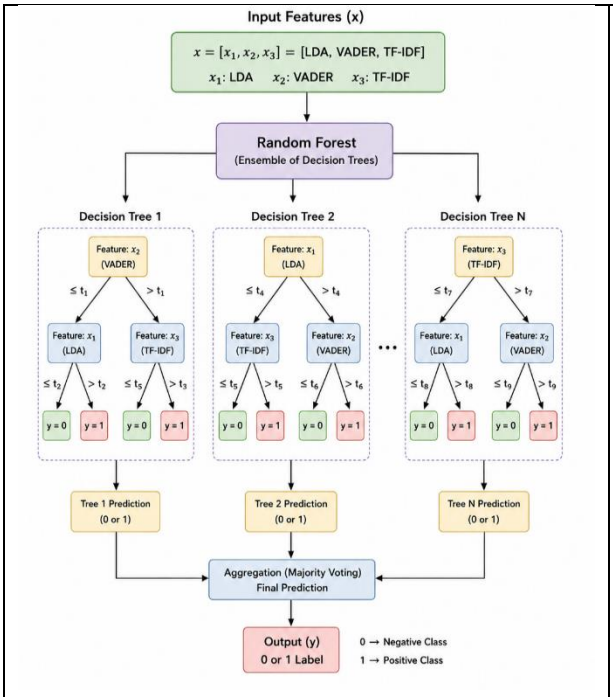


Figure 4a. Working model of random forest as ML classifier.

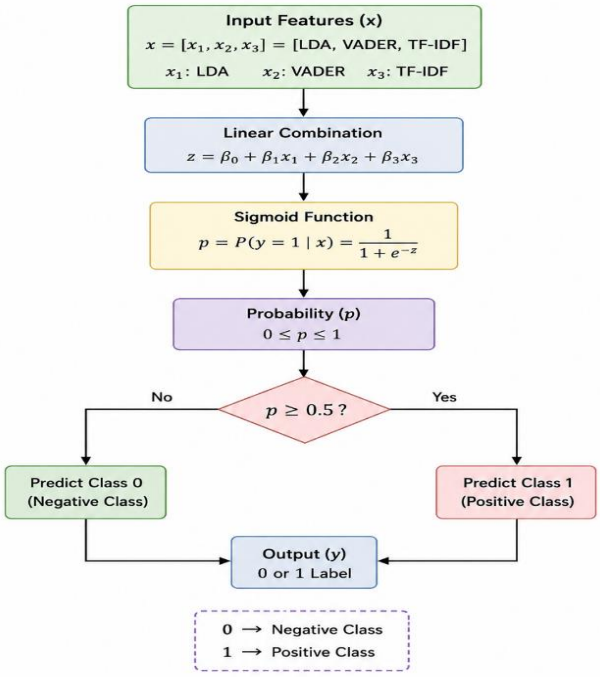


Figure 4b. Working model of Logistic regression as ML classifier.

(b) Logistic Regression

Logistic Regression is a supervised machine learning algorithm. This is used for the classification problems. It is useful to estimates the likelihood of the event of a class label outcome such as yes/no, true/false etc. Unlike linear regression, which

produces any continuous number, Logistic Regression employs the sigmoid function to make the output probability lie between 0 and 1. The algorithm is used for medical diagnosis, fraud detection, sentiment analysis and risk prediction due to its simplicity, efficiency and interpretability.

The mathematical model of LR starts with a linear combination of input features:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where $\beta_0$is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$are model coefficients, and $x_1, x_2, \ldots, x_n$represent input variables. The obtained value is passed through the sigmoid activation function to convert it into probability form:

$$P(y = 1) = \frac{1}{1 + e^{-z}}$$

The sigmoid function gives us values between 0 and 1. Probabilities greater than 0.5 are classified as class 1 and values less than 0.5 are classified as class 0. The model parameters are optimized by Maximum Likelihood Estimation (MLE) that minimizes the classification error on training.

Logistic Regression is efficient over linearly separable data. It has a fast computation and low complexity. And it is less prone to overfitting on small datasets and makes it easy to interpret feature importance with model coefficients. But it could fail to perform well on highly nonlinear and complex datasets which would be better handled by advanced machine learning models.

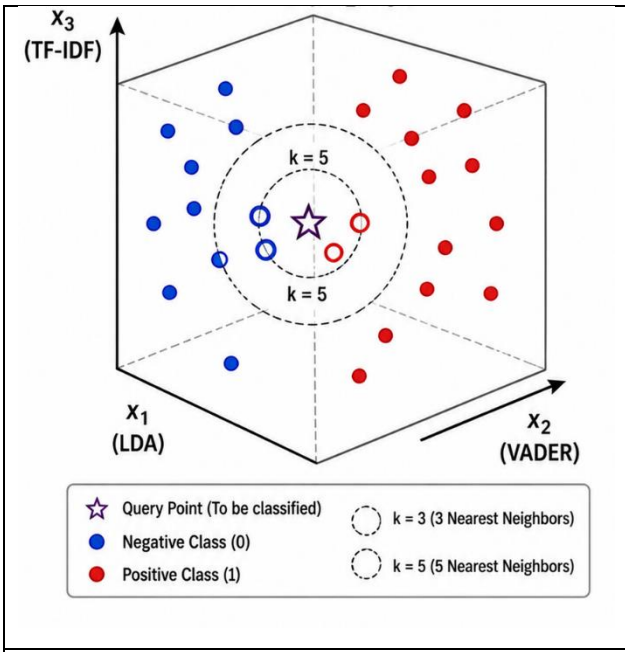


Figure 4c. Working model of K-NNt as ML classifier.

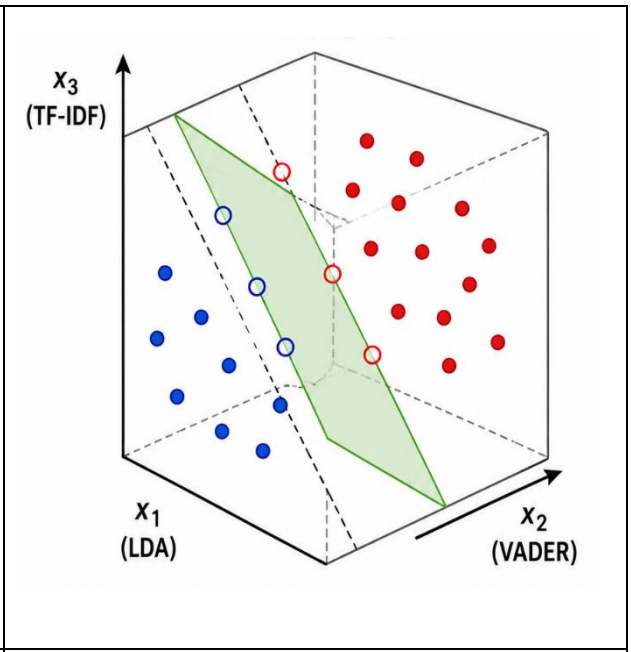


Figure 4d. Working model of SVM as ML classifier.

(c) K-Nearest Neighbors (KNN)

K-Nearest Neighbours (KNN) is a supervised learning algorithm that is used for classification and regression. It is a non-parametric, instance-based learning algorithm where prediction is based on the similarity between data points. To classify a new sample, the algorithm finds the closest neighbouring samples in the training data set. The simplicity and effectiveness of KNN makes it widely used algorithm in pattern recognition, medical diagnosis, recommendation systems, image classification and intrusion detection.

KNN training consists of saving all the training data points in the feature space. When a new input sample is given, the algorithm calculates the distance of the new sample from all the training samples existing. The Euclidean distance is the most common distance metric used and is mathematically expressed as:

Natural Resources for Human Health
eISSN: 2583-1194

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where $x_i$and y_i are feature values of two data points and n is the number of features. After the calculation of distances, the algorithm selects the Knearest neighbors with the smallest distance values. For classification problems the final class is determined by majority voting of the closest neighbors while for regression problems the mean value of the neighboring outputs is used.

The KNN performance highly depends on the choice of K. A very small value of K may cause overfitting and a very large value may reduce the accuracy of classification. KNN does not need a separate training phase, which makes it easy to implement, but it has high computational complexity for large datasets because distance calculations are performed for each prediction. But KNN is still a good algorithm for small and medium size data with well-defined feature space.

Table 4: Top Most Frequently Used Hundred Words

Case: Normal					**Case: Depressed**			
*Bold Fonts: Common Words, **Yellow Highlights: Positive Words					*Bold Fonts: Common Words, **Yellow Highlights: Negative Words			
again	god	Man	Start		**again**	**guy**	**morning**	**Still**
always	**good**	**Miss**	**Still**		already	**happy**	**need**	Stupid
amazing	**great**	**Morning**	**Take**		another	hate	**never**	Suck
awesome	**guy**	Movie	**Talk**		away	**home**	**new**	**Take**
baby	haha	Music	**Tell**		**back**	**Hope**	next	**Talk**
back	**happy**	**Need**	Thank		bad	Hour	**night**	**Tell**
bed	**home**	**Never**	Thanks		**bed**	Hurt	**off**	**Thing**
call	**hope**	**New**	**Thing**		break	Iphone	**old**	**Think**
come	hot	Nice	**Think**		**come**	**Just**	**people**	Though
cool	**just**	**Night**	**Time**		cry	**Keep**	phone	**Time**
day	**keep**	**Off**	**Today**		damn	**Know**	**play**	**Today**
down	**know**	**Old**	**Tomorrow**		**day**	**Last**	**please**	**Tomorrow**
enjoy	**last**	**People**	**Tonight**		die	**Leave**	rain	**Tonight**
ever	late	**Play**	Tweet		**down**	**Let**	**really**	Try
everyone	**leave**	**Please**	**Twitter**		even	**Life**	**right**	**Twitter**
feel	**let**	Ready	**Wait**		**ever**	**Like**	sad	**Wait**
first	**life**	**Really**	**Watch**		**feel**	**Lol**	**say**	**Watch**
follow	**like**	**Right**	**Way**		find	Long	**school**	**Way**
friend	listen	**Say**	**Week**		**friend**	**Look**	**show**	**Week**
fun	live	**School**	Weekend		beat	Lose	sick	**Well**
funny	**lol**	**Show**	**Well**		**get**	**Love**	**sleep**	**Why**
game	**look**	**Sleep**	**Why**		**girl**	**Make**	**someone**	Win
get	lot	Smile	**Work**		**give**	Mean	something	Wish
girl	**love**	**Someone**	World		**good**	**Miss**	soon	**Work**
give	**make**	Song	Yay		**great**	Mom	sorry	Year

(d) Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised ML algorithm used for classification and regression problems. It performs task of binary classification problems, where the aim is to find an optimal decision boundary that separates data points of

different classes. The SVM algorithm has been widely used in the fields of image classification, text classification, medical diagnosis, handwriting recognition and bioinformatics due to its strong generalisation ability and effectiveness of high dimensional data.

The main idea of SVM is to construct a hyperplane with the maximum margin between different classes. Support vectors are the data points that are close to the hyperplane and are important to define the decision boundary. The linear decision function of SVM for a data set with input vector x and class label y is given as:

$$f(x) = w^T x + b$$

where $w$represents the weight vector and $b$denotes the bias term. The optimal hyperplane is obtained by maximizing the margin distance between support vectors and the separating boundary. The margin is defined as:

$$Margin = \frac{2}{\| w \|}$$

SVM solves an optimisation problem by reducing the classification error and increasing the margin. For the non-linear datasets, SVM uses kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid kernels that helps in data mapping to a higher-dimensional feature space where linear separation is feasible. The transformation of the kernel is given by:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

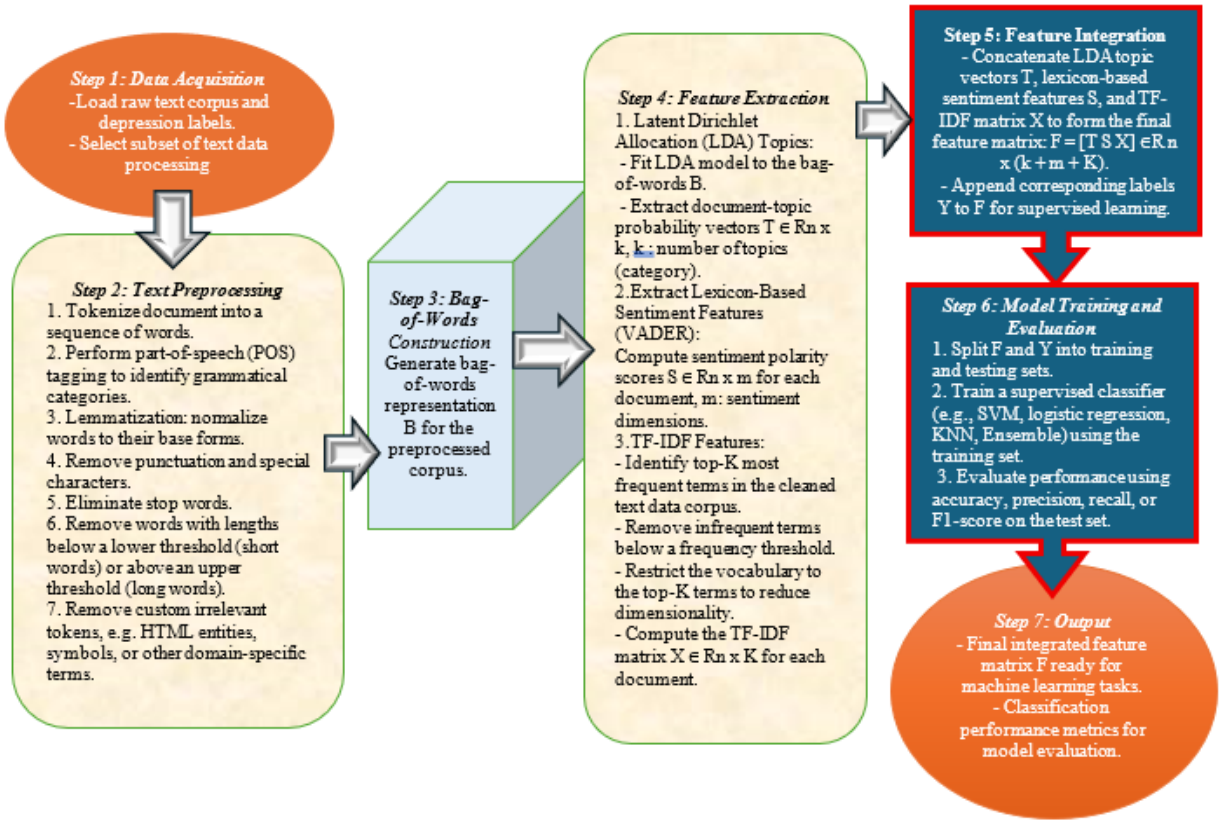
where $\phi(x)$is the mapping function.

SVM has high accuracy and good performance for complex datasets with limited number sample count. It ihas rare issues of overfitting and works well for high dimensional spaces. However, if the datasets is large training complexity is increased, and choosing appropriate kernel parameters become computationally expensive. Despite these limitations, SVM is still one of the most reliable ML algorithms for classification applications.

Abbreviations	**Symbols**	
TF-IDF: Term Frequency-Inverse Document Frequency LDA: Latent Dirichlet Allocation VADER: Valence Aware Dictionary and Sentiment Reasoner POS: Part-of-Speech BOW: Bag-of-Words ML: Machine Learning SVM: Support Vector Machine	D: Input text corpus d_i: i^{th} document in the corpus Y: Label vector for all documents y_i: Label for i^{th} document F: Final feature matrix B: Bag-of-words matrix T: Document-topic probability matrix of LDA	S: Lexicon-based sentiment score matrix X: TF-IDF feature matrix n: Number of documents k: Number of topics in LDA m: Number of sentiment dimensions K: Number of top TF-IDF terms

Natural Resources for Human Health
eISSN: 2583-1194

Algorithm:



4. RESULTS

The methodology explained in previous section is expressed in short as an algorithm. The algorithm is the representative view of the programming written in the Matlab software using the inherited commands of Signal processing, NLP and machine learning tool box. Four different ML models are developed using training dataset by applying several attempts on each ML model type by setting the different values of model parameters [26-29].

The two factors are investigated further to minimize the data requirement and model capacity. The count of number of infrequent words and the number of top words are specified for final unique words extracted from bag of cleaned dataset. The subset values of no. of infrequent words is taken at 5 and 10 (2 combination) while number of top words considered only 30, 50 or 70 (3 combination). For these combination total 3x2=6 experiment sets are designed to figure out the impact on ML model performance on these two factors taken as case id 1 to 6. The performance in terms of accuracy is shown in the table 5. Here each column is the case id and three rows are showing Number of infrequent words (5 or 10), Number of top words (0, 50 or 70) and the respective value of accuracy (%). It may be observed that the percentage accuracy is varying from 86 to 95 percent. The accuracy is highest 95.1% at high value of Number of infrequent words i.e. 10 as well as highest value of top words i.e. 70 under the case id 4.

The results in terms of other performance metrics are further validated using the confusion matrix value and the results are summarized for Number of infrequent words and top word combination of case id 4 and results are shown in table 6 in terms of precision, recall, F1-Score and accuracy. It may be observed that the SVM giving the best result at parameter setting of Gaussian kernel for kernel scale of 7.3. The scond best performance is given by k-NN at 10 neighbors for Euclidian distance with squared inverse weightage to distance.

Table 5: Accuracy comparison at different cases in terms of number of infrequent words and top words

Case Id:	1	2	3	4	5	6
Number of infrequent words	5	5	5	10	10	10
Number of top words	30	50	70	30	50	70
Accuracy (%)	86	92	93.7	95.1	94.8	94.2

Table 6: Summary of best results obtained after testing different experimental cases for Number of Infrequent words =10, Number of Top-words=50, features=54.

	Precision	**Recall**	**F1-Score**	**Accuracy**	**Parameters**
SVM	98.8 %	89.4 %	93.9 %	95.1 %	Kernel: Gaussian, Kernel Scale=7.3
Logistic regression	89.6 %	88.6 %	89.0 %	90.9 %	Linear equation
KNN	94.7%	91.0%	92.7%	94.1%	Number of neighbors=10, Distance= Euclidian, Distance weight= squared inverse
Random Forest	82.9%	68%	74.7%	77%	Number of splits=9999, Number of learners=30, Learning rate=0.1

Figure 5. Bar chart representation of collective demonstration of precision, recall, F1-score and accuracy.

Table 7. Comparison to latest research works based on NLP for sentiment analysis by text data.

Technique	**Precision**	**Recall**	**F1-measure**	**Accuracy**
SABSA [22]	0.858	0.839	0.8485	0.837
SentiVec [21]	0.877	0.858	0.8675	0.861
Ngram+TF-IDF+SVM [23]	0.866	0.846	0.856	0.844
SEML [25]	0.854	0.837	0.8455	0.838
MTMVN [24]	0.817	0.789	0.803	0.792
SVM (Proposed)	0.988	0.894	0.939	0.951

5. CONCLUSION

In this work, a technically rigorous framework for depression detection is presented. It integrates topic modelling, lexicon-based sentiment analysis and statistical weighting into a unified feature matrix. Experimental evaluation confirms that the hybrid approach shows a significant improvement in accuracy, precision and robustness over individual feature extraction techniques. SVM is found to be most effective classifier. The model has a great potential for real-world deployment in large-

scale mental health monitoring systems by systematically preprocessing to reduce noise and incorporating multi-dimensional textual features. Besides its contribution towards the advancement of computational techniques for healthcare, the framework is extensible to domains such as opinion mining, customer behavior analysis and public sentiment tracking where interpretability and high predictive performance are essential. In the future studies, fusing multimodal information such as images, audio and user interaction information with text can be explored for more precise depression detection. Cross-lingual and code-mixed extensions for diverse social media environments may be developed. Furthermore, privacy-preserving mechanisms can improve the scalability and ethical adoption of healthcare systems by enabling real-time deployment.

ACKNOWLEDGEMENT

All the authors of this article are very thankful to the Computer Science and Engineering Department and Research and development centre of Integral University for the guidance and support for this research work. We acknowledge the communication of this article for purpose of publication under the Manuscript communication number (MCN): IU/R&D2026-MCN0004933.

REFERENCES

[1] Naglik, I., & Lango, M. (2024). ASTE-Transformer: Modelling dependencies in aspect-sentiment triplet extraction. *Proceedings of the Association for Computational Linguistics (ACL).*

[2] Bodke, A., Zhang, Y., & Lee, H. (2025). PASTEL: Polarity-aware sentiment triplet extraction with LLM-as-a-judge. *IEEE Transactions on Affective Computing*, Advance online publication.

[3] Peng, J., & Su, B. (2024). Aspect sentiment triplet extraction based on deep relationship enhancement networks. *Knowledge-Based Systems*, 295, 111380.

[4] Zhang, J., Xu, S., Gao, X., & Tang, Z. (2025). Aspect Sentiment Triplet Extraction with Syntax-Semantics Graph Convolutional Network. *International Journal of Computational Intelligence Systems*, *18*(1), 167.

[5] Haocheng Xi, Yutong Wang, Xinyu Wang, Zhitao Li, Bochun Liu, Yihan Wang, Chen Ni (2024). AWA-GCN: Enhancing Chinese sentiment analysis with graph convolutional networks. *Information Processing & Management.*

[6] Çekinel, R. F. (2025). *Multilingual, Multimodal and Explainable Approaches for Automated Fact-Checking Problem* (Doctoral dissertation, Middle East Technical University (Turkey)).

[7] Piao, Y., & Zhang, J. (2024). Text triplet extraction with fused graph neural networks and improved biaffine attention. *Applied Intelligence*, 54(6), 6570–6584.

[8] Zakieldin, K., Khattab, R., Ibrahim, E., Arafat, E., Ahmed, N., & Hemayed, E. (2024). Vitcn: Hybrid vision transformer with temporal convolution for multi-emotion recognition. *International Journal of Computational Intelligence Systems*, *17*(1), 64.

[9] Rasool, Z., Khan, S., & Ahmad, F. (2025). nBERT: Harnessing NLP for emotion recognition in psychotherapy. *Frontiers in Psychology*, 16, 1234567.

[10] Rezapour, M. (2024). Emotion detection with transformers: A comparative study of explanation methods. *Computers in Human Behavior*, 154, 107228.

[11] Creangă, C., & Dinu, L. P. (2024). ISDS-NLP at SemEval-2024 Task 10: Transformer networks for emotion recognition in conversations. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval)*, 145–152.

[12] Patankar, S., & Phadke, M. (2025). A CNN-transformer framework for emotion recognition in code-mixed English–Hindi data. *Discover Artificial Intelligence*, *5*(1), 160.

[13] Zhao, X., Li, J., & Wang, H. (2025). Multi-target opinion words extraction using graph convolutional networks. *Neurocomputing*, 570, 127899.

[14] Zhu, H., Chen, Y., & Liu, Q. (2023). Sentence-guided grid tagging for aspect-sentiment quadruple extraction. *Information Sciences*, 626, 237–251.

[15] Zhang, Y., Wen, S., Zhu, Y., Li, Z., & Wang, X. (2025). Combining large language models with interpretable models for explainable aspect-based sentiment analysis in the medical domain. *Journal of King Saud University Computer and Information Sciences*, *37*(7), 1-19.

[16] Yang, J., Dai, F., Li, F., & Xue, Y. (2023, October). Span-based pair-wise aspect and opinion term joint extraction with contrastive learning. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 17-29). Cham: Springer Nature Switzerland.

[17] Deng, J., & Liu, Y. (2025). Research on sentiment analysis of online public opinion based on RoBERTa–BiLSTM–attention model. *Applied Sciences*, *15*(4), 2148.

[18] Lu, X., Chen, F., & Zhao, Y. (2024). Automatic construction of a Chinese review dataset for aspect sentiment triplet extraction via iterative weak supervision. *Data & Knowledge Engineering*, 149, 102215.

[19] Kalsum, T., Rauf, A., & Ali, N. (2023). A novel lightweight deep CNN model for human emotions recognition in diverse environments. *Multimedia Tools and Applications*, 82(13), 19751–19769.

[20] Venkatesan, S., Kumar, R., & Prasad, M. (2023). Human emotion detection using DeepFace and artificial intelligence. *Journal of Ambient Intelligence and Humanized Computing*, 14(5), 3123–3136.

[21] Zhu L, Li W, Shi Y, Guo K. SentiVec: learning sentiment-context vector via kernel optimization function for sentiment analysis. IEEE Trans Neural Netw Learn Syst. 2021;32(6):2561–72. https://doi.org/10.1109/tnnls.2020.3006531.

[22] Schouten K, van der Weijde O, Frasincar F, Dekker R. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. IEEE Trans Cybern. 2018;48(4):1263–75. https://doi.org/10.1109/tcyb.2017. 2688801.

[23] Ayyub K, Iqbal S, Munir EU, Nisar MW, Abbasi M. Exploring diverse features for sentiment quantification using machine learning algorithms. IEEE Access. 2020;8:142819–31. https://doi.org/10.1109/access.2020.3011202

[24] Bie Y, Yang Y. A multitask multiview neural network for end-to-end aspect-based sentiment analysis. Big Data Mining Analytics. 2021;4(3):195–207. https://doi.org/10.26599/bdma.2021.9020003

[25] Li N, Chow CY, Zhang JD. SEML: a semi-supervised multi-task learning framework for aspect-based sentiment analysis. IEEE Access. 2020;8:189287–97

[26] Sahay, S., Banoudha, A., & Sharma, R. (2014). On the use of ANFIS for predicting groundwater Levels in Hard Rock Area. Int. J. Res. Dev. Appl. Sci. Eng.

[27] On the use of ANFIS for predicting groundwater Levels in Hard Rock Area S Sahay, A Banoudha, R Sharma - Int. J. Res. Dev. Appl. Sci. Eng, 2014

[28] P. Pathak and M. M. Tripathi, "Hybrid machine learning models for post-COVID sentiment public health detection," Nanotechnology Perceptions, vol. 20, S11, pp. 104–119, 2024, doi:10.62441/nano-ntp.v20iS11.8.

[29] Singh, N., Singh, S. K., & Singh, R. (2025). Machine learning-based sentiment analysis for suicide prevention and mental health monitoring in educational institutions. *Journal of Neonatal Surgery*.