Information Processing and Management 60 (2023) 103168

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Fair and Explainable Depression Detection in Social Media

V Adarsh [a], P Arun Kumar [b], V Lavanya [b], G.R. Gangadharan [a,*]

[a] Department of Computer Applications, National Institute of Technology Tiruchirappalli, India
[b] Department of Management Studies, National Institute of Technology Tiruchirappalli, India

ARTICLE INFO

*Keywords:*
Depression
Fairness
Feature Selection
Explainable Artificial Intelligence
Mental Illness
Natural Language Processing
One-shot Decision Approach
Participation Dynamics

ABSTRACT

Detection at an early stage is vital for the diagnosis of the majority of critical illnesses and is the same for identifying people suffering from depression. Nowadays, a number of researches have been done successfully to identify depressed persons based on their social media postings. However, an unexpected bias has been observed in these studies, which can be due to various factors like unequal data distribution. In this paper, the imbalance found in terms of participation in the various age groups and demographics is normalized using the one-shot decision approach. Further, we present an ensemble model combining SVM and KNN with the intrinsic explainability in conjunction with noisy label correction approaches, offering an innovative solution to the problem of distinguishing between depression symptoms and suicidal ideas. We achieved a final classification accuracy of 98.05%, with the proposed ensemble model ensuring that the data classification is not biased in any manner.

## 1. Introduction

Depression is still one of the most significant challenges that the world is confronted with in the modern day, and if it is not handled, it may lead to thoughts of suicide as well as real attempts at suicide. A considerable obstacle exists on both an individual and a social level when it comes to the accurate diagnosis of depression and the determination of the point at which it becomes a risk factor for suicidal thoughts or behaviour. According to Goldman & Lewis (Goldman & Lewis, 2008), depression is a hidden mental disease that affects a person and it is also something that may happen to any individual who seems to be doing absolutely well. Moreover, it is one of the most prevalent forms of mental illness that can be found in our modern world, which is distinguished by a lightning-fast reliance on the progression of technological innovation. Even on the most basic level, the existing approaches fail to address the general population's collective needs for mental health care. It is important to note that no known treatment can completely reverse the effects of this illness. Hence it is of the utmost importance that we find out what is causing it and find a solution to the problem at its core so that it does not become significantly worsen over time.

Patients' suicidal notes and post-processing of the Electronic Health Records (EHR) of patients are generally the methods used in most instances for discovering such illness. However, the amount of information obtained from the said methods is often relatively restricted. In light of the fact that individuals use social media on a widespread scale and have easy access to these platforms in these

*Abbreviations:* CNN, Convolutional Neural Network; DL, Deep Learning; DT, Decision Tree; KNN, K-Nearest Neighbor; LIME, Local Interpretable Model Agnostic Explanations; ML, Machine Learning; NMT, Neural Machine Translator; RF, Random Forest; RNN, Recurrent Neural Network; SVM, Support Vector Machine; TF-IDF, Term Frequency Inverse Document Frequency.

* Corresponding author.
*E-mail address:* geeyaar@gmail.com (G.R. Gangadharan).

https://doi.org/10.1016/j.ipm.2022.103168
Received 20 August 2022; Received in revised form 19 October 2022; Accepted 9 November 2022
Available online 17 November 2022

days, researchers are analyzing individuals' behaviour via social media to battle the scarcity of data. An increasing number of studies are being carried out to comprehend suicidal ideas expressed on social media by applying natural language processing (NLP) and machine learning (ML) techniques (Boettcher, 2021; De Choudhury & De, 2014; Guo, 2011; Kim et al., 2020; Lokala et al., 2022; Low et al., 2020; Slemon et al., 2021; Xu et al., 2020). Reddit is one of the most prominent social media networks as it allows user-generated content to be placed on its site and users can discuss the difficulties which they encounter in their daily lives on the platform. As the Reddit platform enables users to express their problems anonymously and can receive assistance from within the platform, the chances of obtaining a large amount of reliable data on this platform are also significantly high. Using this data, we can figure out how likely someone currently going through the stages of depression will have suicidal thoughts or try to kill themselves. If persons of this sort are provided with early warning information, a significant number of lives may likely be spared. However, many studies are not nearly as accurate as they might be because these studies' data originate mainly from Reddit posts, which may be vulnerable to bias.

People who regularly use the Reddit website, especially those between the ages of 18 and 49, are the most important contributors to the data acquired from Reddit. Reddit users older than this age range have a greater propensity to use social media less regularly. As a result, their contributions to posts on Reddit will be lower when compared to those of users in other age groups (Kim et al., 2020). When this happens throughout the process of training a model, the characteristics shown by those classes having a low population tend to become less evident. Thus, there is a reasonable likelihood that the model would not be able to identify the characteristics shared by the group being considered as part of the analysis. To ensure that the model is trustworthy as well as fair, it is essential to consider the fair distribution of the data over various classes.

When studying a model, it is common practice to focus on the aspects of the model that are considered to be its most important, whether that be its training examples, input data, or deep network layers. To understand the data, one may search for patterns or use the model's features to investigate how the model internally reflects the data. Both of these approaches are valid. This involves looking at the character-to-sentence range of linguistic units and model components that NLP offers. Numerous studies have been carried out on the topic of suicidal thoughts on social media; however, to the best of our knowledge, none of these studies has acknowledged how important it is to disseminate the information that has been gathered in an equitable manner (De Choudhury et al., 2016; Fu et al., 2013; Rabani et al., 2021; Robinson et al., 2016; Roy et al., 2020; Young & Garett, 2018). In the context of our particular circumstance, the model needs to be able to recognise any and all indications and symptoms that are produced by individuals, irrespective of their age group, demography, or any other factor that may be involved. In order to tackle this issue, we utilise the one-shot decision problem as a way to fix the class inequality. The concept of participation dynamics (Zhang, Khaliligarekani, Tekin, & Liu, 2019) is being employed so that every individual who fits into one of the defined groups may be confident that they will have an equal opportunity to participate in the venture under consideration. This action is taken to ensure that all of the classes or groups that are taking part in the process continue to maintain their fair standards of contributions throughout the entirety of the procedure and do not experience any kind of decline in their contributions in any way in the middle of the process.

The purpose of this research is to identify and evaluate individuals at risk for developing a mental disorder and to look for early warning indicators of suicidal thoughts and behaviours. In line with this, it seeks to provide a reliable classifier that can compensate for any inherent bias that may have been introduced during the process of data gathering to generate an objective model. Initially, the data collected for this study was scraped and pre-processed, and then we fixed the class imbalance on the acquired data using the one-shot decision (Guo, 2011) mechanism. The data is processed with semantic network analysis to create a frequent and most important set of words to act as a pre-classifier that can enhance the clustering quality. Then clustering algorithm is employed to cluster the groups into two broader categories: those with suicidal tendencies and those without suicidal tendencies. The clustered groups are then analysed for the participation of the groups and concerning their age and demographics. Further, the participation dynamics are used to verify the contribution of the groups and label correction is employed to enhance the stability. Once the participation is confirmed, it is passed on to the ensemble model for the final classification, and the factors for classification are presented using LIME (Ribeiro et al., 2016).

The remaining parts of the paper are structured as follows. Section 2 discusses the existing research literature in depression to suicidal ideation diagnosis. Section 3 presents the proposed methodology on fair and explainable depression detection and the results are described in Section 4. Section 5 highlights the conclusions and future directions.

## 2. Related Works

People are increasingly turning to social media platforms such as Facebook and Twitter to discuss and seek help for difficulties related to their mental health (Low et al., 2020). Because of this, academics have been inspired to use the data in conjunction with various Natural Language Processing (NLP) and Machine Learning (ML) techniques to aid people who may want assistance. Recently, significant progress has been made in identifying those in danger of committing suicide via social media (Luxton et al., 2012). These developments have been made feasible due to the platform's real-time nature, cheap cost, and ability to reach a vast audience (Braithwaite et al., 2016). It has been shown that people with suicidal thoughts spend more time online, have a more significant predisposition to engage in personal connections online, and participate in online forums more often (Harris et al., 2013). Until recently, a considerable portion of prior studies depended on Twitter data. Now it is switched over to Reddit data, and researchers are focusing on the information from Reddit (Benton et al., 2017; Coppersmith, Dredze, Harman, Hollingshead, et al., 2015; Gkotsis et al., 2017; Husseini Orabi et al., 2018; Kim et al., 2020; Zirikly et al., 2019). Reddit platform allows users to express their difficulties anonymously, and users can receive assistance inside the platform. The likelihood of obtaining a large quantity of reliable data on this platform is also very high. Using this information, we may determine the possibility that someone experiencing the stages of depression will have suicidal thoughts or attempt suicide.

Deep learning, more specifically, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are used for diagnosing depression (Husseini Orabi et al., 2018). A reinforcement learning system based on Recurrent Neural Networks was proposed by (Gui et al., 2019) to identify people suffering from depression. Even though these advanced models based on deep learning can extract higher-level semantic information and have made great progress, they are not yet capable of successfully extracting emotional semantic information. A CNN-based deep learning model with NLP method (Kim et al., 2020) is presented to identify users with potential mental illness based on their Reddit posts.

Since the inception of deep learning in NLP, many studies have been using deep learning models for a depression diagnosis. There have been many different approaches to identifying suicidal ideations from depressive thoughts (Haque et al., 2021), ranging from more traditional NLP techniques to more cutting-edge deep learning methods based on neural network models. Several of the currently available models diagnose depression based on feature engineering. These models include the bag of words (Zhang et al., 2010), Latent Dirichlet allocation (LDA) (Blei et al., 2001), N-gram (Baker, 1990), and the Linguistic Inquiry and Word Count (LIWC) dictionary (Tausczik & Pennebaker, 2010). LIWC was used in the study that looked into the capability of identifying feelings of melancholy (Coppersmith, Dredze, Harman, & Hollingshead, 2015). Support Vector Machines (SVM) were utilised to create a multimodal technique for detecting depression. Kang et al. (Kang et al., 2016) included text analysis, emoticon word analysis, and a support vector machine-based image classifier. To improve the accuracy with which they diagnose depression, the researchers concluded that they needed to construct a mood lexicon and then conducted an analysis of emoticons (Ren et al., 2021).

Moreover, (Coppersmith, Dredze, Harman, Hollingshead, et al., 2015) used character-level language models to assess the likelihood that a user with mental health issues would produce a certain string of characters. This was done to determine how probable the user would construct the string. They did this so that they could determine how likely it was that the user would produce anything. (Benton et al., 2017) compared to a basic regression model, a multilayer perceptron single-task learning model and a neural multitask learning model to diagnose various mental health conditions. Furthermore, to detect a higher level of precision, (Husseini Orabi et al., 2018) attempted to detect depression by combining word embeddings with a number of different neural network models, including CNNs and RNNs. Also, experiments were carried out utilizing Feed Forward Neural Networks, CNNs, Support Vector Machines, and Linear classifiers in order to carry out binary classification on posts relevant to mental health (Gkotsis et al., 2017). Hierarchical attention networks were created by (Sekulic & Strube, 2019) to identify a range of mental health issues such as anxiety, depression, and attention deficit hyperactivity disorder. (Gruda & Hasan, 2019) employed a non-intrusive machine learning approach to scale human anxiety ratings in microblogs to predict perceived anxiety over time. It illustrates user state-anxiety changes over time as well as mean trait anxiety. (Shen & Rudzicz, 2017) explored anxiety disorders using personal anecdotes gathered on Reddit by constructing features that successfully categorise posts linked to binary degrees of anxiety using N-gram language modelling, vector embeddings, topic analysis, and emotional norms on a large data set of usual and anxiety-related posts. (Jiang et al., 2020) gathered Reddit postings from people with eight diseases and classification models were developed based on deep contextualised word representations.

(De Choudhury et al., 2021) constructed a model for diagnosing depression by making use of the data that was obtained from Twitter. This model was developed with the assistance of the data that was collected from Twitter (such as decreased social activities, increased negative emotions and self-consciousness, a higher level and increased expression of religious thoughts, etc.). (Park et al., 2021) questioned 14 active users using a combination of semi-structured and unstructured questions. According to the results of the research, individuals who struggle with depression see social media as a venue in which they may not only exchange information with one another but also express their own emotions.

(Pestian et al., 2010) used machine learning methodologies to predict suicide risk from suicide notes in order to construct a suicide note classifier that is more successful than human psychologists in identifying between real and fake online suicide notes. This was achieved by developing a classifier for suicide notes that was able to determine the level of suicide risk based on the contents of suicide notes. MySpace.com is a well-known website among adolescents and young adults, notably among adolescents belonging to sexual minority groups, and it has more than one billion registered members all over the globe. On the website, lexicon-based keyword matching was used to perform the search for suicide notes that was described in (Huang et al., 2007). This search was conducted with the objective of determining whether or not people have suicidal thoughts or want to take their own lives. (Li, Ng, Chau, Wong, & Yip, 2013) analysed the user posts and comments that were placed in a Chinese internet forum in order to uncover phrases that are associated with suicide. These approaches included both the analysis of textual sentiment as well as summarization of the information. (Masuda et al., 2013) conducted research at Japanese online forums and discovered that a user's chance of thinking about suicide was associated with the number of communities they belonged to, their intransitivity, and the percentage of suicidal neighbours in their social network. In addition, the number of communities a user belonged to was found to have a correlation with their likelihood of thinking about suicide. A statistical technique known as logistic regression was used in the study by (De Choudhury et al., 2016) to investigate the likelihood of users on Reddit moving from sub-communities on Reddit that concentrate on mental health to those that encourage suicide. (Alambo et al., 2019) developed a method to generate four semantic clusters that are similar to one another and that characterise suicide indications, suicide thinking, suicide behaviour, and suicide attempts. The suicide lexicons served as the foundation for these clusters. They merged survey questions with Reddit posts that were semantically related to one another to create clusters and then utilized those clusters to determine the overall severity of the suicide risk provided by a post.

(Hiraga, 2017) recognized sad attitudes by analyzing numerous features of the language, including character n-grams, token n-grams, lemmas, and selected lemmas. This allows to determine whether or not a person was feeling melancholy. (Hussain et al., 2019) developed what is now known as the Socially Mediated Patient Portal via collaborative efforts with their contemporaries.

In this article, our primary focus is on investigating a fair framework for detecting a connection between depressive symptoms and the risk of having suicidal thoughts or the chances that the depressive state of mind leads to suicidal ideations. We are utilizing various fairness metrics to ensure that the system we design is least affected by the bias of any terms, i.e., it tries to avoid any chances of

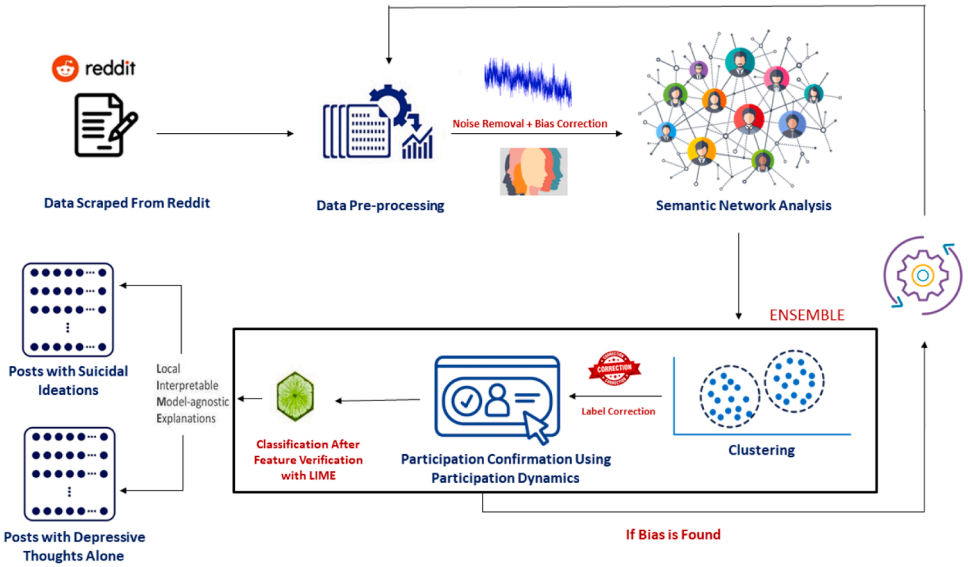



**Fig. 1.** Proposed methodology.

**Table 1**
Properties and attributes of data used in the study.

| Subreddits | Count | Description |
|---|---|---|
| Bipolar disorder | 342 | Communities devoted to discussing topics related to bipolar disorder, in contrast to bipolar and Bipolar Reddit, which are primarily concerned with patients and the support they receive, welcome contributions from individuals in a romantic or sexual partnership with someone who has bipolar disorder. |
| Suicidal Watch | 6442 | Support groups for persons struggling with suicidal thoughts or concerned that someone else may be in danger of taking their own life. |
| Borderline Personality Disorder | 108 | Text Discussion board for those with Borderline Personality Problems, those who are closely linked to someone who has this disorder, or those who are just interested in learning more about this condition. |
| Schizophrenia and Psychosis | 478 | Subreddit for the discussion of schizophrenia and other mental health conditions, such as psychosis, that are connected to schizophrenia |
| Anxiety | 545 | Discussion board for everything associated with anxiety disorders; does not differentiate between sufferers and those linked to those who suffer from anxiety disorders. |
| Depression | 4996 | A community dedicated to supporting those dealing with depression; posting is not limited to people diagnosed by their primary care physician or a doctor working in an inpatient environment, and the major emphasis is on helping others in their struggle with depression. |

inherent biases that may be present in the data or any bias that comes unintentionally. Further, we present a method that uses a clustering-classification approach in the form of an ensemble to produce a model which best fits the task at hand.

## 3. Proposed Methodology

Fig. 1 provides an outline of the procedures that were carried out while conducting this experiment. The data required for the study was obtained from Reddit, utilising the Reddit Application Programming Interface (API). After the data had been gathered and text pre-processing was carried out, we used a basic Neural Machine Translator (NMT) to remove the noise in the collected data. Once the noisy data is cleaned, we examine the data for class imbalance. The imbalance found in terms of participation in the various age groups and demographics is normalised using the one-shot decision approach. This can aid in reducing biases against specific age groups and demographics with lower participation rates.

After the data pre-processing and the bias normalization, the semantic network analysis is performed using the term frequency-inverse document frequency (TF-IDF). By doing this, we can determine the frequency of the terms and sentences that semantically align to the clusters, which acts as a further pre-filtering stage for the later clustering algorithm.

The filtered data is then sent to a clustering algorithm, which is a Gaussian mixture model with K-Means clustering. This algorithm broadly classifies the data into two groups: those with suicidal ideations and those who do not have them. The participation dynamics concept is applied to analyze the two clustered groups to validate the participation concentrations of the groups present in the primary data. If the group retention cannot be verified, it is sent back to the initial stages for further tuning. Suppose the verification turns out to be positive, it is classified using the ensemble to produce a robust classification, and the features contributed to the classification are listed out using LIME.

**Algorithm 1**
Algorithm for cleaning data using NMT.

Input: Noisy data
Output: A Cleaned data
t = 0; T= max(Input)
while $t < T$ do
1 Input ← denoise(NMT)
1 Randomly take data samples from the cleaned set.
1 Compute noise by creating instances from cleaned set.
1 Sort the set by noise scores.
1 Discard the higher value-scored instances.
1 t=t+1
end while

**Algorithm 2**
Algorithm for one shot decision approach.

Input: Data with inherent bias
Output: Normalised Data
1 nput (data)
1 Apply the decision rule $h_\theta(x) = 1(x \geq \theta)$
1 Find a pair $(\theta_a(t), \theta_b(t))$ such that loss is minimised.
1 Apply $\Gamma_{Eq0pt.t} = \int_{\theta_a}^{\infty} f_{a,t}^0(x)dx - \int_{\theta_b}^{\infty} f_{b,t}^0(x)dx$ to define equal opportunity using a false positive rate.
1 Calculate loss and save the ones with minimal loss.

### 3.1. Data Pre-Processing

We employ Natural Language Processing tools to pre-process the dataset (See Table 1) before moving on to the step of picking the features to be trained on and doing the actual training. To begin, we separate each post into its unique token using a method called tokenization. Subsequently, we delete any URLs, punctuation, and stop words that, if left intact, might result in inconsistent results. Following that, we utilize stemming to reduce the words to their most basic form and organize them according to the degree to which they are connected. After that, the models are given features that capture people's linguistic patterns while posting in Reddit forums.

### 3.2. Noise Removal using Neural Machine Translator (NMT)

Neural Machine Translator (NMT) (Bahdanau et al., 2014) is a simple encoder-decoder network. Encoder and decoder are both multilayer recurrent neural networks. When the encoder receives an input sentence, it converts it into a sequence of hidden state representations. Another multilayer recurrent neural network that anticipates a target sequence is the decoder. The current hidden state of the decoder, the context vector, which is a conditional representation of the source sequence relevant to the target position, and the last target token are the three elements that determine the likelihood of producing a token. The primary target token influences this likelihood.

The primary focus of our denoising of NMT training is to train an NMT model using a curriculum that becomes noisier over time (data batches). As a result, training is a dynamic process requiring ongoing adjustment. To reduce the noise level in batches of data throughout several training steps, we use dynamic data selection as part of our implementation of the denoising program. Our approach as described in Algorithm 1 avoids modifying the per-example loss directly to influence the noise-to-training loss ratio via selective data use. In most cases, if the sample is big enough, there will be sufficient instances from which to draw. To be more explicit, the approach leverages trustworthy data to refine the original model at each stage. We can then use this combination of noisy and clean models to assign noise scores to training examples stored in a continuously updated buffer with random data from the sample at each step. Those noise scores have been ranked. The final set of raw data is then arbitrarily selected from the highest-ranked sets of scores (See Fig. 2). The sentence's probability is calculated by multiplying the probabilities of each target token. As a result, the sentence's probability increases. A SoftMax layer is applied to the output of the feedforward network, which generates the probability of each word in the target vocabulary.

Using an attention mechanism, the context vector is arrived at by computing the weighted sum of the encoder's hidden states (Luong et al., 2015). The computation of activations involves applying a scoring function on the relationship between the decoder's current state and each of the encoder's hidden states. This allows the activations to be determined. Finally, the outer summation is the total loss calculated by adding the totals from each target sequence. As a result, moving the pseudo labels closer to the ground truth label makes it feasible to train a more accurate model.

### 3.3. Bias Normalisation using One Shot Decision Approach

We are using the One-Shot Decision to address the issue of group retention to reduce and normalize the bias that arises as a

consequence of varied engagement from people of different demographics, ages, and so on. Two separate demographic groups are designed to be distinct or have meaningful relationships.

We can see that, even in somewhat favourable situations, the disparity in group representation rises with time and that disadvantaged groups may eventually be completely excluded from the system (Heidari & Krause, 2018). When utilizing a standard algorithm to make decisions, it is possible that this criterion will be readily satisfied (for example, seeing the goal as minimizing the total loss). This is because a few universally accepted criteria for fairness have been satisfied (e.g., statistical parity, equal opportunity, etc.).

In this situation, we are utilizing the One-Shot Decision problem as specified in Algorithm 2 to overcome the problem with the group retention (Zhang, Khaliligarekani, Tekin, & Liu, 2019) of individuals in the marginalized section (Hashimoto et al., 2018). Additionally, we are attempting to avoid the class imbalance (Galar et al., 2012) to maintain participation dynamics. The following is an example of how the notion might be modelled:

Consider the following: there are two demographic groupings that are either meant to be differentiated from one another or to be closely connected to one another. These groups are divided into $G_a, G_b$ Subcategories according to sensitive characteristics like gender, age range, geographical area, etc.

It is possible to use a feature called $X \in \mathbb{R}^d$ and a label called $Y \in \{0,1\}$ to represent a participant from any one of the groups. Both of these features are time-varying.

$G_k^j \subset G_k$ is the notation used to refer to the subgroup that has the label $j, j \in \{0,1\}, k \in \{a,b\}, f_{k,t}^j(x),$ which indicates the feature distribution of the subgroup, and $\alpha_k^j(t)$ Which indicates the size of the $G_k^j$ Subgroup as a percentage of the whole population at the time $t$.

The size of $G_k^j$ As a proportion of the population is denoted by $\overline{\alpha}_k(t) := \alpha_k^0(t) + \alpha_k^1(t)$ and the representation disparity between two groups at time step t is measured by the difference between $\overline{\alpha}_a(t)$ and $\overline{\alpha}_b(t)$.

If we denote the percentage of labels $j \in \{0,1\}$ in group k at time t using the symbol $g_{k,t}^j = \frac{\alpha_k^j(t)}{\overline{\alpha}_k(t)}$ then the distribution of $X$ across $G_k$ may be written as follows: $f_{k,t}(x) = g_{k,t}^1 f_{k,t}^1(x) + g_{k,t}^0 f_{k,t}^0(x)$ and $f_{a,t} \neq f_{b,t}$

In this context, we are thinking about a problem of binary classification based on the feature $X \in \mathbb{R}$. The decision rule may be described as $h_\theta(x) = 1(x \geq \theta),$ which is a threshold policy that is parameterised by $\theta \in \mathbb{R}$, and $L(y, h_\theta(x)) = 1(y \neq h_\theta(x))$, which is the 0-1 loss that is suffered by applying decisions on people who have data $(x, y)$.

The purpose of the decision-maker at each time is to select a pair $(\theta_a(t), \theta_b(t))$ according to criteria $\mathscr{C}$ such that the total anticipated loss is minimized.

This poor state endures and has the potential to worsen if feature distributions continue to be damaged and fluctuate over time. If the traits equalised by the fairness criteria do not correlate to what drives user retention, one of the most important considerations is that the difference in treatment will compound representation disparity over time.

We employ the one-shot decision approach to produce a threshold where we categorize the groups and calculate the participation ratio in terms of age and demographics. We then refer to the subgroup with the label and find out the feature distribution and the number of features by which the groups are represented. Once it has been found out, we use a random selection of the top features from both groups equally, and the normalized data is further processed. Using this approach, we eliminate the biases that may reside within the collected data.

### 3.4. Syntax Based Semantic Network Analysis

Semantic analysis relies on humans analyzing and comprehending semantic structures for exploration or reasoning. By mixing domain-specific and common sense information, this technique discovers qualitative features of a semantic network (Helbig, 2006). This difference is created because neither an analyst nor a machine can quickly manage vast volumes of data or evaluate semantics if the extensional meaning is unknown.

After the pre-filtering of the data, the refined words were listed according to their frequency and the term frequency-inverse document frequency values. Then the general features that have been apparent in the texts of r/SuicideWatch and r/Depression were all identified. After that, we utilised the identified words to construct word matrices, which are also known as co-occurrence matrices, in addition to semantic networks. These matrices are used to map the relationships between words. We utilized the frame of words to determine the degree and the eigenvector to sort out significant nodes, or words, in the network. This allowed us to decide which words had the most impact on the overall meaning of the network. Not only did users in the two subreddits devoted to mental illnesses exchange information on the latest medical treatments, but also they discussed their personal experiences, opinions, and emotions. Because of this, we could determine which words were the most significant.

In semantic network analysis, nodes and edges make up a network. Nodes can be global or local network hubs. Subgraphs are network segments that are comparable to a node's surroundings. Strongly coupled nodes form clusters. These elements are used to do quantitative and qualitative semantic network analysis. As part of our research, we have also analyzed and evaluated the lists of positive and negative emotional terms provided by the r/SuicideWatch and r/Depression subreddits, respectively. Moreover, to analyze the emotional terms used, we used linguistic inquiry and word count (LIWC) (Tausczik & Pennebaker, 2010).

Furthermore, the convergence of iterated co-relation (CONCOR) method was used to determine the structural equivalence of networks. This was done to establish whether or not the networks are structurally comparable to one another (Breiger et al., 1975). By utilizing the CONCOR methodology, we were able to discover the semantic characteristics of the texts and the clusters that occurred within the networks.

### 3.5. Clustering

After the syntax-based semantic network analysis has been applied to the processed data and the embedding has been created, we use an unsupervised clustering technique to cluster the processed embedding into two groups. These clusters are those that show no symptoms of suicidal ideations and those that reveal suicidal ideations. This approach aided us in attaching labels to the postings. The clustering approach we use in this particular scenario is known as the Gaussian Mixture Model (GMM) (Reynolds, 2009). The Gaussian Mixture Model considers the probability distribution of continuous data modelled by the parametric probability density function. We use K-Means Clustering, which divides the embedding into a certain number of groups called clusters. This is done to allocate the posts to the cluster where it demonstrates the closest means, with the goal of assigning them to a cluster that displays the lowest within-cluster distance and maximizing the between-cluster distance. Following that, spectral clustering is used, which enables the clustering of semantically coherent data by detecting clusters in different sub-spaces within the dataset (ParsonsLance et al., 2004). After the clustering has been applied, they will be left with two labels attached: the ground-truth labels, which were taken from the subreddit, and the labels that are created by unsupervised clustering.

### 3.6. Participation Dynamics for Participation Confirmation

As a final step to clustering, an approach called confidence-based thresholding is utilized to make adjustments to the ground-truth labels after it is checked for the participation dynamics of the various participating groups.

To have a better grasp of the dynamics of participation (Zhang, Khaliligarekani, Tekin, & Liu, 2019), we can examine how an individual or a user reacts to a certain choice. This reaction is reflected in the dynamics of retention, which can be characterised by having an awareness of the retention rate. In the following, we will provide two instances of different kinds that might occur when the monotonicity criteria are met.

(1) Accuracy of the model determined by the use of a departure: In this scenario, users who are experiencing a lower accuracy or misclassification may be regarded to be a deciding factor for the retention of the group. These users have a greater likelihood of quitting the system. Therefore, this quality can be considered a factor in the retention of the group.

For any strictly decreasing function $\nu(\cdot) : [0,1] \rightarrow [0,1]$, the retention rate at time t may be modelled as:

$$\pi_{k,t}(\theta_k) = \nu\big(L_{k,t}(\theta_k)\big)$$

where $\theta_k$ is the retention rate at time k.

(1) User departure driven by intra-group disparity: participation in the group can be hugely affected by the intragroup disparity that occurs between the users who are from the same demographic groups but with a different set of labels,

i.e., $G_k^j$ For $j \in \{0,1\}$. This can cause a user to leave the group. If we model the intragroup disparity of $G_k$ at time $t$ using the formula

$$Dk,t(\theta_k) = \Pr\big(Y=1, h_{\theta_k}(X) = 1 | K=k\big) - \Pr\big(Y=0, h_{\theta_k}(X) = 1 | K=k\big) \int_{\theta_k}^{\infty} \Big(g_k^1 f_{k,t}^1(x) - g_k^0 f_{k,t}^0(x)\Big) dx$$

Then we can model the retention rate using the formula $\pi_{k,t}(\theta_k) = w_{k,t}(\theta_k)$

The value that was originally obtained during the time of data collection is examined in contrast to the retention rate that was derived by applying the participation dynamics concept. We then use the User departure driven by intra-group disparity to find out the ratio of the participation that we had in the beginning stage and the current stage. If there is a change in the distribution of the data, it is pushed back into the prepossessing stage; Else, it will be forwarded for final classification.

For the label correction, if the clustering algorithm correctly predicts the projected label with a probability that is greater than a predetermined threshold, then the ground truth label will be substituted with the prediction. In the event that the algorithm is unable to provide an accurate forecast of the projected label, the prediction will be utilised instead. It ensures that only predicted labels with a high degree of confidence are used to update the ground truth. Thus, the tuned threshold removes the chance of erroneous adjustments being made.

### 3.7. Classification

The classification is done after the participation of each of the classes is verified using the participation dynamics concept to ensure group retention. Once it is verified, the latter is passed into the proposed ensemble classifier, which classifies the data into two buckets. Here we are utilising a combination of Support Vector Machines (SVM) with K-Nearest Neighbour (KNN) in order to create an ensemble with layered function to check our desired class imbalance issue as well.

Support vector machines, or SVMs (Gkotsis et al., 2017), are a rapid technique to tackle classification and regression problems that have low generalisation error or, in other words, do not suffer from overfitting to the training data set. SVM performs very well on linearly separable data sets, which implies that a hyperplane can be built that divides the instances into two classes in such a manner that examples from one class (almost) entirely fall on one side. Because there are an endless number of possible hyperplanes, the

**Table 2**
Depression and suicidal ideation classification (as in the existing literature).

| Method | References | Accuracy Obtained in % |
|---|---|---|
| **XGBoost** | (Kim et al., 2020) | 71.69 |
| **CNN** | (Husseini Orabi et al., 2018) | 77.48 |
| **SVM** | (Gkotsis et al., 2017) | 76.46 |
| **Naïve Bayes** | (Balani & De Choudhury, 2015) | 62.5 |
| **KNN** | (Balani & De Choudhury, 2015) | 60.7 |
| **DT** | (Balani & De Choudhury, 2015) | 58.8 |

support vector machine (SVM) chooses the hyperplane H based on its distance to the data points in either class that are closest to it. This is referred to as "margin maximisation."

The K-Nearest Neighbours (KNN) (Balani & De Choudhury, 2015) is a non-parametric, supervised learning classifier that employs proximity to classify or predict the grouping of a single data point. While it may be used for either regression or classification issues, it is most often utilised as a classification technique based on the idea that comparable points can be discovered nearby. To assess the degree of similarity between two vectors, the k-nearest neighbour algorithm generally uses the Euclidean distance as the distance metric (points).

Using a voting classifier, we combine the SVM with KNN to produce an ensemble. While making a decision, the viewpoint with the most votes is considered. We limit the number of errors produced by the ensemble system by adding weights to the majority voting process. Here the ensemble consists of a clustering followed by classification. This ensures that the entire classification process gives a much more robust result as it is working on data that is already being clustered into groups.

### *3.8. Explanation using LIME*

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) can assist in the development of explainable artificial intelligence. This strategy seeks to illuminate a machine-learning model and make each algorithmic prediction intelligible on its own. Because it describes the classifier for a specific instance and provides the needed level of information, this method is suitable for usage in local explanations. It can be utilised in regional explanations. LIME alters the input data before producing a series of manufactured data with just a fraction of the original attributes. These changes are made to the data that has been provided as input. LIME's output is the result of all of these adjustments.

In the case of text data, this leads to the generation of several copies of the original text. Each of these variants lacks a certain number of randomly picked words from the original text. Following that, the freshly created fictitious data are grouped into appropriate categories based on the features of the data. As a result, we may reduce the impact of certain keywords on the classification of the selected text depending on whether or not we identified those keywords in the text under consideration. It is now possible to develop models that are more robust and capable of concentrating their adaptation to new data. Aside from that, the gathered insights might be useful in increasing acceptability and confidence in machine learning-enabled applications. In the current scenario, we are able to provide a list of specific words that are used by people with suicidal ideations so that the classified results are acceptable to a general audience, as these give strength in favour of the classification produced.

## 4. Experimentation and Result Analysis

### *4.1. Reddit Mental Health Data Set*

Posts on Reddit are not accompanied by formal clinical diagnoses or other associated covariates. On the other hand, users of subreddit forums create posts under their own names, each of which is organised around a particular subject. It is not possible to get demographic information for any individual subreddit. The majority of individuals who use Reddit are men from the United States who are between the ages of 18 and 49. As part of our work to determine whether or not an individual is suicidal or depressed, we compile a primary dataset. The Python Reddit API was utilised in order to gather this information from the Reddit website.

In particular, we scrape content from r/SuicideWatch and r/Depression. Both of these communities are quite active. More than two thousand messages are included in the collection as a whole. In its totality, the collection comprises more than 60,000 unique data points. Data were collected from two distinct subreddits and then sorted into the categories of depression and suicide. A few of the qualities and properties that make the dataset excellent for usage in our model are the active posts in them. As inputs, we utilise the original post content and the subreddit to which the post was submitted as labels. Both of these variables are derived from the data that was scraped. The posts in r/SuicideWatch that are considered to be suicidal are labelled as such, while the posts in r/Depression that are considered to be depressing are labelled as such.

### *4.2. Noise Removal and Bias Normalisation Analysis*

Table 2 presents the classification of depression and suicidal ideations based on the Reddit postings using the classic methods including SVM, KNN, and CNN (as in the existing literature). Neural Machine Translator for noise reduction consistently removes noise

**Table 3**
Classification performance both before and after noise correction models were applied, expressed as a percentage of the original noisy data.

| Model | Accuracies in % | | |
|---|---|---|---|
| | Noisy Data | Corrected Data | After Fixing Class Imbalance / Applying Bias Normalisation |
| **Proposed Model (SVM+KNN)** | 77.82 | 81.63 | 83.12 |
| **SVM** | 76.46 | 79.43 | 80.50 |
| **DT** | 75.38 | 78.22 | 81.50 |
| **RF** | 78.68 | 80.46 | 82.14 |
| **XGBoost** | 75.42 | 78.46 | 81.46 |
| **CNN** | 76.43 | 80.38 | 82.44 |

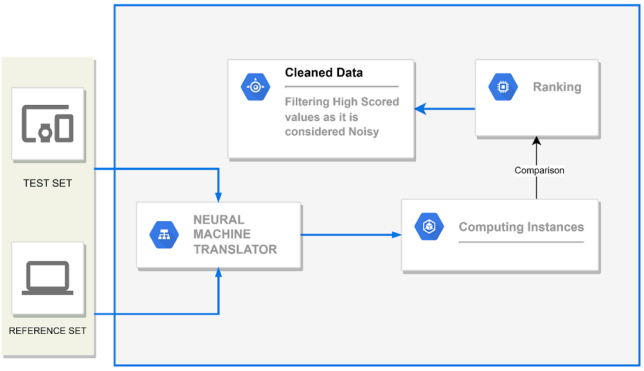



**Fig. 2.** Working of NMT.

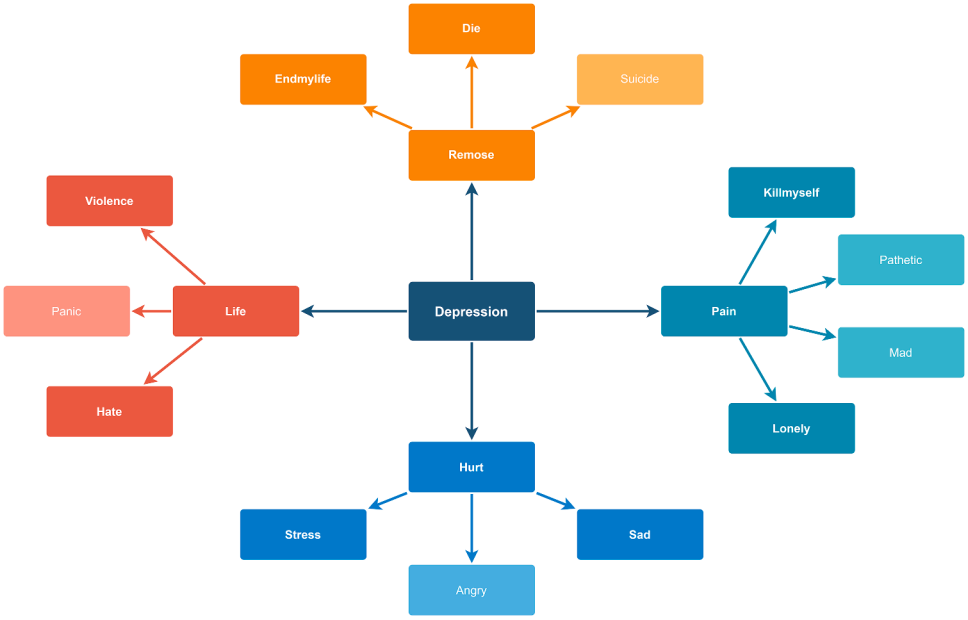



**Fig. 3.** Network representation of the words identified.

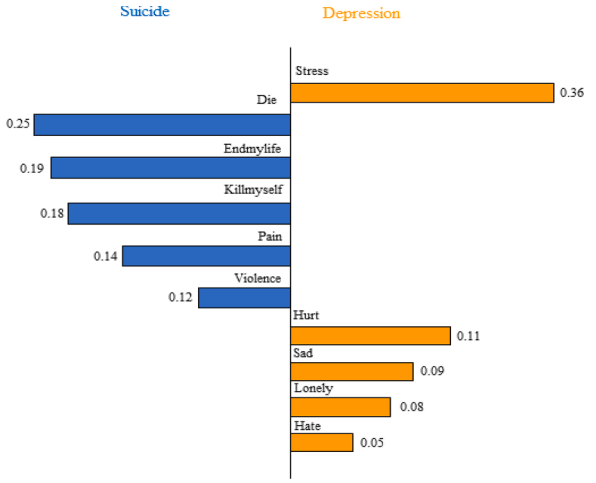
from the data, as presented in Table 3, and minimises a false-positivity rate on the obtained data. Table 3 clearly demonstrates that the prediction accuracies have improved fairly after applying the noise correction, and we can witness a considerable improvement in performance that has increased roughly by 5 percent after the noise has been eliminated. We can also see a further improvement in the prediction accuracy of all the classifiers after the class imbalance has been rectified.

**Table 4**
Evaluation metrics after applying label correction.

| | Proposed Model (SVM+KNN) | Support Vector Machine | Decision Tree | Random Forest | XGBoost | CNN |
|---|---|---|---|---|---|---|
| **Accuracy in %** | 86.28 | 82.60 | 83.57 | 84.98 | 82.46 | 84.77 |
| **Recall** | 87.43 | 81.92 | 77.77 | 83.99 | 84.22 | 88.35 |
| **Precision** | 85.18 | 80.77 | 84.28 | 82.18 | 81.46 | 83.59 |
| **F1-Score** | 84.55 | 81.25 | 80.70 | 82.92 | 81.89 | 83.43 |
| **AUC Score** | 88.83 | 85.43 | 87.11 | 86.35 | 80.56 | 82.36 |

**Table 5**
Major set of words identified using LIME.

| Major Set of Words Identified | Category |
|---|---|
| Stress | #Depression |
| Die | #Suicide |
| Endmylife | #Suicide |
| Killmyself | #Suicide |
| Hurt | #Depression |
| Pain | #Suicide |
| Violence | #Suicide |
| Sad | #Depression |
| Lonely | #Depression |
| Hate | #Depression |




### 4.3. Syntax Based Semantic Network Analysis Visualisation

Semantic Network Analysis was employed to figure out the most frequently used words and the words that had a major impact on the classification task. Fig. 3 depicts the network representation of the words that are identified in our experiments. In Fig. 3, those which are represented in dark shades are the words that came on top of the list and are those which contributed the majority in the clustering process that identified the sentence tone. The rest of the words are the most frequently occurring ones in the r/Depression and r/SuicideWatch on Reddit.

### 4.3. Label Correction and Clustering Analysis

After the pre-filtering of the data, the refined words were listed according to their frequency and the term frequency-inverse document frequency values. Then the general features that have been apparent in the texts of r/SuicideWatch and r/Depression were all identified. After that, we utilised the improved words to construct word matrices, which are also known as co-occurrence matrices, in addition to semantic networks. These matrices are used to map the relationships between words.

Clustering was imposed on each of the two different subreddits pertaining to mental health throughout the training and testing that was done on them. This was done so that the findings could be compared to those of a control group and to the most frequently occurring words and set of domains found using the semantic network analysis. In order to prevent ourselves from overfitting the data, the analysis was limited to including only the two most recent comments contributed by each user. It was established that the ratio of trains to tests as 80-20.

For the purpose of conducting an analysis, we choose a subset of the samples based on a random selection process. We are building a dataset for the binary classification of suicidal text vs. clinically healthy text so that we can compare the effectiveness of our strategy to that of other similar activities and approaches. This will allow us to determine how well our strategy performs in comparison to those other activities and approaches.

The results that are obtained before the inherent noise is reduced are shown in Table 4. It is quite clear that the combination of SVM along with KNN gives the greatest performance in all of the examples that were previously examined, even if the class imbalance was not addressed. As a result, the Label correction process will also be used when we are thinking about applying this approach for future examination.

The first step in deciding whether or not our method of clustering is effective is to demonstrate that the clustering methodology is

**Table 6**
Results of final classification after applying label correction.

| MODELS | Proposed Model (SVM+KNN) | Support Vector Machine | Decision Tree | Random Forest | XGBoost | CNN |
|---|---|---|---|---|---|---|
| **ACCURACY in %** | **98.05** | 84.92 | 86.16 | 86.64 | 88.48 | 89.42 |
| **PRECISION** | **96.77** | 84.66 | 84.16 | 93.43 | 87.42 | 88.69 |
| **RECALL** | **97.55** | 86.47 | 86.09 | 87.17 | 87.38 | 87.58 |
| **F1-SCORE** | **97.33** | 88.43 | 88.61 | 91.56 | 89.47 | 90.47 |
| **AUC** | **97.94** | 82.86 | 83.46 | 83.88 | 87.79 | 89.36 |

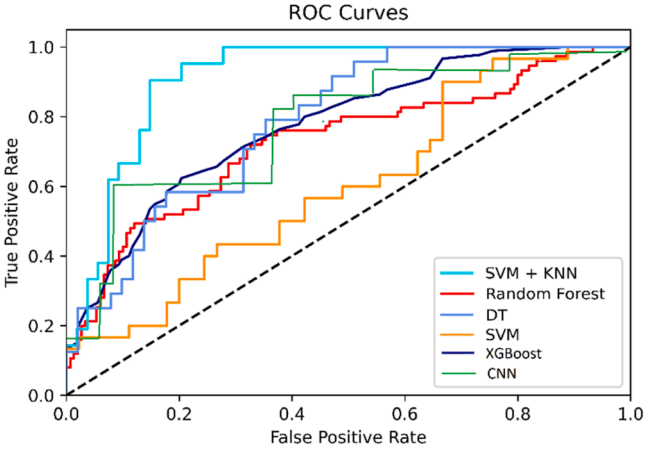



**Fig. 4.** ROC curve of final classification after applying label correction.

trustworthy when it comes to cleaning up noisy labels. It is reasonable to anticipate that the accuracy of classification on a clean test set will significantly decrease with training labels, particularly as it becomes noisy. This is especially the case when the data becomes inconsistent. This is due to the fact that it is logical to predict that training labels would result in the introduction of noise. Therefore, the proper strategy will be successful if, after the application of label correction, our algorithm achieves a higher degree of classification accuracy. This would indicate that the appropriate approach was used. Since the use of label correction, we have obtained a classification that is much more accurate than before, and the level of confidence with which we can provide the clustering has also increased because the class imbalance has been resolved (See Table 4).

Since our study is making use of a dataset that was obtained via the use of the internet, it is crucial to consider that the labels have not been validated by clinical research. This is because we do not have physical access to the labels. We demonstrate that accuracy increases significantly after applying our label correction approach. Our label correction approach is effective for cleaning noisy labels. Furthermore, using a label correction mechanism to clustered data positively affects classifier performance.

### *4.4. Analysis of Posts using LIME*

Prior to classification, we apply LIME on the label corrected data to find the words that have been considered as major factors for classifying the posts into two categories, i.e., with and without suicidal ideations.

Table 5 shows the ten most frequently used words in the postings that were identified by using LIME. Words like "**killmyself**", and "**endmylife**" were the most frequently occurring words used in the posting by those who were actually in a state of suicidal ideation.

### *4.5. Classification Analysis*

The Adam optimizer with a binary cross-entropy loss function allowed us to train the ensemble effectively. Based on the results of the refinement trials, we decided to set the learning rate for all of the studies at 0.85. Accuracy, Precision, Recall, F1-Score, and Area Under Curve (AUC) Scores are the five measurements we use to determine the accuracy of our classifications.

The results of the final classification are shown in Table. 6. The importance of the label correction applied after the class imbalance component was addressed is shown by the fact that every measure is significantly better than the baseline obtained without the label correction in place.

Here, we can see that the proposed ensemble, which is a combination of SVM+KNN algorithms, is able to achieve a classification accuracy of 98.05 per cent, which is substantially greater when dealing with the early stages of data. We can see that bias rectification, and label correction both helped greatly in adjusting the data and optimising the classifier to provide a more robust outcome. It should also be noted that the classification findings may be accepted to a significant degree and implemented in a broad context since we ensured that the results were not biased.

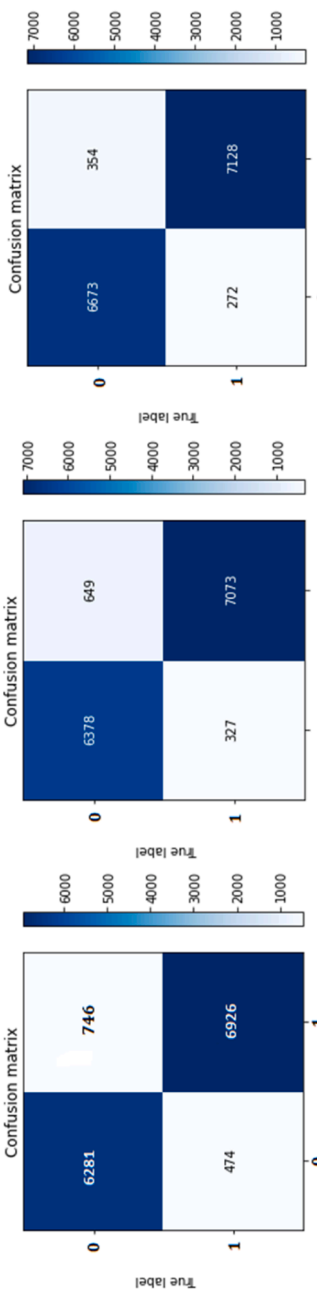



**Fig. 5.** (a)(b)(c): Confusion matrix showing the performance before and after the application of label correction on the proposed method.

The effectiveness of the suggested strategy was evaluated by putting it through its paces on the data from Reddit and analysing the degree of precision with which it categorised the information. It would be good to use labels that were supplied by an expert in mental health in order to offer a comprehensive evaluation of our model. This would ensure that the evaluation is as accurate as possible. Since such labelled data is not available, we are restricted to the available set of labelled data that we obtained from r/Suicide watch as a baseline.

Fig. 5 (a),(b) and (c) present the gain of the classification performance after applying the bias removal and label correction followed by classification. Here we can see a considerable reduction in terms of false positive and false negative predictions.

The class distribution has been determined, as can be seen in Table 1, and the range of possible outcomes is shown in Tables 4 and 6. When we look at those tables, we see that the classification is becoming more accurate as time goes on. This was tested after label correction and class imbalance fixing had been applied. In addition, we can observe that these two changes have resulted in a significant improvement after applying them. As a consequence of this, the strategy that was presented not only provides a higher level of efficiency but also assures that the classes are divided equally and that no one group or collection of groups is biased against the study.

## 5. Conclusion

Early identification is key to treating most diseases, which is also true for those suffering from depression. This paper aimed to identify and evaluate individuals at risk for developing a mental disorder and to look for early warning indicators of suicidal thoughts and behaviours. This paper presents a reliable classifier that can compensate for any inherent bias that may have been introduced during the data gathering process to generate an objective model. Furthermore, the proposed ensemble model which includes clustering and classification presents a better result. A label correction with the NMT further tunes this method, and the bias variance is checked and fixed using the intrinsic explainability methods. We are able to achieve better classification accuracy, as evidenced in the results section, and the so achieved results can be confidently shown since we can ensure the fairness of the facts identified. As this is not a clinical study, the findings should only be utilised for research purposes. However, our approach may be used in a situation where it may give medical practitioners an extra tool for the diagnosis of particular patients. In future, we plan to improve the way mental health themes are categorised by using multiclass classification.

## CRediT authorship contribution statement

**V Adarsh:** Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **P Arun Kumar:** Writing – review & editing, Resources, Visualization. **V Lavanya:** Visualization, Resources, Conceptualization. **G.R. Gangadharan:** Conceptualization, Methodology, Writing – review & editing, Supervision, Resources.

## Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data Availability

The data for this study is publicly available on the Internet.

## Acknowlegdement

We sincerely thank Mr. Anas Abdul Kadher who is a practising clinical psychologist for his suggestions and comments on this paper.

## References

Alambo, A., Gaur, M., Lokala, U., Kursuncu, U., Thirunarayan, K., Gyrard, A., Sheth, A., Welton, R. S., & Pathak, J. (2019). Question answering for suicide risk assessment using Reddit. In *Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019* (pp. 468–473). https://doi.org/10.1109/ICOSC.2019.8665525

Bahdanau, D., Cho, K. H., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1409.0473

Baker, J. K. (1990). Stochastic modeling for automatic speech understanding. *Readings in speech recognition* (pp. 297–307). https://doi.org/10.1016/B978-0-08-051584-7.50028-0

Balani, S., & De Choudhury, M. (2015). Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1373–1378). https://doi.org/10.1145/2702613.2732733

Benton, A., Mitchell, M., & Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 152–162). https://aclanthology.org/E17-1015.

Blei, D., Ng, A., & Jordan, M. (2001). Latent Dirichlet allocation. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems* (Vol. 14). MIT Press. https://proceedings.neurips.cc/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf.

Boettcher, N. (2021). Studies of depression and anxiety using Reddit as a data source: Scoping review. 8(11), e29487. 10.2196/29487.

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health, 3*(2). https://doi.org/10.2196/MENTAL.4822

Breiger, R. L., Boorman, S. A., & Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology, 12*(3), 328–383. https://doi.org/10.1016/0022-2496(75)90028-0

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 1–10). https://doi.org/10.3115/v1/W15-1201

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop* (pp. 31–39). https://doi.org/10.3115/V1/W15-1204

De Choudhury, M., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media, 8(1 SE-Full Papers)* (pp. 71–80). https://ojs.aaai.org/index.php/ICWSM/article/view/14526.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media, 7(1 SE-Full Papers)* (pp. 128–137). https://ojs.aaai.org/index.php/ICWSM/article/view/14432.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2098–2110). https://doi.org/10.1145/2858036.2858207

Fu, K., Cheng, Q., Wong, P. W. C., & Yip, P. S. F. (2013). Responses to a self-presented suicide attempt in social media: A social network analysis. *Crisis: The Journal of Crisis Intervention and Suicide Prevention, 34*, 406–412. https://doi.org/10.1027/0227-5910/a000221

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports, 7*(1), 1–11. https://doi.org/10.1038/srep45141. *2017 7:1*.

Goldman, L., & Lewis, J. (2008). The invisible illness. *Occupational Health, 60*(6), 20–21. https://www.proquest.com/scholarly-journals/invisible-illness/docview/207325578/se-2.

Gruda, D., & Hasan, S. (2019). Feeling anxious? Perceiving anxiety in tweets using machine learning. *Computers in Human Behavior, 98*, 245–255. https://doi.org/10.1016/J.CHB.2019.04.020

Gui, T., Zhang, Q., Zhu, L., Zhou, X., Peng, M., & Huang, X. (2019). Depression detection on social media with reinforcement learning. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11856 LNAI* (pp. 613–624). Springer. doi:10.1007/978-3-030-32381-3_49.

Guo, P. (2011). One-shot decision theory. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 41*(5), 917–926. https://doi.org/10.1109/TSMCA.2010.2093891

Haque, A., Reddi, V., & Giallanza, T. (2021). Deep learning for suicide and depression identification with unsupervised label correction. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12895 LNCS* (pp. 436–447). Springer. doi:10.1007/978-3-030-86383-8_35.

Harris, K. M., McLean, J. P., & Sheffield, J. (2013). Suicidal and online: how do online behaviors inform us of this high-risk population? *Death Studies, 38*(6), 387–394. https://doi.org/10.1080/07481187.2013.768313

Hashimoto, T., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. Eds. In J. Dy, & A. Krause (Eds.), *80. Proceedings of the 35th International Conference on Machine Learning* (pp. 1929–1938). PMLR https://proceedings.mlr.press/v80/hashimoto18a.html.

Heidari, H., & Krause, A. (2018). Preventing disparate treatment in sequential decision making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 2248–2254). International Joint Conferences on Artificial Intelligence Organization. 10.24963/ijcai.2018/311.

Helbig, H. (2006). Knowledge representation and the semantics of natural language. *Cognitive Technologies, 9*. https://doi.org/10.1007/3-540-29966-1

Hiraga, M. (2017). Predicting depression for Japanese blog Text. In *Proceedings of ACL 2017, Student Research Workshop* (pp. 107–113). https://aclanthology.org/P17-3018.

Huang, Y., Goh, T.-T., & Liew, C. L. (2007). Hunting suicide notes in Web 2.0 - Preliminary findings. In *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)* (pp. 517–521). https://doi.org/10.1109/ISMW.2007.43

Hussain, J., Satti, F. A., Afzal, M., Khan, W. A., Bilal, H. S. M., Ansaar, M. Z., Ahmad, H. F., Hur, T., Bang, J., Kim, J. I., Park, G. H., Seung, H., & Lee, S. (2019). Exploring the dominant features of social media for depression detection: *10.1177/0165551519860469*, 46(6), 739–759. 10.1177/0165551519860469.

Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M., & Inkpen, D. (2018). Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 88–97). https://doi.org/10.18653/v1/W18-0609

Jiang, Z., Levitan, S. I., Zomick, J., & Hirschberg, J. (2020). Detection of mental health from Reddit via deep contextualized representations. In *EMNLP 2020 - 11th International Workshop on Health Text Mining and Information Analysis, LOUHI 2020, Proceedings of the Workshop* (pp. 147–156). https://doi.org/10.18653/V1/2020.LOUHI-1.16

Kang, K., Yoon, C., & Kim, E. Y. (2016). Identifying depressive users in Twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)* (pp. 231–238). https://doi.org/10.1109/BIGCOMP.2016.7425918

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. 10(1), 1–6. 10.1038/s41598-020-68764-y.

Li, Tim M. H., Ng, Ben C. M., Chau, Michael, Wong, P. W. C., & Yip, P. S. F. (2013). Collective Intelligence for Suicide Surveillance in Web Forums. In *Pacific-Asia Workshop on Intelligence and Security Informatics* (pp. 29–37). Springer. https://doi.org/10.1007/978-3-642-39693-9_4.

Lokala, U., Srivastava, A., Dastidar, T. G., Chakraborty, T., Akthar, M. S., Panahiazar, M., & Sheth, A. (2022). *A computational approach to understand mental health from reddit: knowledge-aware multitask learning framework*. https://doi.org/10.48550/arxiv.2203.11856

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: Observational study. *Journal of Medical Internet Research, 22*(10), e22635. https://doi.org/10.2196/22635

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing* (pp. 1412–1421). https://doi.org/10.18653/V1/D15-1166

Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social media and suicide: A public health perspective. *American Journal of Public Health, 102 Suppl*(Suppl 2), S195–S200. https://doi.org/10.2105/AJPH.2011.300608

Masuda, N., Kurahashi, I., & Onari, H. (2013). Suicide ideation of individuals in online social networks. *PLOS ONE, 8*(4), e62262. https://doi.org/10.1371/JOURNAL.PONE.0062262

Park, M., McDonald, D., & Cha, M. (2021). Perception differences between the depressed and non-depressed users in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media, 7(1 SE-Full Papers)* (pp. 476–485). https://ojs.aaai.org/index.php/ICWSM/article/view/14425.

Parsons, Lance, Haque, Ehtesham, & Liu, Huan (2004). Subspace clustering for high dimensional data. *ACM SIGKDD Explorations Newsletter, 6*(1), 90–105. https://doi.org/10.1145/1007730.1007731

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights, 2010*(3), 19–28. https://doi.org/10.4137/bii.s4706

Rabani, S. T., Khan, Q. R., & Khanday, A. M. U. D. (2021). Quantifying suicidal ideation on social media using machine learning: A critical review. *Iraqi Journal of Science, 62*, 4092–4100. https://doi.org/10.24996/ijs.2021.62.11.29

Ren, L., Lin, H., Xu, B., Zhang, S., Yang, L., & Sun, S. (2021). Depression detection on reddit with an emotion-based attention network: Algorithm development and validation. *JMIR Med Inform, 9*(7), e28754. https://doi.org/10.2196/28754. *2021;9(7):E28754* https://Medinform.Jmir.Org/2021/7/E28754.

Reynolds, D. (2009). Gaussian mixture models. *Encyclopedia of biometrics* (pp. 659–663). https://doi.org/10.1007/978-0-387-73003-5_196

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., & Herrman, H. (2016). Social media and suicide prevention: a systematic review. *Early Intervention in Psychiatry, 10*(2), 103–121. https://doi.org/10.1111/eip.12229

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *Npj Digital Medicine, 3*(1), 78. https://doi.org/10.1038/s41746-020-0287-6

Sekulic, I., & Strube, M. (2019). Adapting deep learning methods for mental health prediction on social media. In *W-NUT@EMNLP 2019 - 5th Workshop on Noisy User-Generated Text, Proceedings* (pp. 322–327). https://doi.org/10.18653/V1/D19-5542

Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety through reddit. In *Proceedings Ofthe Fourth Workshop on Computational Linguistics and Clinical Psychology* (pp. 58–65). https://doi.org/10.18653/V1/W17-3107

Slemon, A., McAuliffe, C., Goodyear, T., McGuinness, L., Shaffer, E., & Jenkins, E. K. (2021). Reddit users' experiences of suicidal thoughts during the COVID-19 pandemic: A qualitative analysis of r/Covid19_support posts. *Frontiers in Public Health, 9*, 1175. https://doi.org/10.3389/FPUBH.2021.693153

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Xu, Z., Pérez-Rosas, V., & Mihalcea, R. (2020). Inferring social media users' mental health status from multimodal information. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, May* (pp. 6292–6299). France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.772.

Young, S. D., & Garett, R. (2018). Ethical issues in addressing social media posts about suicidal intentions during an online study among youth: Case study. *JMIR Ment Health, 5*(2), e33. https://doi.org/10.2196/mental.8971

Zhang, Xueru, Khaliligarekani, Mohammadmahdi, Tekin, Cem, & Liu, Mingyan (2019). Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness. In *, 32. Advances in Neural Information Processing Systems*. NeurIPS. https://arxiv.org/abs/1905.00569.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1–4), 43–52. https://doi.org/10.1007/S13042-010-0001-0

Zirikly, A., Resnik, P., Uzuner, Ozlem, & Hollingshead, K (2019). CLPsych 2019 Shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (pp. 24–33). https://doi.org/10.18653/V1/W19-3003