Cost-sensitive Boosting Pruning Trees for depression detection on Twitter

Lei Tong, Zhihua Liu, Zheheng Jiang , Feixiang Zhou, Long Chen, Jialin Lyu, Xiangrong Zhang *Senior Member, IEEE*, Qianni Zhang, Abdul Sadka *Senior Member, IEEE*, Yinhai Wang, Ling Li and Huiyu Zhou

***Abstract*—Depression is one of the most common mental health disorders, and a large number of depressed people commit suicide each year. Potential depression sufferers usually do not consult psychological doctors because they feel ashamed or are unaware of any depression, which may result in severe delay of diagnosis and treatment. In the meantime, evidence shows that social media data provides valuable clues about physical and mental health conditions. In this paper, we argue that it is feasible to identify depression at an early stage by mining online social behaviours. Our approach, which is innovative to the practice of depression detection, does not rely on the extraction of numerous or complicated features to achieve accurate depression detection. Instead, we propose a novel classifier, namely, Cost-sensitive Boosting Pruning Trees (CBPT), which demonstrates a strong classification ability on two publicly accessible Twitter depression detection datasets. To comprehensively evaluate the classification capability of CBPT, we use additional three datasets from the UCI machine learning repository and CBPT obtains appealing classification results against several state of the arts boosting algorithms. Finally, we comprehensively explore the influence factors for the model prediction, and the results manifest that our proposed framework is promising for identifying Twitter users with depression.**

***Index Terms*—data mining, boosting ensemble learning, online depression detection, online behaviours.**

I. Introduction

Depression is one of the most common mental illnesses. It is estimated that nearly 360 million people suffer from depression [1].In Britain, 7.8% of people meet the criteria of depression diagnosis, and 4-8% will experience depression in their lifetime [2]. Andrade et al. [3] reported that the probability for an individual to encounter a major episode of depression within a period of one year is 3-5% for males and 8-10% for females. Because of depression, about one million of people committed suicide annually in the world [1].

Depressed people may have a variety of symptoms: having troubles in going to sleep or sleeping too much, lacking of passion or feeling disappointed [4]. In clinical exercises, psychological specialists are looking for reliable methods to detect and prevent depression. Yang et al. [5] investigated the relation between vocal prosody and changes in depression severity over time. Alghowinem et al. [6] examined human behaviours such as speaking behaviours and eye activities associated with major depression. Diagnostic and Statistical Manual of Mental Disorders [7] is an important reference for psychological doctors to diagnose depression. There are nine classes of depression symptoms recorded in the menu, describing the distinguishing behaviours in our daily life. Nevertheless, the symptoms of depression disorders evolve over time and it has been advised to dynamically update the criteria of depression diagnosis [1].

On the other hand, depression sufferers who do not receive timely psychotherapy will develop worse conditions. More than 70% of people in the early stage of depression do not consult psychological doctors, and their conditions deteriorated [7]. González-Ibánez et al. [8] reported that people are somehow ashamed or unaware of depression which makes them miss timely treatment. Choudhury et al. [9] and Neuman et al. [10] proposed to explore the correlation of depression sufferers with their online behaviours on social networks. With the explosive growth of computer network applications, social networks have become an indispensable part of many people's daily lives. 62% of the American adults (age 18 and older) use Facebook, whilst the majority of the users (70%) visit Internet daily and a large portion of the users access to Internet multiple times each day [11]. There are 1.10 billion posts on Facebook every day. Twitter and Tumblr also have 500 and 77.5 million users who are active per day, where 70% of the Twitter users log in every day [11]. Therefore, social networks provide a means for capturing behavioural attributes that are relevant to an individual's thinking, mood, communication, activities and socialisation [9]. Research studies reveal that collecting social networking information for analysing human physical

L. Tong, Z. Liu, Z. Jiang, F. Zhou, L. Chen, J. Lyu and H. Zhou are with School of Computing and Mathematical Sciences, University of Leicester, United Kingdom. E-mail: {lt228;zl208;zj53;fz64;lc408;jl766;hz143}@leicester.ac.uk. H. Zhou is corresponding author.

X. Zhang is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, China. E-mail: xrzhang@mail.xidian.edu.cn.

Q. Zhang is with School of Electronic Engineering and Computer Science, Queen Mary, University of London, United Kingdom. E-mail: qianni.zhang@qmul.ac.uk.

A. Sadka is with Centre for Media Comms Research, Brunel University London, United Kingdom. E-mail: abdul.sadka@brunel.ac.uk.

Y. Wang is with Discovery Sciences, AstraZeneca R&D, Darwin Building, Cambridge Science Park, Milton Road, Cambridge CB4 0WG. E-mail: yinhai.wang@astrazeneca.com.

L. Li is with School of Computing, University of Kent. E-mail: C.Li@kent.ac.uk.

Manuscript received on August 2020; revised xxxx.

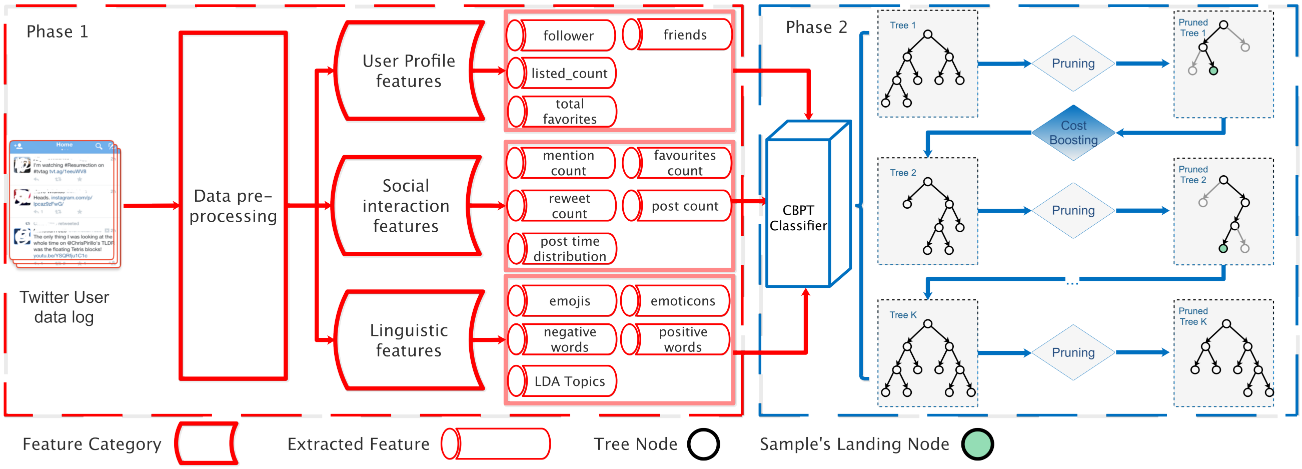


Fig. 1: Proposed framework. In phase 1, we conduct data preprocessing and extract various discriminative features of Twitter users. In phase 2, the CBPT classifier combines the power of K pruned trees. The cost-sensitive boosting structure relies on the landing position of samples in the pruned tree structure and an example of the sample decision path is highlighted in dark black in the diagram.

and mental wellness is possible [12], [13]. Neuman et al. [10] developed working methods for recognising associated signals in the user's posts on social networks, which suggest whether or not clinical diagnosis is required, based on his/her naturally occurring linguistic behaviours. Salawu et al. [14] detected cyber-bullying on social networks by comparing textual data against the identified traits. Nguyen et al. [15] utilised psycholinguistic clues to conduct sentiment analysis on users' posts to detect depressed users online. Hence, it is feasible to detect depression via social networks.

Our proposed framework is shown in Fig. 1. In the first phase, we conduct data preprocessing and extract discriminative features from the posts of Twitter users, while the second phase presents a new Cost-sensitive Boosting Pruning Trees method based on the Discrete Adaboost [16] to classify the users. Our new contributions reported in this paper are:

(1) We propose a novel resampling weighted pruning algorithm which dynamically determines optimal depths/layers and leaves of a tree model. The pruning procedure can support the boosting training and improve the robustness of the base tree estimator.
(2) We combine the proposed pruning process with a novel cost-sensitive boosting structure within an ensemble framework, namely Cost-sensitive Boosting Pruning Trees (CBPT). By introducing cost items into the learning procedure of the boosting paradigm, we highlight the uneven identification importance among the samples so that the boosting paradigm intentionally biases the learning towards the samples associated with higher identification importance.
(3) We conduct comprehensive experiments to justify the significance of our proposed framework against two Twitter depression detection datasets, i.e. Tsinghua Twitter Depression Dataset (TTDD) and CLPsych 2015 Twitter Dataset (CLPsych2015). The experimental results demonstrate that the prediction results are explainable against the ground-truth and our proposed framework can effectively identify Twitter users with depression.

II. Related Work

In the literature, questionnaire or online interview is one of the common means used in depression diagnosis. Lee et al. [17] investigated whether or not interviewees have depressive trends using a choice questionnaire. Park et al. [18] conducted a face-to-face interview with 13 active Twitter users to explore their depressive behaviours. These questionnaires and interviews have several limitations. For example, they are time-consuming and hard to be generalised. On the other hand, because of the explosive growth in the popularity of social networks, online depression detection has attracted large interests in recent years.

Many research studies for online depression detection have focused on feature detection. Choudhury et al. [9] introduced measures (e.g. egocentric social graphs and description of anti-depressant medications) to quantify the online behaviors of an individual for a year before s/he reports the onset of depression. Park et al. [19] explored the use of languages in describing depressive moods using real-time moods captured from Twitter users. Saha et al. [20] analysed the content information of depressed users' posts by extracting topical features. Most recently, Shen et al. [7] extracted six groups' features such as user profile and engagement with online application programming interface (API) to interpret the online behaviours of depression users. However, most previous research studies focus on exploring new features of depression behaviours whilst ignoring the fitness of the classification models.

Shen et al. [7] presented a multi-modal depressive dictionary learning model (MDDL) which combines sparse dictionary learning with logistic regression to identify depressed users. Nadeem et al. [21] conducted experiments to classify Major Depression Disorder (MDD) using four binary classifiers, e.g. decision tree and naive bayes. Also, Choudhury et al. [9] and Shuai et al. [22] proposed a depression detection framework based on support vector machine. Nevertheless, these established classifiers cannot achieve consistent performance due to noise or errors in the data.

In recent years, deep learning based methods for online depression detection attracted the attention of researchers. For example, Shen et al. [23] proposed a cross-domain depression detection framework which transfers the knowledge of Twitter to classify the instances of Weibo. Their proposed method aims to improve the recognition performance in the poorly labeled target domain (Weibo) utilising the rich data of the source domain (Twitter). Ray et al. [24] proposed a multi-level attention network that combines the text, audio and video features to classify depressed people. Gamaarachchige et al. [25] proposed a multi-task, multi-channel and multi-input framework that fuses multiple input features (e.g. emotion labels, tokens) and learns knowledges from multiple classification tasks. Their proposed method achieved good performance in the CLPsych 2015 dataset [26]. Orabi et al. [27] proposed a word embedding optimisation method which combines multiple word embedding features (e.g. Skip-Gram, CBOW). They used this technique to extract text features from the Twitter users' posts and identified the depressed users. These deep learning based method can achieve promising performance on depression detection especially on multi-level feature fusion and knowledge transfer. However, these methods lack clear interpretations to the model predictions as of which specific factor influences the predicted depression risk.

Decision trees based ensemble learning brings up the possibility of developing a powerful and interpretable model. Decision trees can reveal the feature effects to the prediction and ensemble learning uses multiple learning algorithms to obtain better predictive performance than that of using any of the constituent learning algorithms alone [28]–[31]. Our framework is based on Adaboost which is one of the typical ensemble meta-algorithms to reduce biases and variances in supervised learning [32]. In general, Adaboost employs decision dump as its base estimator. However, decision dump cannot fit well the training data because of its simple structure. Adaboost with decision dump does not perform well in complex datasets [30]. Boonyanunta et al. [33] proposed a method to improve Adaboost's performance by averaging the estimators' weights or reordering estimators. Based on Adaboost, Friedman et al. [34] reported Gradient Boost Decision Trees (GBDT) which is the generalisation of boosting to arbitrary differentiable loss functions. Unfortunately, GBDT can be over-fitting if the data is noisy and the training process of GBDT is time consuming. Chen et al. [35] introduced an advanced Gradient Boost algorithm (called 'XGboost') based on GBDT in 2016. Although XGboost is more flexible and efficient than GBDT, it has many parameters that are hard to tune.

In this paper, we propose a novel classification algorithm based on Adaboost that can mitigate the influence of noise or errors and have a strong fitness and generalisation ability. We introduce the details of the proposed algorithm in Section 4. In addition, we summarise the discussed classification methods in Table S4, Supplementary A.

III. Data Preprocessing and feature extraction

In this paper, we intend to analyse depression users' online behaviours. As the scripts on social networks may be random and unpredictable, features with different noise may be obtained and influence the detection accuracy. Before feature extraction is implemented, we carry out the following preprocessing procedure: (1) Minimisation of the influence of noisy samples. Inspired by the work of Yazdavar et al. [36], we remove the noisy samples from the dataset where the posting number of the samples is less than five. These samples cannot provide sufficient information for analysing the users' behaviours or topic modelling. (2) Processing of irregular words. The words on social networks may look irregular because of mistaken spelling or abbreviations. We use the Textblob API reported in [37] (commonly used in natural language processing tasks) to remedy the wrong type of words. (3) Stemming. We expect to perform statistical analysis on commonly used words of control and depressed users separately and conduct topic modelling on the users' posts. Words must be of unified representations regardless of tense and voice. Hence, we utilise the SnowballStemmer algorithm reported in [38] to deal with these words. For instance, "accepting" and "accepted" can be converted to "accept". Afterwards, we extract three feature categories as follows and the proposed framework is shown in Phase 1 of Fig.1.

(1) User's Profile Features: The user's profile features contain the user's individual information on social networks. We collect 4 different features here: $total\_favourites$ reflects the number of posts that this particular user favours during his/her account's lifetime; $listed\_count$ shows the number of the public list that this user holds a membership within. We collect the number of the user's $friends$ and $followers$ which well characterise the author's egocentric social networks.

(2) Social Interaction Features: Park et al. [19] discovered that depressed users are less active in social networks, and depressed users regard social networking as a tool for social awareness and emotional interaction. Thus, we extract $retweet\,count$, $mention\,count$ (e.g. @someone) and $favourites\,count$ (indicating how many times this post has been favoured by the other users) to describe the behaviours of the user interacting with others. Besides, we collect the $posting\,number$ and $time\,distribution$ to demonstrate the user's activeness on social networks.

(3) Linguistic Features: The content of the posts on social networks can intuitively reflect a person's mood and attitude. Depressed users may post more negative words than control users [7], [9], [19], [39]. Hence, we count the numbers of $negative$ and $positive\, words$ in the tweets using the NLTK toolkit [40]. In addition, we collect the numbers of $emoj$i and $emoticons$ from the texts to form relevant features. In order to comprehensively explore the semantics, Resnik et al. [41] examined the difference of the concerned topics between depressed and control users by topic modeling and observed that topic modeling might be effective for depression detection. In our work, we utilise the Latent Dirichlet Allocation (LDA) approach presented in [42] to extract $topic\, distributions$ from the tweets.

Finally, the extracted feature sets are used to train our proposed classifier CBPT, which is shown in Phase 2 of Fig.1 and we provide the details of the extracted feature dimensionality in Table S1, Supplementary A.

IV. Proposed Method

A. Discrete Adaboost

Our classification algorithm is built upon the discrete Adaboost algorithm proposed by Freud et al. [16]. Algorithm 1 presents the baseline scheme of the discrete Adaboost that combines many simple hypotheses (called weak learners) to form a strong classifier for the task [30]. The algorithm can be summarised as follows: (1) Training multiple base classifiers sequentially and assigning a weight value $ln(\beta_m)$ according to its training error ε_m. (2) The samples misclassified by the preceding classifier are assigned a higher weight $w_{m+1,i}$, which will let the classifier pay more attention to these samples. (3) Finally, combining all the weak classifiers with their weights to obtain an ensemble classifier $G(X)$. As we have discussed above, Adaboost may not perform well on a complex dataset, and hence we propose the CBPT algorithm to improve the performance of Adaboost in two aspects: (1) We improve the fitting and generalisation ability of the base classifier. (2) We propose a novel boosting structure to strengthen the sample re-weighting process.

Algorithm 1 Discrete Adaboost algorithm.

Input: A training set $D = \{(X_i, y_i)\}_{i=1}^{N}$.
Output: A model $M_K(X)$ which is based on K decision trees with their corresponding weight.

```
procedure ADABOOST($D$)
    Initialise sample weight distribution $W = \left\{ \left( w_k^{(i)} \right) \right\}$.
    Set each sample's weight $w_k^{(i)}$ to $\frac{1}{N}$.
    for $k \in (1, K)$ do
        Fit an estimator $M_k(X)$ to the training data with $W_k$.
        Let $u_i = 1$ if the i-th case is classified incorrectly, otherwise zero.
        Compute training error $\varepsilon_k = \sum_{i=1}^{N} w_k^{(i)} u_i$.
        Update sample's weight $w_{k+1}^{(i)} = \frac{w_k^{(i)} \beta_k}{\sum_{i=1}^{N} w_k^{(i)} \beta_k}$, where $\beta_k = \frac{(1-\varepsilon_k)}{\varepsilon_k}$.
        $M_k(X) \leftarrow M_{k-1}(X) + \log_e(\beta_k) M_k(X)$.
    end for
    return $M_K(X)$
end procedure
```

B. Cost-sensitive Boosting Pruning Trees

In this section, we propose an ensemble method that combines an improved Adaboost algorithm with pruned decision trees for classification. Here, we still employ a decision tree as the base estimator because of its flexibility and interpretability. Decision dump often suffer from underfitting whilst a full tree has a high variance. We here consider pruning trees in order to increase system generalisation. In our algorithm, we firstly apply all the training samples and allow a decision tree to fully grow, and then use the cost-complexity pruning method reported in [43] to prune certain branches of the trees and use the modified criterion to evaluate the system performance with the pruned trees and update the weights. Afterwards, the above steps will be executed iteratively till the maximum number of the trees is reached. To formulate our algorithm, we here declare the used notations in advance. In particular, we denote the training dataset as $D = \{(X_i, y_i)\}_{i=1}^{N}$, and $X_i^{(v)} \in \mathbb{R}^{N \times V}$ is the sample feature vector where N represents the set size and V is the feature dimension. y_i represents the training target. We employ $W = \left\{ \left(w_k^{(i)} \right) \in \mathbb{R}^N \right\}_{k=1}^{K}$ to represent the set of the sample weight distribution. K is the number of the estimators (iterations) and each sample's weight is initialised to $\frac{1}{N}$ in the first iteration during the normalisation. Furthermore, we use θ_k and $M_K(X)$ to denote the k-th estimator's weight and the ensemble classifier.

1) Resampling Weighted Pruning Algorithm: In most of the previous boosting algorithms [34], [35], [44], except $num\ trees$, $max\ depth$ and $num\ leaves$ are two key hyperparameters which affect the classifier's performance significantly. Manually tuning the hyperparameter combinations is a heavy task and it is hard to find the best parameter combinations for different datasets. Therefore, we propose a novel technique called resampling weighted pruning to automatically prune redundant leaves and produce robust tree models, where weights are used to establish a relationship between the pruning and boosting practice.

Firstly, we denote the original learning sample set D which is divided randomly into S subsets, $\{D_s\}_{s=1}^{S}$ and the training set of each subset is $D^{(s)} = D - D_s$. The tree T_{max} comes from the original set D and we build a complete tree on each subset $D^{(s)}$. We present the cost function of the decision trees as follows:

$$\mathcal{L}(T; w_k) = \sum_{|\tilde{T}|} \left[1 - \sum_{c=1}^{C} \left(\frac{\sum_{i_c} w_k^{(i_c)}}{\sum_i w_k^{(i)}} \right)^2 \right] \tag{1}$$

where $\left|\tilde{T}\right|$ is the leaves' number, C denotes the class number and the sample of class c is defined as i_c. The loss of the decision trees is the sum of all the leaf nodes' gini impurity [45]. A complete tree's loss $\mathcal{L}(T_{max}; w_k)$ is zero because each leaf node only includes a single class's samples. But $\mathcal{L}(T; w_k)$ will increase in the pruning process where the pruned nodes are merged with their parents' nodes. Therefore, the present cost function is not a good measure of selecting a subtree because it always favours large trees. Thus, the penalty term, regularization parameter α and the tree leaves $\left|\tilde{T}\right|$ are added to the cost function. The new cost function is defined as follows:

$$\mathcal{L}_\alpha(T; w_k) = \mathcal{L}(T; w_k) + \alpha\left|\tilde{T}\right| \tag{2}$$

The penalty term favours a simple tree when α is constant and $\left|\tilde{T}\right|$ decreases with pruning.

Now, the variation in the cost function is given by $\mathcal{L}_\alpha(T-T_t; w_k) - \mathcal{L}_\alpha(T; w_k)$, where T_t represents a branch with the node at t and a tree pruned at node t would be $T - T_t$. Next, the cost of the pruning on the internal nodes is calculated by equating $\mathcal{L}_\alpha\left(T - T_t; w_k\right)$ to that of the branch at node t:

$$\begin{aligned}
&\mathcal{L}_\alpha(T - T_t; w_k) - \mathcal{L}_\alpha(T; w_k) \le 0\\
&\Rightarrow \mathcal{L}_\alpha\left(t; w_k\right) - \mathcal{L}_\alpha\left(T_t; w_k\right) \le 0\\
&\Rightarrow \mathcal{L}(t; w_k) + \alpha - \mathcal{L}(T_t; w_k) - \alpha\left|\tilde{T}_t\right| \le 0\\
&\Rightarrow \frac{\mathcal{L}(t; w_k) - \mathcal{L}(T_t; w_k)}{\left|\tilde{T}_t\right| - 1} \le \alpha
\end{aligned} \tag{3}$$

We define:

$$g\left(t\right) = \frac{\mathcal{L}(t; w_k) - \mathcal{L}(T_t; w_k)}{\left|\tilde{T}_t\right| - 1} \tag{4}$$

We will prune branch T_t with the decrease of the cost function value when $\alpha \ge g\left(t\right)$. The order of pruning is performed by setting $\alpha = \arg\min g(t)$ in order to find the suitable branch, which should be pruned, and the process will be repeated until the tree is left with the root node only. This provides a sequence of subtrees $\left\{(T_j^{(s)});\right\}_{j=1}^{J}$ with the associated cost-complexity parameters $\{(\alpha_j); \forall\alpha \in \mathbb{R}\}_{j=1}^{J}$ where J is the length of the subtree sequence.

For α, we apply the pruned tree $T_j^{(s)}$ to predicting the estimations in the s-th test set, resulting in the following error rate:

$$\mathcal{TE}_j^{(s)} = \frac{\sum_{i_{miss}} w_k^{(i_{miss})}}{\sum_i w_k^{(i)}} \tag{5}$$

where i_{miss} denotes the index of the misclassified sample's weight, $w_k^{(i)}$ is the sample's weight of the test set D_s and $\mathcal{TE}_\alpha^{(s)}$ represents the misclassified rate of set D_s. Hence, the average misclassified rate of S is:

$$\mathcal{TE}_j = \frac{1}{s}\sum_{s=1}^{S}\mathcal{TE}_\alpha^{(s)} \tag{6}$$

and we define

$$\alpha^* = \arg\min_{\alpha_j}\mathcal{TE}_j, \quad \exists\alpha_j > 0 \tag{7}$$

which is the best pruned tree obtained by pruning T_{max} till $\mathcal{L}_{\alpha^*}(T_{max}; w_k)$ reaches the minimum. The pseudocode of our resampling weighted pruning algorithm is shown in Algorithm 2.

Algorithm 2 Resampling Weighted Pruning Algorithm.

```
Input: A training set $D$ with corresponding weight $W_k$.
Output: A pruned tree estimator $M_k\left(X\right)$.
function BestPrunedTree($D, W_k$)
    Randomly split the learning samples $D$ into $S$ folds, $\{D_s\}_{s=1}^{S}$.
    Grow a decision tree $T_{max}$ on the whole set $D$.
    for $s \in [1, S]$ do
        Fit a decision tree $T^{(s)}$ to subset $D^{(s)}$.
        Generate subtree sequence $\left\{(T_\alpha^{(s)}); \forall\alpha \in \mathbb{R}\right\}$ by Eq. (3).
        Generate subtree sequence $\left\{(T_j^{(s)});\right\}_{j=1}^{J}$ ←
            { 1. Calculate $g\left(t\right)$ using Eqs. (3)-(4)
            { 2. Set $\alpha = \arg\min\, g\left(t\right)$ and prune the branch $T_t$
            { 3. Recursively repeat till the tree only has root nodes
        Calculate $\mathcal{TE}_j^{(s)}$ ← Eq. (5).
    end for
    Compute average error rate $\mathcal{TE}_j$ against each substree.
    $\alpha^* = \arg\min_{\alpha_j}\mathcal{TE}_j; (\exists\alpha > 0)$.
    The best pruned tree estimator $M_k\left(X\right)$ ←
        Prune $T_{max}$ till $\mathcal{L}_{\alpha^*}(T_{max}; w_k)$ becomes minimal.
    return $M_k\left(X\right)$.
end function
```

2) Tree-based Cost-sensitive Boosting Structure: As shown at steps 7 and 8 of Algorithm 1, Adaboost employs the training error ε_m as the evaluation criterion of the base estimator's performance, to set up the estimator's weights and update the samples' weights. All the misclassified samples receive the same weights in each iteration. In general, we assume that misclassified samples should be given different weights according to the "hardness" of the samples - harder samples are of more weights. We now propose a novel boosting architecture namely Tree-based Cost-sensitive Boosting which utilizes the tree model to assess the "hardness" of the training samples and optimize the boosting process.

In the first step, we apply a complete decision tree to the training data D and prune it in order to obtain the best tree estimator $M_k(X)$. A complex tree model has more depths. Similarly, the deeper the landing node of a sample is, the harder the sample can be classified. Here, we present a new and effective depth penalty term as follows:

$$\mathcal{DP}_k^{(i)} = \frac{\psi_d(\sigma_k^{(i)} - \min(\sigma_k))}{\max(\sigma_k) - \min(\sigma_k)} + \eta_d;\ \psi_d \in \mathbb{N}^+, \eta_d \ge 1 \tag{8}$$

where $\sigma_k^{(i)}$ represents the landing node depth of sample i, $\max(\sigma_k)$ and $\min(\sigma_k)$ are the maximum and minimum values in the node depth array σ_k. ψ_d and η_d are two hyper-parameters where ψ_d is the percentage of data scaling, and η_d is the lower limit of the penalty term. The depth penalty term is a coefficient that is multiplied with the original sample's weight to enable hard samples to gain more weights in the next iteration.

The landing node's depth can be regarded as the global evaluation of the "hardness" of a sample associated with the tree structure. In the pruning procedure, the pruned samples are included in the parent node of the pruned branch. Here, we use node impurity to represent the local evaluation of the "hardness" of a sample. For instance, when two samples land in different leaf nodes but with the same depth, the sample of low node impurity will be given more weights as the sample is separated from the most samples of the same class in the feature space. Hence, the impurity penalty term $\mathcal{IP}_k^{(i)}$ is defined as follows:

$$\text{\oe}_k^{(i)} = \frac{N-\mu_k^{(i)}}{2N} - E_{p(x)}^{(i)}\left[\log q(x)\right]\frac{\mu_k^{(i)}-N}{N} \tag{9}$$

$$\begin{aligned}\mathcal{IP}_k^{(i)} =& (\left\|\mathcal{DP}_k^{(i)}\right\|_{+\infty} - \left\|\mathcal{DP}_k^{(i)}\right\|_{-\infty})\frac{(\text{\oe}_k^{(i)} - \min(\text{\oe}_k))}{\max(\text{\oe}_k)-\min(\text{\oe}_k)} \\ &+ \left\|\mathcal{DP}_k^{(i)}\right\|_{-\infty}\end{aligned} \tag{10}$$

Eq. (9) is an inverse transformation of the impurity value, where μ_k^i is the sample number in the landing node, $E_{p(x)}^{(i)}\left[\log q(x)\right]$ is the impurity value either Cross Entropy or Gini Impurity, $p(x)$ and $q(x)$ are the predicted probability distributions of the sample X_i. Similarly, in Eq. (10), we employ the data scaling for $\text{\oe}_k^{(i)}$ and obtain the impurity penalty term $\mathcal{IP}_k^{(i)}$, $\left\|\mathcal{DP}_k^{(i)}\right\|_{+\infty}$ and $\left\|\mathcal{DP}_k^{(i)}\right\|_{-\infty}$ are positive and negative infinity norms of the depth penalty vector which are used to limit the range of data scaling.

The proposed two penalty terms mainly rely on the landing positions of the samples in the pruned tree structure. Algorithm 3 returns the learning sample position by recursively following the sample's decision path as follows:

Algorithm 3 Recursively Find Landing Node.

Input: Node id l, a learning sample (X_i, y_i)
Output: Node depth $\sigma_k^{(i)}$, node sample number $\mu_k^{(i)}$

```
function TreeRecurse($l, (X_i, y_i)$)    ▷ Find the tree node where the sample land
    if $Node_l == Leaf$ then    ▷ Check if node $l$ is a leaf
        return $\sigma_k^{(i)}$, $\mu_k^{(i)}$
    else
        if $X_i^{(v)} < Threshold^{(v)}$ then    ▷ Determine if the sample flow down to left or right child
            return TreeRecurse $(a_l, (X_i, y_i))$
        else
            return TreeRecurse $(b_l, (X_i, y_i))$
        end if
    end if
end function
```

In each iteration k, the ensemble boosting aims to minimise an exponential loss function, described by:

$$\widetilde{L}(M) = \sum_{i=1}^{N} \exp[-y_i(M_{k-1}(X_i) + \theta_k M_k(X_i))] \tag{11}$$

where $M_{k-1}(X_i)$ represent the $k-1$ trained pruned trees and $w_k^{(i)} = \exp(-y_i M_{k-1}(X_i))$, θ_k is the estimator weight of the k-th pruned tree. We can calculate the first order partial derivative of $\widetilde{L}(M)$ with respect to the estimator weight θ_k:

$$\begin{aligned}\frac{\partial \widetilde{L}(M)}{\partial \theta_k} &= \frac{\partial}{\partial \theta_k}\sum_{i=1}^{N} w_k^{(i)} \exp(-y_i \theta_k M_k(X_i)) \\ &= \frac{\partial}{\partial \theta_k}\sum_{i|y_i \neq M_k(X_i)} w_k^{(i)} e^{\theta_k} + \sum_{i|y_i = M_k(X_i)} w_k^{(i)} e^{-\theta_k} \\ &= \frac{\partial}{\partial \theta_k}((1-\varepsilon_k)e^{-\theta_k} + \varepsilon_k e^{\theta_k}) \\ &= (\varepsilon_k - 1)e^{-\theta_k} + \varepsilon_k e^{\theta_k}\end{aligned} \tag{12}$$

By taking zero to the left hand side of Eq.(12), we have:

$$\theta_k \propto \frac{1}{2}(\log\frac{(1-\varepsilon_k)}{\varepsilon_k} + \log(C-1)) \tag{13}$$

where ε_k is the training error of pruned tree $M_k(X)$. $\log(C-1)$ is a regularisation term and C is the number of classes.

The updating process of the new sample's weights is defined as:

$$\begin{aligned}w_{k+1}^{(i)} &= \frac{\exp[-y_i(M_{k-1}(X_i) + \theta_k M_k(X_i))]}{Z_{k+1}} \\ &= \frac{w_k^{(i)}}{Z_{k+1}}\exp[2\theta_k \log(\mathcal{DP}_k^{(i)})\log(\mathcal{IP}_k^{(i)})1_{y_i \neq M_k(X_i)}]\end{aligned} \tag{14}$$

where Z_k is a normalisation factor, and

$$\begin{aligned}Z_{k+1} = &\sum_{i|y_i \neq M_k(X_i)} w_k^{(i)}\mathcal{DP}_k^{(i)}\mathcal{IP}_k^{(i)}\frac{(1-\varepsilon_k)(C-1)}{\varepsilon_k} + \\ &\sum_{i|y_i = M_k(X_i)} w_k^{(i)}\end{aligned} \tag{15}$$

The two penalty terms are taken as the interference factors to influence the updating of the sample's weights and the misclassified samples employ different weights according to their landing positions in the pruned tree structure. The iterative training of the cost-sensitive boosting will stop if it converges (i.e. ε_k reaches zero) or we reach the maximum iteration number K. The whole algorithm is illustrated in Algorithm 4, linked with the two proposed functions.

Algorithm 4 Cost-sensitive Boosting Pruning Trees Algorithm.

Input: A training set $D = \{(X_i, y_i)\}_{i=1}^{N}$ with sample distribution $W = \left\{\left(w_k^{(i)}\right) \in \mathbb{R}^N\right\}_{k=1}^{K}$

Output: A Cost-sensitive Boosting Pruning Trees model $M_K(X)$

```
procedure CostBoosting($D, W$)
    Initialize sample weight distribution $W = \left\{\left(w_k^{(i)}\right)\right\}$.
    Set each sample's weight $w_k^{(i)}$ to $\frac{1}{N}$.
    for $k \in (1, K)$ do
        $M_k(X) \leftarrow$ BestPrunedTree$(D, W_k)$
        for $i \in (1, N)$ do
            $\sigma_k^{(i)}$, $\mu_k^{(i)}$ $\leftarrow$TreeRecurse$(0, (X_i, y_i))$ ▷ Start from the root node.
            Calculate depth penalty coefficient $DP_k^{(i)}$ using Eq. (8).
            Calculate impurity penalty coefficient $IP_k^{(i)}$ using Eqs. (9)-(10).
        end for
        Update the estimator weight using Eq. (13).
        Update each sample's weight $w_{k+1}^{(i)}$ using Eq. (14).
        $M_k(X) \leftarrow M_{k-1}(X) + \theta_k M_k(X)$
    end for
    return $M_K(X)$.
end procedure
```

V. Experimental Setup

To demonstrate the effectiveness of the proposed CBPT for Twitter depression detection, we conduct experiments on two publicly accessible datasets: the Tsinghua Twitter Depression Dataset (TTDD) and the CLPsych 2015 Twitter Dataset (CLPsych2015). All experimental procedures have been approved by the Ethical Review body of University of Leicester. In this section, we describe the setup details of our evaluation.

TTDD[1]: The Twitter database was collected by Shen et al. [7] in 2017 for depression detection. The Twitter database has three parts: (1) **Depression Dataset D1**: The dataset was created based on the tweets collected between 2009 and 2016, where the users were labelled as depression if their anchor tweet satisfied the pattern "(I'm/I was/I am/I've been) diagnosed depression". (2) **Depression Dataset D2**: This dataset contains Twitter messages where users were labelled as non-depressed if they had never posted any tweets containing the character string "depress". (3) **Depression Dataset D3**: Shen et al. [7] constructed an unlabelled large dataset D3 for depression candidate. Based on the tweets shown in December 2016, this unlabelled depression candidate dataset was established where the user were recorded if their anchor tweet loosely contained the character string "depress". There are 2558, 5304 and 58810 samples stored in D1, D2, D3, respectively. Each sample of these three datasets contains one-month post information of a Twitter user . In this paper, we employ the well labelled datasets D1 and D2 to evaluate our classification algorithm's performance and analyse the online behaviours of depression users.

CLPsych 2015[2]: The dataset was established by John Hopkins University for a depression detection task in 2015 [26]. The dataset contains public Twitter users' posts between 2008 and 2013 via the Twitter application programming interface (API). Similarly, possible mental disease sufferers are labeled as depression or post-traumatic stress disorder (PTSD) according to their self statement of diagnosis, such as "I was just diagnosed with depression or PTSD...". Furthermore, they conducted careful pre-preprocessing and anonymisation operations, such as filtering the users whose tweets are fewer than 25 and removing individual information. Finally, they manually examined and refined the annotation of each collected Twitter user's logs by using a semi-supervised method. The processed dataset consists of 477 depressed users, 396 PTSD (an anxiety disorder caused by very stressful, frightening or distressing events) users and 873 control users. For each user, up to their most recent 3000 public tweets were included in the dataset.

Implementation Details: We implement the proposed CBDT and other benchmark experiments using the Scikit-learn framework [46] and deploy all the experiments on a 8-core Intel Xeon skylake 2.6GHz CPU with 64GB RAM. The source code will be publicly accessible[3].

VI. Experimental Results

In this section, we present both quantitative and qualitative experimental results of different trials. We first conduct an ablation study of our method to show the impact of the pruning procedure and the cost-sensitive boosting scheme on the classification performance. We also compare our proposed Twitter depression detection framework with several state-of-the-art methods using the aforementioned two Twitter datasets. Finally, we justify the signification factors for depression prediction by our model.

A. Ablation Studies

In order to evaluate our proposed CBPT comprehensively, besides the two Twitter datasets, we also use three publicly accessible datasets (e.g. LSVT, Statlog, Glass) from the UCI machine learning repository [47] to examine our method's classification performance. We compare our method with Real Adaboost [48], XGboost [35], LogitBoost [49], LightBoost [50] and KiGB [44], which are state-of-the-art Boosting methods. We also investigate the performance of the

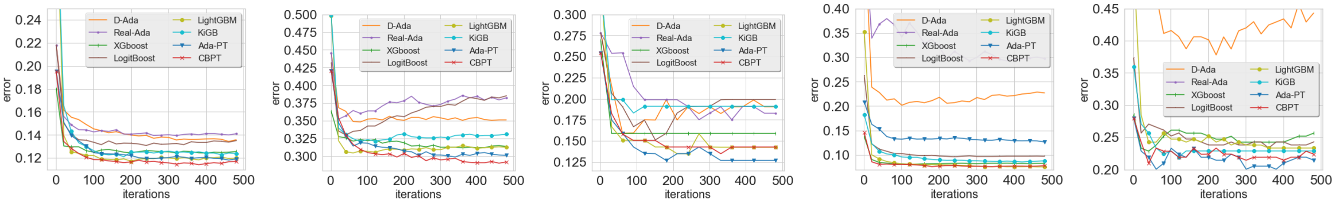
[1] http://depressiondetection.droppages.com/

[2] http://www.cs.jhu.edu/ mdredze/clpsych-2015-shared-task-evaluation/

[3] https://github.com/BIPL-UoL/Cost-Boosting-Pruning-Trees-for-depression-detection-on-Twitter

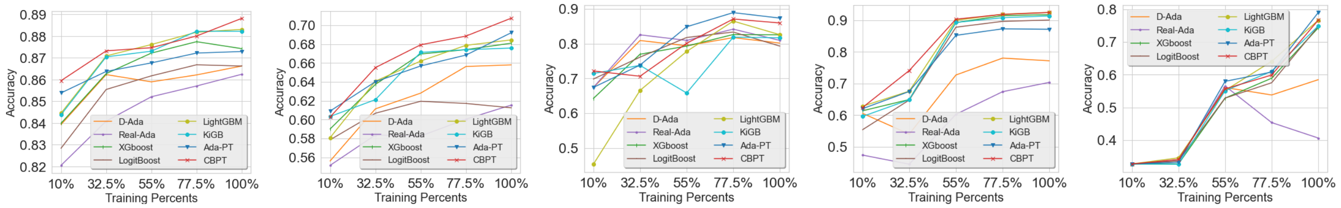
TABLE I: Classification Results: [Mean Accuracy/F1 Score±Standard Deviation] by eight boosting classifiers for five public datasets. The Best results are shown in bold.

	TTDD		CLPsych 2015		LSVT		Statlog		Glass	
Algorithm	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Discrete Adaboost	86.48±0.93	84.88±1.02	64.76±2.02	61.28±2.48	80.15±4.39	75.45±6.71	77.17±0.82	71.15±1.08	58.07±8.84	48.92±7.45
Real Adaboost	85.79±0.85	84.21±0.98	61.42±3.75	57.70±3.45	81.72±3.30	78.05±5.09	70.34±4.29	62.54±4.26	40.64±11.68	29.72±19.01
XGboost	87.43±0.56	86.00±0.57	68.62±2.62	64.66±3.25	84.12±2.54	79.66±5.76	91.74±0.79	90.13±0.85	74.36±10.83	69.56±11.55
LogitBoost	86.54±0.22	85.01±0.28	61.48±3.24	57.32±3.79	80.09±6.80	76.00±5.65	90.33±0.63	88.23±0.59	75.27±6.74	71.84±8.98
LightGBM	87.69±0.72	86.49±0.67	68.62±1.66	64.46±2.30	85.75±3.87	79.90±10.72	**92.46±0.62**	90.90±0.59	76.67±8.87	72.67±10.42
KiGB	87.73±0.68	86.29±0.68	67.06±2.05	62.79±2.27	81.69±5.53	77.76±4.83	91.40±0.70	89.71±0.59	77.13±8.94	67.87±12.84
Adaboost+PT (Ours)	87.70±0.77	86.34±0.83	69.71±2.74	65.71±3.34	**86.52±5.37**	**82.45±7.98**	87.13±1.05	85.04±1.08	**79.02±9.24**	**72.70±9.06**
CBPT (Ours)	**88.39±0.60**	**86.90±0.62**	**70.69±1.84**	**66.54±2.42**	85.72±4.03	81.26±6.24	92.21±0.31	**91.20±0.38**	77.63±8.58	70.66±9.55



(a) TTDD (b) CLPsych 2015 (c) LSVT (d) Statlog (e) Glass

Fig. 2: Convergence Rate: Testing error per iteration/tree.



(a) TTDD (b) CLPsych 2015 (c) LSVT (d) Statlog (e) Glass

Fig. 3: Learning curves for different training sets.

standard Discrete Adaboost and combine the Discrete Adaboost structure with the pruning procedure (Adaboost+PT) as a comparison method to validate the effectiveness of our newly added components. We summarise the datasets' details in Table S5, Supplementary B.

For the performance comparison, we use Accuracy and F1-score as evaluation metrics. The UCI datasets have supplied feature vectors and the ground truth, so we use the same feature extraction procedure (aforementioned in Section 3) to extract features vectors from the two Twitter datasets. We use 5-fold cross-evaluation on the five datasets, where the training size is 75% and the test size is 25%. To seek a fair comparison, we have evaluated different settings of the hyperparameters for the compared methods and the best results on the test set are recorded. Some key hyperparameters include: (1) $Num\ leaves \in \{64, 128, 256\}$, which control the size of each tree. (2) $Max\ depth \in \{5, 10, 15\}$, which limit the maximum depth of each tree. (3) $Learning\ rate \in \{0.1, 0.5, 1\}$, which determine the weight coefficient of each tree. (4) We fix the $tree\ number$ in all the classifiers to 500 in order to obtain converging results. More details of the parameter setting are listed in Table S6-10, Supplementary B.

The results of classification on the five datasets are presented in Table I. We observe that CBPT obtains the best performance in the two Twitter datasets and achieves 92.21% accuracy and a F1-score of 91.20% in the Statlog dataset. But in the LSVT and Glass datasets, the 'ablation' method Adaboost+PT results surpass CBPT by 1% and 2% separately. The reason is that the cost-sensitive boosting structure may be weak in the small-scale datasets. The Adaboost+PT outperforms the baseline Discrete Adaboost in the five datasets, confirming the effectiveness of our proposed pruning procedure. In general, the classification performance of CBPT for the five datasets is better than the other boosting methods except Adaboost+PT. To find out why this occurs, we undertake the following experiments.

Fig. 2(a)-(e) show the testing errors per iteration of the boosting classifiers for the five datasets. We observe that CBPT uses fewer trees to produce a comparable testing error in the TTDD, CLPsych 2015, and Statlog datasets. Comparing Adaboost+PT with CBPT, we witness the cost-sensitive boosting structure is effective to speed up the convergence of the algorithm in the TTDD, CLPsych 2015, and Statlog datasets. In the LSVT and Glass datasets, the cost-sensitive boosting structure is not helpful to improve the testing accuracy. As the LSVT and Glass datasets only have 128 and 214 samples respectively, we examine that in the cost-sensitive boosting structure, the newly added two penalty terms accelerate the weight updating and increase

TABLE II: Detection performance compared with the SOTA frameworks for the TTDD dataset. The best results are shown in bold.

	TTDD	
Method	Accuracy	F1-score
Shen et al. [7]	85%	85%
Pedregosa et al. [46]	73%	71%
Song et al. [51]	82%	81%
Rolet et al. [52]	76%	76%
CBPT (Ours)	**88.39**%	**86.90**%

TABLE III: Detection performance compared with the SOTA frameworks for the CLPsych 2015 dataset. The best results are shown in bold. Columns: depression vs. control (DvC), depression vs. PTSD (DvP) and PTSD vs. control (PvC).

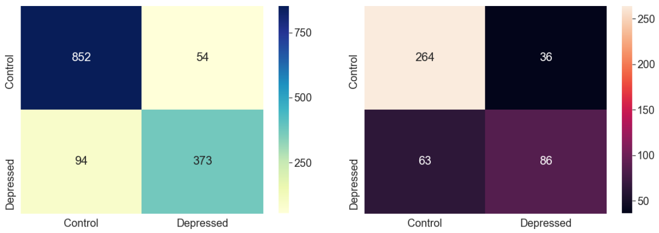
	CLPsych 2015		
Method/Problem AUC	DvC	DvP	PvC
Resnik et al. [53]	0.860	**0.841**	0.893
Preoţiuc-Pietro et al. [54]	**0.862**	0.839	0.860
Pedersen et al. [55]	0.730	0.780	0.710
Coppersmith et al. [26]	0.815	0.821	0.847
CBPT (Ours)	0.840	0.812	**0.898**

the variance in the small-scale datasets. To validate our assumption, we look at Fig. 3(a)-(e). The accuracy of CBPT and Adaboost+PT increase as more training samples are added. In spite of being trained with small data, CBPT and Adaboost+PT still outperform the baseline Discrete Adaboost, which verifies the pruning procedure effectively improves the models' generalization ability. From Fig. 3(a), (b) and (d), CBPT outperforms Adaboost+PT after having been trained with 32.5% or more training data. We summarise that in the case of sufficient training data, the proposed cost-sensitive boosting structure can improve the robustness of the model.

B. Comparison with the SOTA Depression Detection Frameworks

In the above discussion, we have verified our proposed classifier CBPT outperforms the other SOTA boosting algorithms in the two Twitter depression detection datasets. We employ the same feature extraction procedure to extract features from the two Twitter datasets. We obtain 38 dimensional feature vectors from the TTDD dataset and 40 dimensional vectors from the CLPsych 2015 dataset (i.e. age and gender information are available so we extract the extra two features from the CLP dataset). The two feature matrixes are used to train CBPT.

Tables II and III show the comparison results of depression detection. From Table II, it is obvious that our framework achieves the best performance and surpasses the SOTA method of Shen et al. [7] by 3.39% on accuracy and 1.69% on F1-score. In the CLPsych 2015 leader-board, the detection performance is evaluated against three separate classification tasks, i.e. Depression vs. Control, Depression vs. PTSD and PTSD vs. Control. In Table III, the CBPT results are competitive and better than the other methods in the PvC task. Another advantage of our framework is that the dimensionality of our extracted feature is far less than that of the other methods. For example, Resnik et al. [53] employed a complicated Supervised LDA model to extract document vectors and combine these with large vocabularies (feature dimensionality is about 500). Preoţiuc-Pietro et al. [54] applied the unigram word features of 41687 dimensions to training their model. Our method only uses few features and achieves competitive performance for the CLPsych 2015 dataset. From the two comparison experiments, we can verify our proposed depression detection framework has satisfactory robustness on different datasets.



(a) TTDD

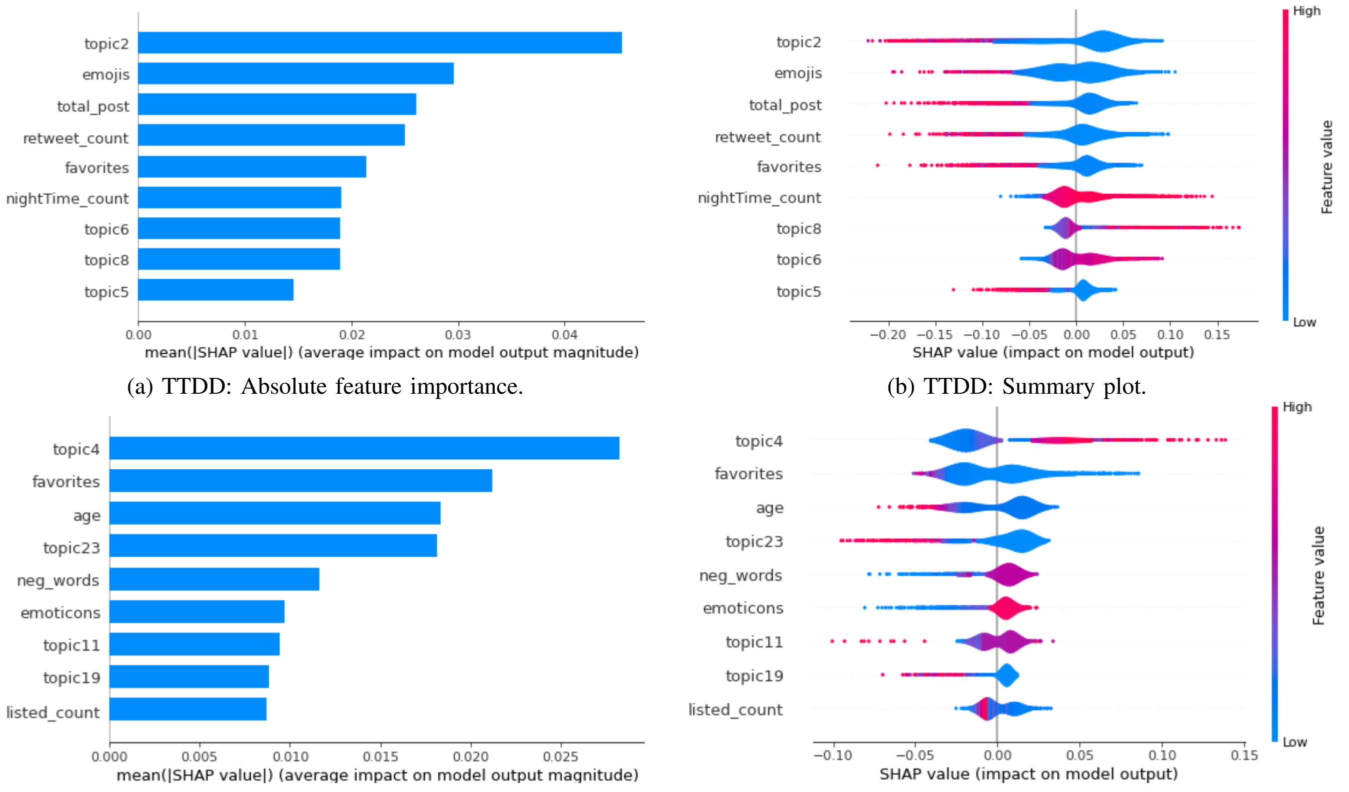
(b) CLPsych 2015

Fig. 4: Confusion Matrixes for the two depression detection datasets.

C. Explainable Depression Detection

Previous research studies [7], [21], [56] have widely analysed online behaviours of depressed users through examining features' distributions or mean values and variances. But they have not explored which specific factors contribute to depression detection. Tree Shapley Additive Explanation (TreeSHAP) [57] is a game approach to explain the output of decision trees based models. The goal of TreeSHAP is to explain the prediction of any instance by measuring the contribution of each feature to the prediction. TreeSHAP treats Shapley Values [58] as the features' contributions and uses all the advantages of Shapley Values: (1) TreeSHAP has a solid theoretical foundation in the game theory. (2) The prediction is fairly distributed over the features' values. (3) TreeSHAP gives contrastive explanations that compares the prediction with the model's expectation [28]. Hence, we integrate our framework with TreeSHAP to comprehensively investigate the influencing factors for the prediction results. We use the subset CvD of the CLPsych 2015 and the TTDD dataset for evaluating the depression risk factors, and other results (e.g. DvP, PvC subsets) are shown in Supplementary C. Besides, we list the related formulas of TreeSHAP in Section A, Supplementary D and we give an example for the calculation of the Shapley Values via decision trees in Section B, Supplementary D.

In Section A, Supplementary D, we describe that $\phi_{X_i^v}(f)$ represents the contribution of feature X_i^v to the classifier's prediction for instance X_i. In our depression detection datasets, we aim to explore the influencing factors for the predicted depression risk of Twitter users, so the value of $\phi_{X_i^v}(f)$ represents how much the predicted depression probability for instance X_i has been affected by feature X_i^v.



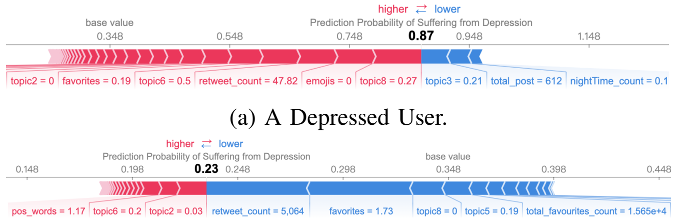
(a) TTDD: Absolute feature importance.

(b) TTDD: Summary plot.

(c) CLPsych 2015: Absolute feature importance.

(d) CLPsych 2015: Summary plot.

Fig. 5: Top 9 significant features for depression detection. (a) and (c): Average feature importance. (b) and (d): Summary Plots. Each point is a Shapley Value $\phi_{X_i^v}(f)$ corresponding to a feature and an instance. Overlapping points are jittered on the y-axis direction.

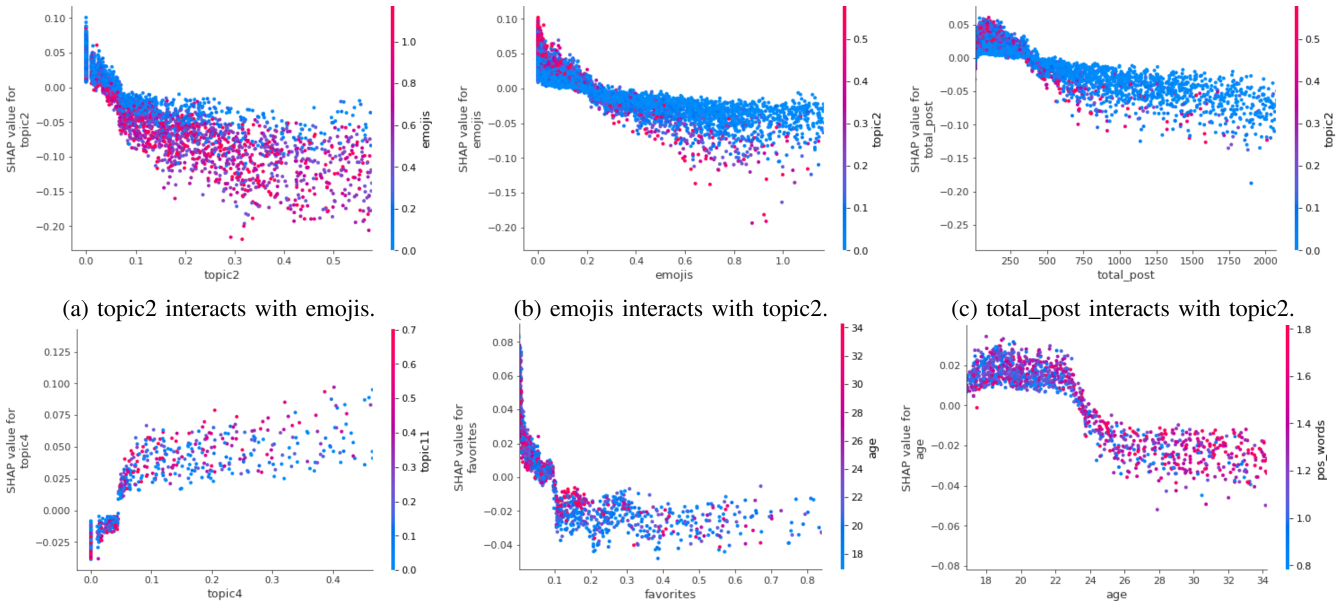


(a) A Depressed User.

(b) A Control User.

Fig. 6: Additive force plots for two Twitter users. The bold text is the predicted depression probability. Each arrow (red or blue) is a single feature of the instance and the arrow length represents the feature's Shapley Value.

Fig. 4 shows two confusion matrices of the prediction results of CBPT over the two depression detection datasets. From these figures, we know how many depressed or control users has been classified. Here, we use feature importance to analyse which feature significantly affects global depression detection. Feature importance is computed by $I^v = \frac{\sum_{i=1}^{N} \left|\phi_{X_i^v}(f)\right|}{N}$ (N is the number of data instances). Fig. 5(a) and (c) show top 9 significant features for depression detection in the two Twitter datasets. In these two figures, features with large absolute Shapley Values are important. For example, topic2 stands in the most critical position in Fig. 5(a) and topic2 changes the predicted depression probability by 4% on average for all the instances. Although the feature importance plot is useful, there is no more information beyond the importance. For more information, we use summary plots (Fig. 5(b) and (d)) to further analyse the significant features. In the summary plots, each point is a Shapley Value $\phi_{X_i^v}(f)$ corresponding to a feature and an instance. Overlapping points are jittered on the y-axis direction so each row is the distribution of Shapley Values. In Fig. 5(b), topic2 with a high feature value (red points) stands for decreasing depression risk and a low value of topic2 (blue points) refers to increasing depression risk. In Table S13-S14, Supplementary C, we show the top 10 words that are the most likely to occur in each LDA topic. Topic2 includes words such as 'trump', 'obama', 'russia' that infers topic2 may be related to 'politics'. The feature value of topic2 is the occurrence probability of topic2 in the posting texts. If a user posts many tweets on the theme of politics, his/her predicted depression risk will be decreased. Similarly, if a user posts many tweets with emojis that receive many retweets, the user is less likely to be depressed. Depressed users seem to lack of communication with others that depressed users are more likely to post tweets during midnight and their posted tweets are barely retweeted or favoured by other users. And it is an



(a) topic2 interacts with emojis.

(b) emojis interacts with topic2.

(c) total_post interacts with topic2.

(d) topic4 interacts with topic11.

(e) favourites interacts with age.

(f) age interacts with pos_words.

Fig. 7: Feature dependence Plot: The point's position on the X axis represents the target feature value, and the values on the Y axis are the Shapley Values for the target feature. Color bar is the value range of the interactive feature. (a)-(c) are the dependence plots of the TTDD dataset. (d)-(f) are the dependence plots of the CLPsych 2015 dataset.Please note that the x-axis and color bar range have been trimmed to the 5th and 95th percentile of the data in order to avoid the x-axis or color bar being too board because of the outliers.

interesting phenomenon that the posting texts of depressed users may involve the content of ’film’ (topic8) or ’policy’ (topic6) but without ’band’ (topic5) information. In Fig. 5(d), the most important feature topic4 is related to the theme of ’mental health’ (shown in Table S14, Supplementary C). In the CLPsych 2015 dataset, depressed users are more likely to undertake the following behaviours: (1) Their posted tweets are related to the topics of ’mental health’ (topic5) or ’news’ (topic11) and include many emoticons and negative words. (2) They are young and they do not take many Twitter activities. (3) Their posted tweets may not be favoured by others and their tweets’ content is not related to ’friend’ (topic19) and ’autism’(topic23).

Then, we use the additive force plots to explain why a user is predicted as depressed or control. Using two instances from the TTDD dataset, Fig. 6(a) is the prediction visualization of a depressed user. In Fig. 6(a), the bold text 87% is the predicted depression probability and the base value 34.8% is the classifier's expectation $\phi_0(f)$ referring to Eq. (2), Supplementary D. Features pushing the prediction higher are shown in red, while those pushing the prediction lower are shown in blue. For example, the emoji number of this user is 0 which is lower than the average value 0.34 (shown in Table S2, Supplementary A) and it contributes 8% probability to the depression prediction. The total_post feature (= 612 that is larger than 457.31)[4] reduces the predicted risk about 4%. This supports our finding in Fig. 5(a)-(b) that few emojis lead to higher predicted depression risk and posting many tweets leads to less risk. Similarly, for a control user shown in Fig. 6(b), this user's posting content may not be relevant to ’politics’ (topic2=0.03 that is less than 0.11) that increases the depression risk by 1%. This users’ tweets are frequently retweeted by others (retweet_count=5064 that is over 1843.14) and this behaviour decreases the user's predicted depression risk. The predicted depression probability for the control user drops from the base probability 34.8% to 23%.

Finally, we use the dependence plots to show the detailed interpretation of the features’ impacts. Fig. 7 includes 6 dependence plots for the most important three features with their most interactive features in the two datasets. The interactive feature can be selected arbitrarily and we decide the most interactive features depending on Eq. (6), Supplementary D. This equation calculates the correlation coefficient between the Shapley Values of the target feature and the values of the other features. In Fig. 7(a), the predicted depression risk decreases with the increasing of the values of topic2 and emojis. This suggests that posting tweets on the theme of politics with many emojis leads to lower predicted depression risks and vice versa. In Fig. 7(b), topic2 is also the most interactive feature of emojis. This figure shows a similar trend to Fig. 7(a). In Fig. 7(c), the predicted depression risk shows a decreasing value at total_post=400. This suggests that control users are more likely to post many tweets and share politics news than the depressed users. Similarly, from Fig. 7(d)-(f), we observe

[4] 612 is the feature value of total_post and 457.31 is the average value of this feature (shown in Table S2, Supplementary A).

that posting tweets about 'mental health' (topic4) and 'news' (topic11) is proportionally related to the predicted depression risk. Depressed users' tweets are hard to receive favourites from others. The predicted depression risk for the Twitter users is decreasing at age=23 and using the positive or negative words will change the depression risk of the users. By the above feature dependence analysis, we have shown the influences of the feature interactions on the classifier's predicted depression probability and revealed the difference of the online behaviours between the depressed and control users.

VII. Conclusion

In this paper, we have made an attempt to automatically identify potential Twitter depressed users. As we have known, most of the established works mainly focused on exploring new features of depression behaviours whilst ignoring the fitness of the classification models. Considering the complexity of Twitter data, in order to improve the robustness of the decision tree based estimator, we proposed a novel resampling weighted pruning algorithm which dynamically determines optimal depths/layers and leaves of a tree model. Taking into account the "hardness" of different misclassified samples, we also proposed a cost-sensitive boosting structure to hierarchically update the instances' weights in the pruned trees. We combined the proposed pruning process with the novel cost-sensitive boosting structure within an ensemble framework, namely Cost-sensitive Boosting Pruning Trees (CBPT) to classify control and depressed users.

CBPT outperformed the other depression detection frameworks in the two Twitter datasets. In the meantime, we conducted the convergence analysis of our proposed CBPT through comprehensive experiments. Moreover, we utilised three UCI datasets to evaluate the classification ability of our method quantitatively, which shows our method performs better than the other SOTA boosting algorithm. We then integrated CBPT with TreeSHAP in order to explain the predicted depression risks of Twitter users by investigating the contribution of each feature to the prediction. We used three different types of figures, i.e. additive force plot, summary and dependence plots, to explain the contributions of individual features to the predicted depression risks.

Taking a close look at the above experimental results, we found that the features extracted from the tweet content were really important for depression prediction. Features including LDA topics, negative/positive words and emojis play a key role in online depression risk detection. In the future, we will develop a robust topic model methodology to summarise posting text content of depressed users with clearly explainable topics. We will also attempt to mine similar information over other social networks, e.g. Facebook, Instagram, and Tumblr, for sentiment analysis.

Acknowledgements

The work of Huiyu Zhou is supported in part by the Royal Society-Newton Advanced Fellowship under Grant NA160342.

References

[1] R. Cheng, S. Juo, J. Loth, J. Nee, I. Iossifov, R. Blumenthal, L. Sharpe, K. Kanyas, B. Lerer, B. Lilliston *et al.*, "Genome-wide linkage scan in a large bipolar disorder sample from the national institute of mental health genetics initiative suggests putative loci for bipolar disorder, psychosis, suicide, and panic disorder," *Molecular psychiatry*, vol. 11, no. 3, p. 252, 2006.

[2] S. McManus, H. Meltzer, T. Brugha, P. Bebbington, and R. Jenkins, *Adult psychiatric morbidity in England: Results of a household survey*. Health and Social Care Information Centre, 2009.

[3] L. Andrade, J. J. Caraveo-Anduaga, P. Berglund, R. V. Bijl, R. D. Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R. C. Kessler *et al.*, "The epidemiology of major depressive episodes: results from the international consortium of psychiatric epidemiology (icpe) surveys," *International journal of methods in psychiatric research*, vol. 12, no. 1, pp. 3–21, 2003.

[4] L. S. Radloff, "The ces-d scale: A self-report depression scale for research in the general population," *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.

[5] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013.

[6] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, 2016.

[7] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pp. 3838–3844, 2017.

[8] R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pp. 581–586. Association for Computational Linguistics, 2011.

[9] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media." *ICWSM*, vol. 13, pp. 1–10, 2013.

[10] Y. Neuman, Y. Cohen, D. Assaf, and G. Kedma, "Proactive screening for depression through metaphorical and automatic text analysis," *Artificial intelligence in medicine*, vol. 56, no. 1, pp. 19–25, 2012.

[11] M. Zaydman, "Tweeting about mental health," Ph.D. dissertation, PARDEE RAND GRADUATE SCHOOL, 2017.

[12] M. Akbari, X. Hu, L. Nie, and T.-S. Chua, "From tweets to wellness: Wellness event detection from twitter streams." in *AAAI*, pp. 87–93, 2016.

[13] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 51–60, 2014.

[14] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–1, 2017.

[15] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, no. 3, pp. 217–226, 2014.

[16] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *Icml*, vol. 96, pp. 148–156. Citeseer, 1996.

[17] Y. Lee, M.-J. Yang, T.-J. Lai, N.-M. Chiu, and T. Chau, "Development of the taiwanese depression questionnaire," *Chang Gung medical journal*, vol. 23, pp. 688–694, 2000.

[18] M. Park, D. W. McDonald, and M. Cha, "Perception differences between the depressed and non-depressed users in twitter." *ICWSM*, vol. 9, pp. 217–226, 2013.

[19] M. Park, C. Cha, and M. Cha, "Depressive moods of users portrayed in twitter," in *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, vol. 2012, pp. 1–8. ACM New York, NY, 2012.

[20] A. Hussain, J. Heidemann, and C. Papadopoulos, "A framework for classifying denial of service attacks," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pp. 99–110. ACM, 2003.

[21] M. Nadeem, "Identifying depression on twitter," *arXiv preprint arXiv:1607.07384*, 2016.

[22] H.-H. Shuai, C.-Y. Shen, D.-N. Yang, Y.-F. C. Lan, W.-C. Lee, S. Y. Philip, and M.-S. Chen, "A comprehensive study on social network mental disorders detection via online social media mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1212–1225, 2018.

[23] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T. S. Chua, and W. Hall, "Cross-domain depression detection via harvesting social media." International Joint Conferences on Artificial Intelligence, 2018.

[24] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 81–88, 2019.

[25] P. K. Gamaarachchige and D. Inkpen, "Multi-task, multi-channel, multi-input learning for mental illness detection using social media text," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pp. 54–64, 2019.

[26] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, DOI 10.3115/v1/W15-1204, pp. 31–39. Denver, Colorado: Association for Computational Linguistics, Jun. 5 2015. [Online]. Available: https://www.aclweb.org/anthology/W15-1204

[27] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "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, 2018.

[28] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[29] D. Ignatov and A. Ignatov, "Decision stream: Cultivating deep decision trees," in *Tools with Artificial Intelligence (ICTAI), 2017 IEEE 29th International Conference on*, pp. 905–912. IEEE, 2017.

[30] G. Leshem, "Improvement of adaboost algorithm by using random forests as weak learner and using this algorithm as statistics machine learning for traffic flow prediction. research proposal for a ph. d," *Research proposal for a Ph. D. thesis, the Hebrew university of Jerusalem*, 2005.

[31] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. With Applications in R. New York: Springer, 2013.

[32] L. Breiman, "Bias, variance, and arcing classifiers (technical report 460)," *Statistics Department, University of California*, 1996.

[33] N. Boonyanunta and P. Zeephongsekul, "Improving the predictive power of adaboost: A case study in classifying borrowers," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 674–685. Springer, 2003.

[34] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[35] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

[36] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 1191–1198. ACM, 2017.

[37] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey *et al.*, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.

[38] M. Porter and R. Boulton, "Snowball stemmer," 2001.

[39] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 47–56. ACM, 2013.

[40] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 31. Association for Computational Linguistics, 2004.

[41] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, "Beyond lda: exploring supervised topic modeling for depression-related language in twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 99–107, 2015.

[42] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[43] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[44] H. Kokel, P. Odom, S. Yang, and S. Natarajan, "A unified framework for knowledge intensive gradient boosting: Leveraging human experts for noisy sparse domains." in *AAAI*, pp. 4460–4468, 2020.

[45] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[47] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[48] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[49] P. Li, "Robust logitboost and adaptive base class (abc) logitboost," *arXiv preprint arXiv:1203.3491*, 2012.

[50] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, pp. 3146–3154, 2017.

[51] X. Song, L. Nie, L. Zhang, M. Akbari, and T.-S. Chua, "Multiple social network learning and its application in volunteerism tendency prediction," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 213–222, 2015.

[52] A. Rolet, M. Cuturi, and G. Peyré, "Fast dictionary learning with a smoothed wasserstein loss," in *Artificial Intelligence and Statistics*, pp. 630–638, 2016.

[53] P. Resnik, W. Armstrong, L. Claudino, and T. Nguyen, "The university of maryland clpsych 2015 shared task system," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 54–60, 2015.

[54] D. Preoţiuc-Pietro, M. Sap, H. A. Schwartz, and L. Ungar, "Mental illness detection at the world well-being project for the clpsych 2015 shared task," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 40–45, 2015.

[55] T. Pedersen, "Screening twitter users for depression and ptsd with lexical decision lists," in *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pp. 46–53, 2015.

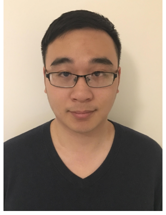
[56] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pp. 32–40, 2017.

[57] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[58] L. S. Shapley, *Notes on the N-person Game–II: The Value of an N-person Game*. Rand Corporation, 1951.

Lei Tong is currently pursuing the Ph.D. degree with the School of Informatics, University of Leicester, Leicester, U.K. His research interests include computer vision, social network analysis and data mining.

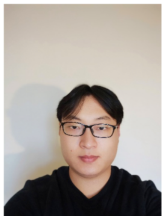
Zhihua Liu is currently pursuing the Ph.D. degree with the School of Informatics, University of Leicester, Leicester, U.K. His research interests include machine learning, deep learning and computer vision.



Zheheng Jiang has been awarded his Ph.D. degree in Computer Science from University of Leicester, Leicester, U.K. He is currently the Senior Research Associate at the Computing and Communications, Lancaster University, Lancaster, U.K. His current research interests include machine learning for vision, object detection and recognition, video analysis and event recognition.

Feixiang Zhou is currently pursuing the Ph.D. degree with the School of Informatics, University of Leicester, Leicester, U.K. His current research interests include Computer Vision, Machine Learning and their applications on video understanding.

Long Chen is currently pursuing the PhD degree with the School of Informatics, University of Leicester, U.K. His research interests are in the areas of Computer Vision and Machine Learning.

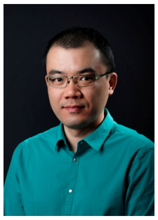


Jialin Lyu is an MPhil student at the School of Informatics, University of Leicester. His current project is Automated Classification of Alzheimer's Disease and Mild Cognitive Impairment Using Deep Neural Networks. His research interests include machine learning and medical image analysis.

Xiangrong Zhang received the B.S. and M.S. degrees from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, in 2006. Currently, she is a professor in the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, China. She has been a visiting scientist in Computer Science and Artificial Intelligence Laboratory, MIT between Jan. 2015 and March 2016. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.

Qianni Zhang received the Ph.D. degree from Queen Mary University of London, U.K., in 2007. She is currently a Senior Lecturer (Associate Professor) at the School of Electronic Engineering and Computer Science, Queen Mary University of London. She has authored over 50 technical papers and book chapters, and has actively contributed to several European funded research projects. Her research interests include multimedia processing, semantic inference and reasoning, machine learning, image understanding, 3D reconstruction, and immersive environments. She has served as a Guest Editor of a special issue in Journal of Multimedia and a Reviewer of journals including the IEEE Transactions on CSVT, Image Processing, Multimedia, Sensors, Signal Processing: Image Communication, and various conferences and workshops including the IEEE ICIP, ICASSP, ICME and ACM Multimedia. She has served as an organiser, a session chair, or a member of the technical program committee of several international conferences, workshops, or special sessions.

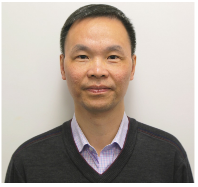
Abdul Sadka is the former Head of the Department of Electronic and Computer Engineering (ECE) at Brunel University London (Oct 06 - Jul 12). He has a leadership-focused and research-inspired academic career underlined by a commitment to industry engagement and high-profile academic output. He has 200+ publications in refereed journals and conferences, 3 patents and a seminal textbook entitled ”Compressed Video Communications” published by J. Wiley in 2002. He has thus far managed to attract circa £13m worth of research grants and contracts and supervised nearly 50 Research Assistants and PhD students to full completion. He frequently serves on influential advisory boards and international evaluation panels and provides expert consultancy services to the Telecom/ICT industry as well as corporate Law firms in the area of 2D/3D video compression and multimedia processing. He served as elected member (2011-14) on the Steering Board of the NEM European technology platform which heavily influences the strategic research agenda of the European Commission Framework Programmes. He is a Chartered Engineer (CEng), Fellow of HEA, IET and BCS.



Yinhai Wang received the B.Eng. degree in computer software from Jinan University, Guangzhou, China, in 2001, the M.Sc. degree in software engineering from Napier University, Edinburgh, U.K., in 2002, and the Ph.D. degree in electronics, electrical engineering, and computer science from Queen's University, Belfast, U.K., in 2008. He is currently an Associate Director at Data Science & Quantitative Biology, AstraZeneca (Cambridge, UK) focusing on the use of image analysis and data sciences in biological applications.

Ling Li is the Director of Internationalisation at the School of Computing and also the founding coordinator of Laboratory of Brain | Cognition | Computing (BC2 Lab) of the school responsible for coordinating multidisciplinary research between Computing, Sports and local NHS hospitals. She had six-year research experience at Imperial College London with a focus to understand body sensor data (EEG, EMG, ECG, eAR-sensor, and etc.). She participated in large scale projects. She also involved in projects from government and industry (i.e. Samsung GRO award). She now serves at the editorial board of Brain Informatics and the secretary of IEEE Computing Society in UK and Ireland.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3145634, IEEE Transactions on Affective Computing



Huiyu Zhou received a Bachelor of Engineering degree in Radio Technology from Huazhong University of Science and Technology of China, and a Master of Science degree in Biomedical Engineering from University of Dundee of United Kingdom, respectively. He was awarded a Doctor of Philosophy degree in Computer Vision from Heriot-Watt University, Edinburgh, United Kingdom. Dr. Zhou currently is a full Professor at School of Informatics, University of Leicester, United Kingdom. He has published over 350 peer-reviewed papers in the field. His research work has been or is being supported by UK EPSRC, ESRC, AHRC, MRC, EU, Royal Society, Leverhulme Trust, Puffin Trust, Invest NI and industry.