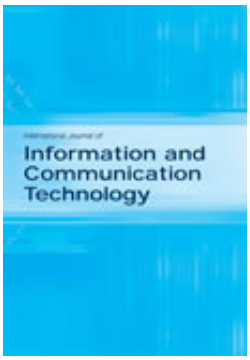
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Social psycho-emotional characterisation of college students based on semi-supervised learning

Weihua Li

Article History:

Received:	27 October 2024
Last revised:	22 November 2024
Accepted:	23 November 2024
Published online:	20 January 2025

Social psycho-emotional characterisation of college students based on semi-supervised learning

Weihua Li

School of Culture and Arts,
Xinjiang Institute of Engineering,
Urumqi, 830023, China
Email: 13899980090@163.com

Abstract: Texts generated by college students through social sharing are characterised by emotional richness and psychological vulnerability. To address the issue that existing social-psycho-emotional profiling methods for college students rely on the size of the labelled dataset and have unsatisfactory classification results, this article first optimises semi-supervised learning (MSASL), which composes the samples and incorporates a smoothness loss while imposing consistency constraints. The BERT model is then used to obtain a semantic representation of the social text, using interactive attention to capture important feature information related to the opinion tendency of the topic words. Finally, semi-supervised GAN (MSASL-GAN) is applied to optimise the text feature representation, and the sentiment feature classification results are output through the fully connected layer. The experimental results show that the classification accuracy of the proposed method is improved by 5.01%–11.51% compared with the comparison model.

Keywords: sentiment profiling; semi-supervised learning; SSL; BERT model; interactive attention; generative adversarial network; GAN.

Reference to this paper should be made as follows: Li, W. (2025) 'Social psycho-emotional characterisation of college students based on semi-supervised learning', *Int. J. Information and Communication Technology*, Vol. 26, No. 1, pp.117–130.

Biographical notes: Weihua Li received Bachelor's degree from the Lanzhou Jiaotong University in 2007. She is currently a Lecturer at the Xinjiang Engineering College. Her research interests include interpersonal processes, as well as the integration of psychology and culture.

1 Introduction

At present, the mental health of college students in our country is increasingly becoming the focus of social attention, contemporary college students are in a psychological quality of rapid growth of the adolescent stage, their minds open and active, and at the same time, emotionally rich and strong desire to express (Galay and Aizman, 2023). However, psychological immaturity usually makes them susceptible to the influence of external factors and lead to emotional disorders, such as love, exams and employment pressure will cause a certain degree of negative impact on college students, if they cannot make timely self-adjustment will often form depression and anxiety and other mental health

problems (Moeller et al., 2020). With the increasing popularity of the Internet, the number of college students on online social platforms is gradually increasing, and the number of texts produced is also countless, and people are more likely to express their emotions on online social platforms such as microblogging, WeChat circle of friends, or QQ space, and how to analyse the emotional characteristics of the social text data of the college students is a work of practical significance.

Traditional psycho-text sentiment profiling is based on the method of sentiment lexicon and rule base, Zhou et al. (2018) used positive, neutral and negative sentiments towards things in the text as a classification criterion and created a text retrieval tool for psycho-emotional analysis. The mutual information approach proposed by Shamoi et al. (2022) extracts emotion words from text and then determines the similarity between words based on lexical rules. However, the construction of sentiment lexicon costs a lot of human resources, whereas machine learning based methods utilise statistical knowledge without constructing sentiment lexicon, saving human resources. Kumar and Garg (2019) used semantic information to extract psycho-emotional features and combined it with speech information to propose a multi-category feature extraction method to effectively reduce the interference of noise in the context. Isnain et al. (2021) used K-nearest neighbours and simple Bayes to classify the sentiment of a collection of college students' microblog comments, but the classification accuracy of the method is not high.

Labelled data is difficult to obtain in ML-based methods for psychosentiment analysis, while the generalisation ability of traditional models gradually decreases as the size of training data becomes larger. Liu (2020) used CNN to initialise word vectors and to design and fine-tune a two-channel CNN model. Although CNN has achieved good results in text sentiment analysis, the model ignores the information of word order in text. Durga and Godavarthi (2023) proposed the use of recurrent neural network (RNN) to capture word order information in social text. However, the gradient of this model becomes smaller and smaller during the derivation process, and the model parameters cannot be updated. Kokab et al. (2022) proposed psycho-emotional feature classification based on the transformer model without the use of RNN and CNN, which uses only the attention mechanism.

The application of DL in various industries is gradually expanding, but the problem of its need for a large amount of manually labelled data is gradually coming to the fore, and if the dataset if too small in size, it will produce problems such as overfitting. The characteristics of semi-supervised learning (SSL) using a small number of less labelled samples and a large number of unlabeled datasets have attracted the attention of many researchers. Barreto et al. (2023) proposed the transductive support vector machine (TSVM), as well as Chen et al. (2019) proposed the co-training algorithm, after which SSL became associated with DL. Ruz et al. (2022) used a Bayesian network model to select the labels with the highest and lowest confidence in each iteration of a self-training algorithm as the classification result for textual sentiment features, but the classification efficiency was not satisfactory. Menaouer et al. (2022) applied spectral clustering and direct push SSL methods together to build a document sentiment analysis model, but there were prediction errors. Compared to traditional methods, generative adversarial networks (GANs) greatly improve the efficiency of characterisation due to their powerful generative ability to generate signatures similar to real data from random noise, while the discriminator tries to differentiate between the generated data and real data. Riyadh and

Shafiq (2022) proposed GAN-CBOW based on the idea of SSL, which utilises the advantages of CBOW and semi-supervised GAN.

Based on the analysis of existing studies, it is known that the current research work suffers from the high cost of data labelling and ignores the semantic mining of college students' social texts, which makes the effect of sentiment feature classification general. For this reason, this paper suggests a SSL-based approach for social psychological sentiment feature analysis of college students. Firstly, to address the issue that the SSL method ignores the data structure information, the SSL is optimised by fusing consistency regularity with streaming regularity (MSASL), which imposes consistency constraints while framing the samples and adding a smoothness loss to optimise the model's smoothness within the neighbourhood of each sample point as well as between sample points. Then the BERT model is introduced for social text representation, and important feature information is extracted by adding self-attentive head with syntax awareness to the BERT model, and then using interactive attention. Finally, by adding noise vectors to the text representation vectors and then generating new samples, semi-supervised GAN (MSASL-GAN) is applied to optimise the text feature representations, and the information from the features is used as the input to the sentiment classifier, and the results of the sentiment feature classification are obtained through the fully connected layer. The experimental outcome indicates that the suggested method has a significant improvement in classification accuracy compared to the other three approaches.

2 Relevant theoretical foundations

2.1 Semi-supervised learning

In the field of psycho-emotional characterisation of college students, the training of the model often requires a large amount of high-quality labelled data, and the acquisition of labelled data has a high cost of manpower and time, and in today's social platforms are flooded with a huge amount of social text information of college students, and the SSL can achieve good classification results in the face of a small amount of labelled data. An example of SSL is shown in Figure 1.

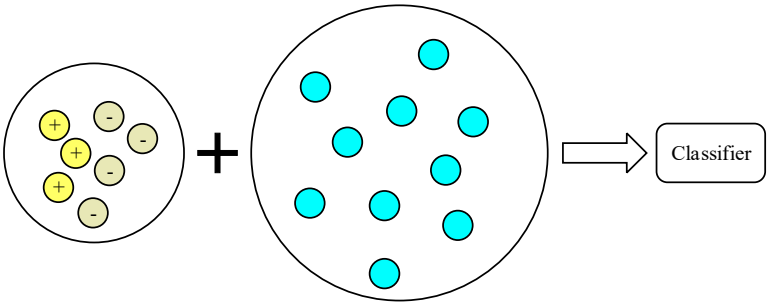
Formally the SSL classification algorithm can be expressed as follows: for a given input dataset $X = \{X_L \cup X_U\}$ with n samples, the input dataset consists of two parts, where a part of the dataset containing labels is called the labelled dataset $X_L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$, and the number of X_L is l, where the unlabeled dataset is $X_U = \{x_{i+1}, x_{i+2}, \ldots, x_{i+j}\}$, and the number of X_U is j, and at the same time the value of K, i has to be much less than j.

SSL relies on the model's assumptions on the data structure, which are the smoothing assumption, the clustering assumption, and the streaming assumption (Matuszyk and Spiliopoulou, 2017), among which the streaming assumption is able to effectively utilise the geometrical structural information in the high-dimensional data, which can improve the model's performance and generalisation ability. Conventional ML algorithms are not able to label the nodes that do not appear in the graph, along with the mislabelled and noisy points, which can be solved by stream regularisation. The objective function of stream regularisation is as follows.

$$f^* = \arg\min \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{1}$$

In addition to the loss function in the formula, the function contains two regularisers at the same time, a practice that allows for smoother variations in samples outside of the training set. Most of the existing SSL algorithms are based on one or more of the above data structure assumptions. In this paper, we will optimise the SSL algorithm based on the smoothing assumption and the streaming assumption.

Figure 1 SSL example diagram (see online version for colours)



2.2 Generating adversarial network

GAN was initially used in the image domain, but in the past few years, more and more scholars have focused on applying GAN to natural language processing (Vlachostergiou et al., 2018). Compared to neural network algorithms such as CNNs, GANs are capable of generating very realistic samples that do not rely on dataset size or a priori assumptions, and are able to learn distributions directly from the data, which greatly improves the performance of heart-sentiment profiling. GAN is constructed by using the idea of game, first of all, a generative model (G) and a discriminative model (D) are constructed, and the two models are in opposition to each other, and the ultimate purpose of the model is to make the samples generated by G 'deceive' D, so that D cannot judge the authenticity of the samples, as shown in Figure 2.

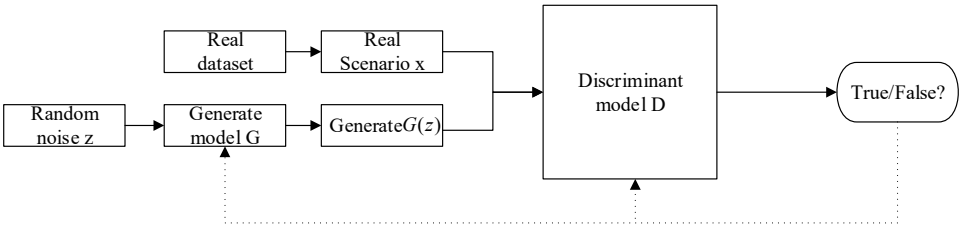
Take the input data as text data as an example, G is a network that generates text data, it can receive a random noise z, G uses the received z to generate a text sample $G(z)$, D can determine whether a text is real, the real input x is the data sample, the output $D(x)$ is the probability that the input D sample is real text data, if $D(x) = 1$ then the data is real, if $D(x) = 0$ then the input data is not real must not be real.

$$\min_G \max_D V(D, G) = E_{x-Pdata(x)}\left[\log D(x)\right] + E_{x-PZ(Z)}\left[\log\left(1 - D\left(G(Z)\right)\right)\right] \tag{2}$$

Throughout the training process, the GAN performs small batch stochastic gradient descent training. For the number of training iterations and k, D is first trained with a small batch of m noisy samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from the noise $p_g(z)$ a priori. m samples from the small batch of samples $\{x^{(1)}, \ldots, x^{(m)}\}$ form the data generating distribution $p_{data}(x)$, and since D wants $V(D, G)$ to be as large as possible, D is updated by boosting its stochastic gradient.

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \right] \tag{3}$$

Figure 2 Model structure of GAN

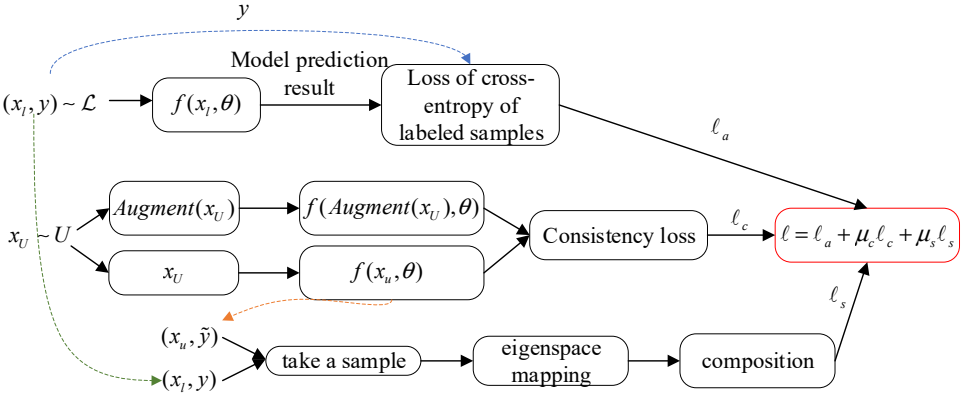


3 Optimisation of SSL algorithms based on consistency regularity and flow regularity

Most of the existing SSL algorithms are based on the smoothing assumption, but ignore the connectivity between data points and do not fully utilise the information in the data structure of the training set, such as clusters or manifolds, resulting in smoothing in the vicinity of each sample point, but not between points.

For the goal of dealing with the above issues, a new deep SSL algorithm (MSASL) combining the smooth hypothesis and manifold hypothesis in SSL is proposed in this paper. Taking into account the connection between sample points in the data, partial configuration information of the data is utilised to further strengthen the robustness of the model among different classes. The overall flow of the algorithm is shown in Figure 3.

Figure 3 The overall flow of MSASL (see online version for colours)



For a total of N samples in the dataset D, where L is labelled samples and the rest U are unlabeled samples, then the labelled dataset $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{L}$, labelled $y_i \in \gamma = \{1, 2, \ldots, K\}$, a total of K categories, the unlabeled dataset $U = \{x_i\}_{i=L+1}^{N}$, and $f(x_i; \theta)$ denote the predicted results of the classifier with parameter θ for the samples as a predefined loss of

cross-entropy. The training process represents a batch and the adjacency matrix represents the similarity measure between the samples.

The SSL algorithm (SASL) based on the smoothing assumption only imposes smoothing constraints on the perturbations in the local region of the samples, and ignores the information of the cluster structure among the sample points. Therefore, the MSASL algorithm is proposed by integrating the SASL algorithm with the flow regularisation, and the overall loss function of this algorithm is as follows.

$$\ell = \ell_a + \mu_c \ell_c + \mu_s \ell_s \tag{4}$$

where ℓ_a is the cross-entropy loss, ℓ_c is the consistency loss, and ℓ_s is the smoothness loss. μ_c and μ_s are the weights of balancing the three losses. Next, the specific calculation of each loss is introduced.

1 Consistency loss: In the SASL algorithm, in addition to performing the basic data augmentation introduced into the interpolation consistency training ICT in the data augmentation method mix-up, and the model's prediction of the interpolation x' between the sample points x_1 and x_2 should be consistent with the model's interpolation y' of its prediction.

$$\begin{cases} x' = \lambda x_1 + (1-\lambda) x_2 \\ y' = \lambda y_1 + (1-\lambda) y_2 \end{cases} \tag{5}$$

where λ is a weight parameter that obeys the beta distribution. If x_1 and x_2 belong to different clusters and have different labels, they are interpolated closer to the categorisation plane, and ℓ_c is calculated as follows.

$$\ell_c(\theta, \gamma, U) = \frac{1}{|B|} \sum_{i \in B} d\left(f\left(x_i'; \theta\right), y'\right) \tag{6}$$

where B is the set of sample points and d(., .) measures the difference between the classifier's prediction of the interpolated samples' labels and the labelled interpolations.

2 Smoothness loss: SASL is a distance measurement of the input space, but the data structure information of the sample will be lost, making the smoothness loss term meaningless. Therefore, this paper, referring to the SNTG model of Zhu et al. (Li et al., 2020), will solve the above problems from the calculation of composition and adjacency matrix and the calculation of slip loss function.

The similarity between the samples is first measured using the label space of the samples, and the weight matrix is calculated as follows.

$$W_{ij} = \begin{cases} 1 & \text{if } \hat{y}_i = \hat{y}_j \\ 0 & \text{if } \hat{y}_i \neq \hat{y}_j \end{cases} \tag{7}$$

where $\hat{y}_i = \arg\max_k [f(x_i)]_k$, $f(x_i)$ are pseudo-labels, and $\hat{y}_i$ takes the category with the highest probability among the soft labels as the unique hot label for that sample. W_{ij} is the similarity measure of x_i and x_j. If $W_{ij} = 1$, then x_i and x_j are close neighbours and vice versa.

Then the characteristic representation of the sample is considered, and assuming that $h(x)$ is a low-dimensional representation of the sample, the Euclidean distance between $h(x_i)$ and $h(x_j)$ is more appropriate than the distance of the probability $f(x)$, which represents the class to which the sample belongs, and thus the loss of smoothness can be obtained as follows.

$$\ell_s\left(\theta, \mathcal{L}, U, W_{ij}\right)=\begin{cases}\sum\limits_{x_i, x_j \in D}\left\|h\left(x_i\right)-h\left(x_j\right)\right\|^2 & \text{if } W_{ij}=1 \\ \sum\limits_{x_i, x_j \in D} \max\left(0,\left\|m-h(x_i)-h(x_j)\right\|^2\right) & \text{if } W_{ij}=0\end{cases} \tag{8}$$

Combining the above ℓ_c, ℓ_s and the definition of cross-entropy loss, the overall loss function of MSASL is obtained as follows.

$$\begin{aligned}\ell=&-\frac{1}{|B|}\sum_{i\in(B\cap\mathcal{L})}\log\left[f\left(x_i;\theta\right)\right]y_i+w_c(t)\mu_c\frac{1}{|B|}\sum_{i\in B}d\left(f\left(x'i;\theta\right),y'\right)\\&+w_s(t)\mu_s\frac{1}{|S|}\sum_{i\in S}\ell_s(\theta,S)\end{aligned} \tag{9}$$

where $w_c(t)$ and $w_s(t)$ are linearly increasing weight functions.

For a mini-batch of size b, the time complexity required to compute the adjacency matrix between samples, W, is n^2. The time complexity required to draw samples to compute S, based on the fact that W will be very slow in computing the parameters of the gradient descent update, is S^2. Due to the fact that $S << n$, the time complexity will be greatly reduced.

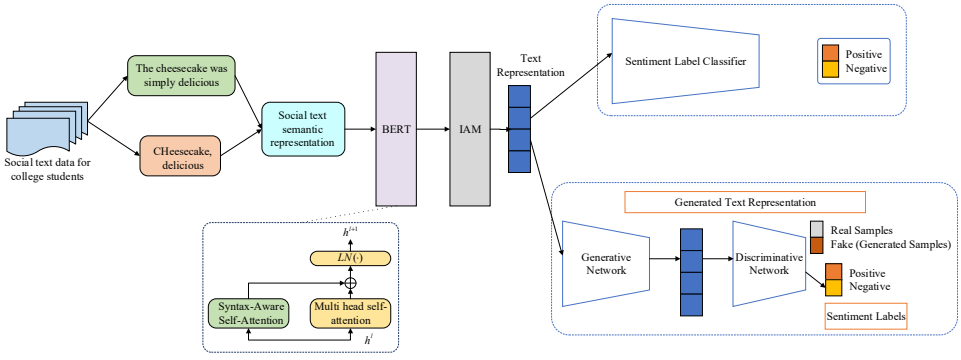
4 SSL-based social-psycho-emotional characterisation of college students

4.1 Semantic representation of social text for college students based on BERT modelling

To solve the issue that the acquisition of a larger amount of labelled data in existing studies needs to consume more resources, which leads to the unsatisfactory efficiency of psycho-emotional feature analysis, this paper proposes a model for analysing the social psycho-emotional features of college students based on the MSASL algorithm mentioned above, as shown in Figure 4.

The model encodes the incoming social texts of college students and obtains the corresponding representations. By adding a syntax-aware self-attention header to the BERT model, and using the interactive attention layer to capture important feature information related to the opinion tendency of the topic words, the feature extraction capability can be improved. By adding noise vectors to the text representation vectors and then generating new samples, the idea of semi-supervised GAN is applied to optimise the text feature representation, and the information from the features is used as the input of the emotion classifier to achieve the classification of extracting the psychological and emotional features of college students.

Figure 4 A model for analysing social psychological-emotional characteristics of college students based on MSASL (see online version for colours)



This article obtains semantic representations of college students' social texts based on the BERT model (Acheampong et al., 2021). BERT utilises its unique Transformer structure to help solve the problem of word polysemy that exists in social texts. The input of the model is the pre-processed social text of college students, and the vector representation matrix of sentences is obtained. On this basis, the input can be in the form of a whole sentence or a sentence pair consisting of a sentence and a topic word, where a sentence may contain multiple topic words, and the whole sentence input affects the subsequent categorisation of affective features. Based on this, this chapter adopts the input of sentence pairs for the task of psycho-sentiment analysis, and the representation of sentence pairs is shown below.

$$x = ([CLS] + d_1, \ldots, d_n + [SEP] + a_1, \ldots, a_m + [SEP]) \tag{10}$$

where a is the topic word, d is the length, $[CLS]$ is the categorisation label, and $[SEP]$ is used to separate the topic word from the sentence, is a sentence separator, and also marks the end of the sentence.

4.2 Social text feature extraction for college students based on interactive attention mechanism

After obtaining the input of sentence pairs, a pre-trained BERT model is used as a feature encoder to obtain the shared feature space of different topic words. This paper improves the BERT model (OBERT) by incorporating external dependent syntactic knowledge (Liang et al., 2021) into the pre-trained BERT model and combining it with the interactive attention mechanism (IAM) to improve the extraction of shared features. OBERT is mainly formed by stacking layers of transformer encoder components. The transformer encoder consists of a multi-head self-attention (MAM) and a feed-forward network layer (FNN) connected by means of residuals and normalisation, the MAM is constructed through multiple self-attention structures to obtain contextual semantic relations, and the FNN layer is used to perform linear transformations.

Three different inputs, query vector Q, key vector K and value vector V, are constructed in MAM, which are computed by inner product and then normalised by the following formula.

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \tag{11}$$

The attention feature information is then computed and output as $f_i = A_i V_i$, where f_i is the matrix of $T \times T$ and d is the embedding dimension, and all the attention header features f_i are spliced and passed into the FNN to obtain $SA = FF(f_1, f_2, \ldots, f_i)$. In the improvement process, p_{parse} is first generated by the syntax-dependent parser, and the product is done on the basis of the previous MAM structure to train the self-attention to obtain syntactic perception. The resulting syntax-aware formula is as follows.

$$A_{parse} = \text{softmax}\left(\frac{\left(Q_{parse} K_{parse}^T\right) * p_{parse}}{\sqrt{d_k}}\right) \tag{12}$$

As in the original MAM structure, the output of the attention head through the FNN is shown in equation (13). Finally, the BERT function is modified to add a syntax-aware self-attentive head to each attention layer of the BERT model in parallel, as shown below.

$$SA_{parse} = FF\left(A_{parse} V_{parse}\right) \tag{13}$$

$$h^{l+1} = LN\left(h^l + SA(h^i) + SA(h^i)\right) \tag{14}$$

where LN is the normalisation function and h^l is the hidden representation of the l^{th} level.

After obtaining the hidden representation of the sentence, the IAM is used to make the sentence feature representation pay more attention to the features that have a higher degree of association with the topic word. Firstly, the interaction matrix of the topic word and context word features is calculated, then the rows in the matrix are normalised to get α, and the columns are normalised to get α. After obtaining the feature representation, then the interaction operation is performed to obtain the context attention score η which is more relevant to the topic word.

$$\eta = \frac{1}{n} \sum_i \alpha \cdot \beta^T \tag{15}$$

4.3 Classification of college students' social psychological-emotional characteristics based on semi-supervised GANs

Existing deep learning-based mental-emotional feature models are limited by the size of the labelled dataset, and the emotion classification ability is reduced when sufficient labelled datasets are not available. While the MSASL algorithm and GAN proposed in Section 3 utilise a small amount of labelled data and a large amount of unlabeled data for training, so this paper is based on MSASL-GAN to complete the mental-emotional feature classification.

1 Generator: Generator G is a multi-layer perceptron (MLP) consisting of an input layer, a hidden layer, and an output layer, whose purpose is to generate a fake vector representation of text similar to social text. The input to G is a noise vector z, and the output is a false text vector representing h_{fake}. Firstly, 100-dimensional z is randomly

extracted from the normal distribution $N(\rho, \sigma^2)$; then, z is fully connected through the hidden layer. Finally, the generated h_{fake} is obtained in the output layer.

$$h_{fake} = g(V * z + b) \tag{16}$$

where V is the hidden layer neuron weight coefficient matrix, b is the bias matrix, and g is the LeakyRelu activation function as follows, where a denotes the learning rate.

$$g(x) = \begin{cases} ax & x < 0 \\ x & x \geq 0 \end{cases} \tag{17}$$

2 Discriminator: Discriminator (D) is another MLP whose task is to classify psycho-emotional features of real social texts and fake social texts. The input of D is the vector h^*, h^* can be h_{fake} or h, and the output is a vector of $k + 1$ dimensional logits. The specific process is as follows: firstly, h^* is fully connected through the hidden layer, and then the vector representation obtained from the obscured level is input to the softmax classification layer for psychological and emotional feature classification. The obscured level in D performs the fully connected operation as in equation (16), and the softmax classification layer is formulated as shown below.

$$y' = P\left(p_i \middle| h\right) = \text{softmax}\left(V^T h + b\right) \tag{18}$$

$$\text{softmax}\left(x_i\right) = \frac{\exp\left(x_i\right)}{\sum_{i=1}^{k+1} \exp\left(x_i\right)} \tag{19}$$

where V^T is the parameter matrix and b is the bias.

After obtaining the classification function, the weight coefficients are updated by cross entropy as shown in equation (20), where N is the sample size, y is the true sentiment label, and y' is the predicted sentiment label.

$$\ell_t = \frac{1}{N} \sum_{1}^{N} y \log\left(y'\right) \tag{20}$$

Thus the total loss of combining MSASL and GAN for psycho-emotional feature classification is shown in equation (21).

$$\ell = \ell_t + \mu_c \ell_c + \mu_s \ell_s \tag{21}$$

where ℓ_c and ℓ_s are the consistency loss and smoothing loss as shown in equations (7) and (9), respectively, μ_c and μ_s are the balancing factors.

5 Experimental results and analyses

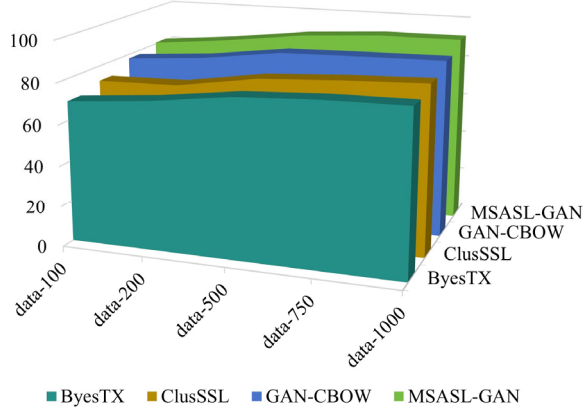
This paper utilises the data of 20,975 social text emotional comments of 935 students in higher education crawled from online social platforms in the literature (Saha et al., 2022) as a dataset, which is labelled with four categories of psycho-emotional features, namely,

positive, negative, ordinary, and confused. During the experiment all the data are categorised into labelled data, unlabeled data and test data, while the dataset is cut into five categories, data-100, data-200, data-500, data-750 and data-1000, as shown in Table 1. The algorithm is based on python language and third-party libraries, using pytorch framework, the experiment is really Linux system, using GTX 1080Ti graphics card and NVDIA launched CUDA 11.0 under the execution of the algorithm. The parameters of the experiment are batch size is set to 32, epoch is 100, the initial learning rate is 0.01, and the optimiser selected is Adam.

Table 1 Slice and dice table for social text dataset for college students

Dataset	*Data-100*	*Data-200*	*Data-500*	*Data-750*	*Data-1000*
Labelled data	100	200	500	750	1,000
Unlabeled data	3,479	3,398	3,172	2,928	2,593
Test data	571	571	571	571	571

Figure 5 Categorisation accuracy with different label data volumes (see online version for colours)



Students' psycho-emotional characterisation belongs to the field of text categorisation, and in this paper, the performance is measured by using the common metrics of accuracy (Acc), precision (Prec), recall (Rec) and F1-value in the field of text categorisation. The classification accuracies of ByesTX method, ClusSSL method, GAN-CBOW method and the proposed method MSASL-GAN with different amount of labeled data are shown in Figure 5. The accuracy of MSASL-GAN is 80.69% at a labelled data volume of 100, which improves 11.38%, 8.06%, and 2.77% compared to ByesTX, ClusSSL, and GAN-CBOW, respectively. The accuracy of the proposed method exceeds the accuracy of ByesTX, ClusSSL, and GAN-CBOW at a tag volume of 500 at a tag volume of 750. Also at a labelling amount of 750, the accuracy of the proposed method exceeds the accuracy of ByesTX, ClusSSL, and GAN-CBOW at a labelling amount of 1,000. It can be seen that the proposed method can achieve the classification results of other models when using a larger amount of data labels while reducing the cost of data labelling.

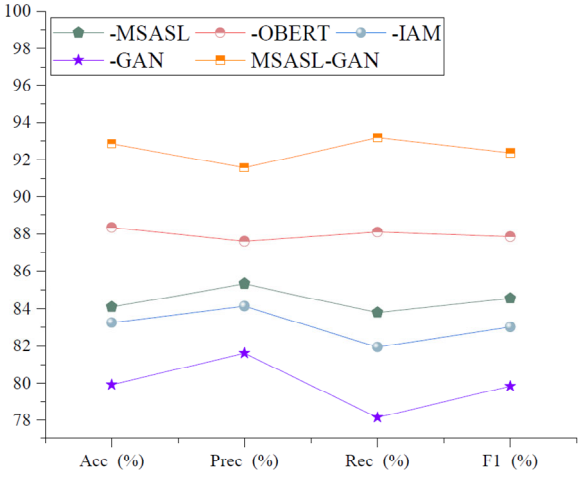
Comparisons of the overall classification performance of the different methods are shown in Table 2. F1 integrates the model's ability to check all and to check accuracy, and the combined consideration of Acc and F1 can more accurately assess the model's

classification effectiveness. The Acc and F1 of MSASL-GAN are 92.84% and 92.36%, respectively, which is an improvement of 11.51% and 11.25% compared to ByesTX, 9.62% and 8.89% compared to ClusSSL, and 5.01% and 4.33% compared to GAN-CBOW. MSASL-GAN not only optimises the SSL algorithm using consistency regularity and streaming regularity, but also improves the BERT model, combines the IAM, highlights the key features of college students' social texts, and combines MSASL and GAN to classify the psychological emotional features, which is able to efficiently complete the task of classifying college students' social texts' emotions while reducing the cost of data labelling.

Table 2 Classification performance of psycho-emotional profiling methods

Method	*ByesTX*	*ClusSSL*	*GAN-CBOW*	*MSASL-GAN*
Acc (%)	81.33	83.22	88.51	92.84
Prec (%)	79.72	82.04	85.38	91.56
Rec (%)	82.55	84.96	89.42	93.17
F1 (%)	81.11	83.47	87.35	92.36

Figure 6 Results of ablation experiments with different components in MSASL-GAN (see online version for colours)



To further validate the impact of each component in MSASL-GAN on classification performance, this paper conducts ablation experiments to study MSASL-GAN, where -MSASL denotes semi-supervised deep learning using traditional SSL, -OBERT denotes social text semantic representation using unoptimised BERT model, and -IAM denotes removing the IAM module, and directly using the IBERT representation of feature vectors as model input. -GAN denotes removing the semi-supervised module and using only GAN for psycho-emotional feature classification. The ablation results for different components are shown in Figure 6.

As can be seen in Figure 6, -OBERT outperforms all other groups of models and the gap between them and the full model is small, which indicates that the classification accuracy can be improved to some extent by incorporating external dependent syntactic knowledge into the BERT model. -MSASL performs sub-optimally, with its F1 value

reduced by 7.8% compared to MSASL-GAN, indicating that performance optimisation using the introduction of flow regularisation on the SASL algorithm is necessary. -GAN has the worst performance in classification. This is because for a large number of datasets, GAN requires a large amount of computational resources in the case of unlabeled samples, whereas by introducing MSASL into GAN, the computational consumption can be greatly reduced and thus the classification accuracy can be improved. The classification performance of -IAM is also lower than that of MSASL-GAN, indicating that IAM pays more attention to features with higher correlation with subject words, reduces feature redundancy, and improves classification accuracy. Therefore, MSASL-GAN integrated with all modules achieves the best classification performance in the psychological and emotional feature analysis task.

6 Conclusions

Social platform is an essential channel for college students to express their emotional attitudes, and the emotional information contained in its texts also provides data support for college students' psycho-emotional characterisation. Intending to the issues of high cost of data tagging and insufficient semantic extraction of social texts in the current psycho-emotional characterisation methods, this paper suggests a SSL-based psycho-emotional characterisation method for college students. Firstly, to address the disadvantage of losing the structural information of the dataset based on the consistency semi-supervised algorithm, a MSASL algorithm fusing the consistency regularity with the streaming regularity is proposed to ensure its anti-local perturbation while adding the streaming regularity to fuse more information. Then the BERT model is used to semantically represent the social text, and then the IAM is used to capture the key feature information related to the opinion tendency of the topic words, so as to improve the ability of feature extraction. Finally, by adding noise vectors to the text representation vector and then generating new samples, MSASL-GAN is applied to optimise the text feature representation, and the feature information is used as the input to the sentiment classifier, which outputs the classification results of psychological sentiment features through the fully connected layer. Experimental outcome indicates that the offered approach improves the accuracy of sentiment feature classification while reducing the cost of data labelling.

To some extent, the suggested approach provides valuable contributions to the field of emotional feature analysis, but due to time constraints, there are still some shortcomings. In the future, the chatting emoticons, pictures, voices and other data generated by college students on social platforms will be added for multi-modal emotional feature analysis, so as to further improve the generalisation ability of the suggested approach.

References

Acheampong, F.A., Nunoo-Mensah, H. and Chen, W. (2021) 'Transformer models for text-based emotion detection: a review of BERT-based approaches', *Artificial Intelligence Review*, Vol. 54, No. 8, pp.5789–5829.

Barreto, S., Moura, R., Carvalho, J. et al. (2023) 'Sentiment analysis in tweets: an assessment study from classical to modern word representation models', *Data Mining and Knowledge Discovery*, Vol. 37, No. 1, pp.318–380.

Chen, J., Feng, J., Sun, X. et al. (2019) 'Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts', *Symmetry*, Vol. 12, No. 1, p.8.

Durga, P. and Godavarthi, D. (2023) 'Deep-sentiment: an effective deep sentiment analysis using a decision-based recurrent neural network (D-RNN)', *IEEE Access*, Vol. 11, pp.108433–108447.

Galay, I. and Aizman, R. (2023) 'Gender features of the psychosocial state of students in the period of adaptation to the educational environment of the pedagogical university', *Research and Advances in Education*, Vol. 2, No. 3, pp.23–27.

Isnain, A.R., Supriyanto, J. and Kharisma, M.P. (2021) 'Implementation of K-nearest neighbor (K-NN) algorithm for public sentiment analysis of online learning', *Indonesian Journal of Computing and Cybernetics Systems*, Vol. 15, No. 2, pp.121–130.

Kokab, S.T., Asghar, S. and Naz, S. (2022) 'Transformer-based deep learning models for the sentiment analysis of social media data', *Array*, Vol. 14, p.100157.

Kumar, A. and Garg, G. (2019) 'Sentiment analysis of multimodal twitter data', *Multimedia Tools and Applications*, Vol. 78, pp.24103–24119.

Li, Y., Zhao, Z., Sun, H. et al. (2020) 'Snowball: iterative model evolution and confident sample discovery for semi-supervised learning on very small labeled datasets', *IEEE Transactions on Multimedia*, Vol. 23, pp.1354–1366.

Liang, Y., Meng, F., Zhang, J. et al. (2021) 'A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis', *Neurocomputing*, Vol. 454, pp.291–302.

Liu, B. (2020) 'Text sentiment analysis based on CBOW model and deep learning in big data environment', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 11, No. 2, pp.451–458.

Matuszyk, P. and Spiliopoulou, M. (2017) 'Stream-based semi-supervised learning for recommender systems', *Machine Learning*, Vol. 106, pp.771–798.

Menaouer, B., Zahra, A.F. and Mohammed, S. (2022) 'Multi-class sentiment classification for healthcare tweets using supervised learning techniques', *International Journal of Service Science, Management, Engineering, and Technology*, Vol. 13, No. 1, pp.1–23.

Moeller, R.W., Seehuus, M. and Peisch, V. (2020) 'Emotional intelligence, belongingness, and mental health in college students', *Frontiers in Psychology*, Vol. 11, p.499794.

Riyadh, M. and Shafiq, M.O. (2022) 'GAN-BElectra: enhanced multi-class sentiment analysis with limited labeled data', *Applied Artificial Intelligence*, Vol. 36, No. 1, p.2083794.

Ruz, G.A., Henríquez, P.A. and Mascareño, A. (2022) 'Bayesian constitutionalization: Twitter sentiment analysis of the Chilean constitutional process through Bayesian network classifiers', *Mathematics*, Vol. 10, No. 2, p.166.

Saha, K., Yousuf, A., Boyd, R.L. et al. (2022) 'Social media discussions predict mental health consultations on college campuses', *Scientific Reports*, Vol. 12, No. 1, p.123.

Shamoi, E., Turdybay, A., Shamoi, P. et al. (2022) 'Sentiment analysis of vegan related tweets using mutual information for feature selection', *PeerJ Computer Science*, Vol. 8, p.e1149.

Vlachostergiou, A., Caridakis, G., Mylonas, P. et al. (2018) 'Learning representations of natural language texts with generative adversarial networks at document, sentence, and aspect level', *Algorithms*, Vol. 11, No. 10, p.164.

Zhou, K., Zeng, J., Liu, Y. et al. (2018) 'Deep sentiment hashing for text retrieval in social CIoT', *Future Generation Computer Systems*, Vol. 86, pp.362–371.