Sentiment-guided Transformer for depression detection with severity-aware contrastive learning

Tianlin Zhang[a,*], Kailai Yang[a] and Sophia Ananiadou[a,b]

[a]National Centre for Text Mining, Department of Computer Science, The University of Manchester, Manchester, M1 7DN, UK
[b]The Alan Turing Institute, London, NW1 2DB, UK

ARTICLE INFO

Keywords:
depression severity detection
natural language processing
sentiment analysis
contrastive learning

ABSTRACT

Depression is a common mental disorder worldwide. Early identification of depression is beneficial to public health surveillance and disease treatment. Recently, advances in natural language processing (NLP) have played an essential role in supporting the analysis of user-generated posts on social media, including early mental illness detection. There are many models that mainly treat the detection as a binary classification task, such as detecting whether a user is depressed. However, identifying users' depression severity levels from posts on social media is more clinically useful for future prevention and treatment. Existing severity detection methods mainly model the semantic information of posts while ignoring the relevant sentiment information, which can reflect the user's state of mind and could be helpful for severity detection. In addition, they treat all severity levels equally, making the model difficult to distinguish between closely-labeled categories. We propose a sentiment-guided Transformer model, which efficiently fuses social media posts' semantic information with sentiment information. Furthermore, we also utilise a supervised severity-aware contrastive learning framework to enable the model to better distinguish between different severity levels. Experimental results show that our model achieves new state-of-the-art results on two public datasets, while further analysis proves the effectiveness of all proposed modules.

1. Introduction

According to the World Health Organization (WHO)[1], depression is a common mental disorder and a leading cause of disability across the world. It is estimated that around 350 million people suffer from depression and over 70% of people do not receive timely treatment [1]. Therefore, if depression is detected at an early stage, it could be useful for disease intervention and treatment. In addition, with social media becoming increasingly popular, people often share their feelings and mental states through social media platforms such as Twitter and Reddit, making social media posts critical for mental illness detection, including depression detection [2, 3].

In recent years, advances in natural language processing (NLP) have played an increasingly essential role in supporting the analysis of user-generated contents from social media, including mental illness surveillance [3, 4, 5]. There are many methods leveraging NLP technologies for automated depression detection. Early works utilise feature extraction and traditional machine learning models such as SVM [6], Logistic Regression [7] to detect depression. Recent methods employ deep learning-based models to capture latent semantic information for depression detection without feature engineering. Convolutional Neural Networks (CNN) [8, 9] or Recurrent Neural Networks (RNN) [10, 11] have been explored for that purpose. Furthermore, the Pre-trained Language Models (PLMs) such as BERT [12], RoBERTa [13] and MentalRoBERTa [14] are widely used for mental health prediction given their context modelling ability, achieving good performance on several mental illness detection tasks. Nevertheless, existing depression detection methods mainly treat detection as a binary classification task due to the limitation of existing annotated datasets. According to the Beck's Depression Inventory (BDI) [15], not all cases of depression are the same; depression could be classified as: *minimal*, *mild*, *moderate* and *severe*. For our methods to be translational, detecting the severity level of depression is expected to be more important for further prevention and treatment [16]. For instance, evidence-based psychological therapies could be appropriate for mild or moderate depression [17], while users with severe level might be provided more resource-intensive interventions [18].

Previous psychological studies [19, 20] have examined the correlation between sentiment and depression, which have illustrated that sentiment is able to reflect the user's state of mind and help to identify depressed people [21]. For example, a depressed user who posts a comment "*Today was a really bad day. I had no energy and thought about suicide all day*" with a negative sentiment and those who are not depressed tend to express a positive or neutral attitude, like "*idk why I'm doing this. I guess I just am. Maybe it will help someone else*". However, existing severity detection methods mainly model the semantic information of posts while ignoring the relevant sentiment information. Meanwhile, traditional classification models with cross-entropy loss treat all severity levels equally, making it difficult to distinguish between closely-labeled categories. For example, '*moderate*' and '*severe*' are semantically closer in severity level than '*mild*' and '*severe*', making them more difficult to distinguish. Therefore, in order to tackle the above challenges, we propose a sentiment-guided Transformer model with

*Corresponding author
tianlin.zhang@postgrad.manchester.ac.uk (T. Zhang)
ORCID(s): 0000-0003-0843-1916 (T. Zhang); 0000-0003-3142-2516 (K. Yang); 0000-0002-4097-9191 (S. Ananiadou)

[1]https://www.who.int/health-topics/depression

severity-aware contrastive learning for depression severity detection. For the first challenge, we leverage post encoders MentalRoBERTa and SentiLARE [22] to obtain the semantic as well as sentiment hidden features and then design a sentiment-guided Transformer model to better fuse these two features. For the second challenge, we implement a supervised severity-aware contrastive learning framework, which helps the model to adaptively differentiate the weights between different negative samples according to the label information. We validate our method on two depression severity detection datasets. Experimental results show that our model consistently outperforms other baselines, such as DepressionGCN [17] and MentalRoBERTa [14]. Our main contributions can be summarized as follows:

(1) We propose a sentiment-guided Transformer fusion model to fuse sentiment information into the depression severity detection task. This network can better combine semantic information and sentiment information compared with existing PLMs-based models.

(2) We implement a supervised severity-aware contrastive learning framework to enable our model to better capture class-specific features and distinguish between different severity levels.

(3) We conduct experiments and further analysis on mainstream public datasets. Experimental results illustrate that our model achieves competitive performance compared with others and each of the proposed modules is effective in depression detection.

2. Related work

2.1. Depression detection using social media

With the fast growing numbers of social media users, using user-generated posts to detect depression manually is no longer practical. Automated NLP and text mining technologies provide new opportunities for mental health analysis [4, 5, 23].

Early researchers [24, 25, 26] extracted linguistic features (e.g., Linguistic Inquiry and Word Count (LIWC), Bag-of-Words) and statistical features (e.g., n-gram, Term Frequency-Inverse Document Frequency (TF-IDF)) from posts, then employed them in traditional machine learning methods such as SVM [6], Logistic Regression [7] and Random Forest [27] to detect depression.

Recently, deep learning-based models have been used for depression detection, allowing models to automatically capture valuable features without feature engineering. Some works have shown that CNN [8, 9] and RNN [10, 11], including Long Short Term Memory (LSTM) [28] and Gated Recurrent Unit (GRU) [29] are effective. Furthermore, PLMs such as BERT [12], RoBERTa [13] are also widely utilised to many mental illness detection tasks due to strong context modelling ability. In particular, Ji et al. [14] trained domain-specific language model MentalRoBERTa for mental healthcare, which introduces domain-related knowledge and outperforms basic PLMs on several mental health datasets. The depression severity identification task is becoming increasingly attractive due to the practical clinical implications, so some studies have emerged [30, 31, 32]. Although the severity level is measured on an ordinal scale, current methods treat it as a traditional multi-classification task. [17] first reformulate depression identification as an ordinal classification problem via ordinal loss and employ Graph Convolutional Neural Network (GCN) to capture contextual information, which achieves state-of-the-art performance on depression severity datasets.

2.2. Sentiment analysis and contrastive learning

Sentiment analysis is one of the key technologies in NLP, which aims to analyse people's sentiments or opinions expressed in texts (usually classified as positive, negative and neutral). Sentiment analysis is widely applied to reviews, social media, newswire, and medical informatics [33, 34]. Recently, PLMs-based models have achieved competitive performances on sentiment analysis. For example, Yin et al. [35] propose SentiBERT, a variant of BERT model that effectively captures compositional sentiment semantics. Ke et al. [22] present SentiLARE with context sentiment mechanism and knowledge integration, which facilitates sentiment understanding.

Contrastive learning is a recent technique for enhancing the performance of models by learning different representations of contrasting samples, which first originates from computer vision (CV) [36]. The main purpose is to generate negative pairs using data augmentation and minimise the contrastive loss of the positive pairs. Similar methods are transferred to NLP tasks, like ConSERT [37] for better unsupervised sentence representation learning. Supervised contrastive learning can leverage label information for representation learning in a supervised setting [38] , which improves the ability of the model to differentiate between samples with different labels. Recent work also uses supervised contrastive learning for depression detection tasks [39, 40], illustrating its effectiveness.

3. Method

3.1. Task definition

We focus on detecting the depression severity level of users by analysing posts in social media. Formally, given a post $P = \{s_1, s_2, ..., s_i\}$, where s_i is the i-th sentence of the post, our goal is to classify each post into a corresponding depression severity level label $y \in Y$, where Y is the list of increasing depression severity levels. For example, according to the definition of different datasets described in Section 4.1, Y is {*minimal*, *mild*, *moderate*, *severe*} or {*not depressed*, *moderately depressed*, *severely depressed*}. This task can be formalised as a multi-class text classification problem.

3.2. Depression severity detection model

Our model architecture (Figure 1) consists of a post encoder, a sentiment-guided Transformer and a supervised

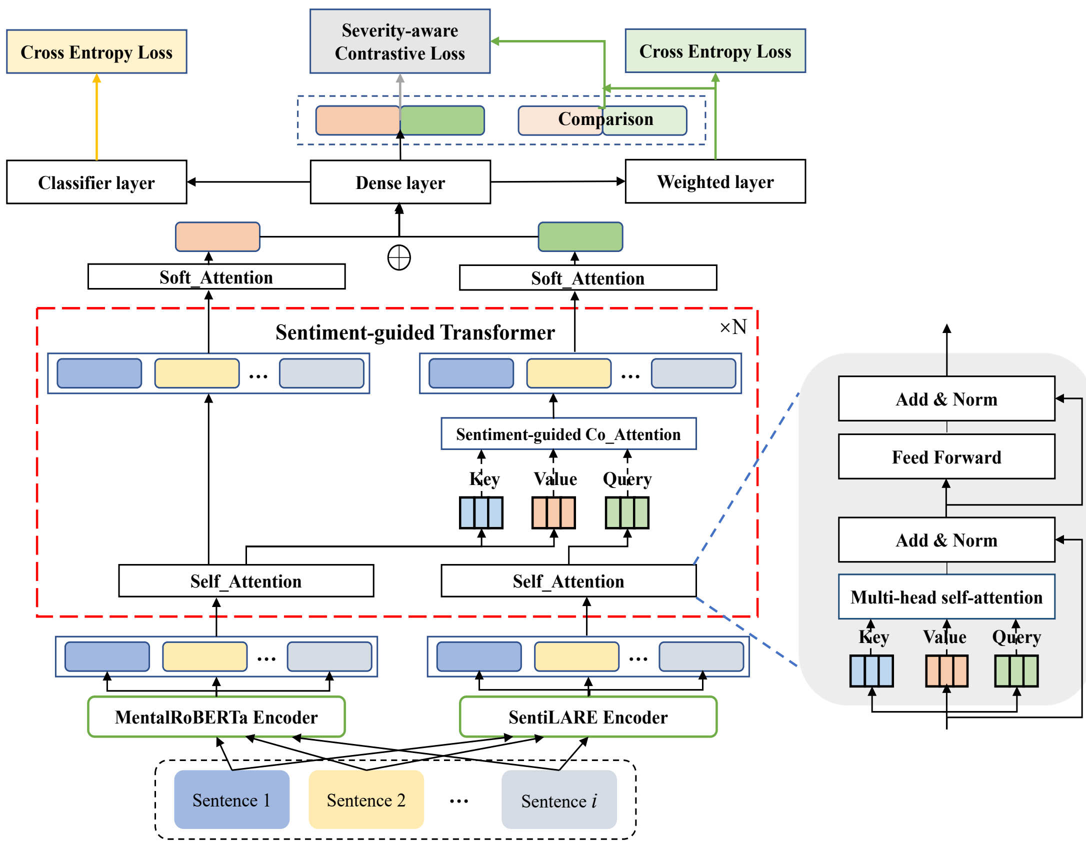

Figure 1: The overall architecture of our proposed model.

severity-aware contrastive learning component. We will introduce the details of each component as follows.

3.2.1. *Post encoder*

With the huge success of PLMs in several NLP tasks, PLMs are widely used as sentence encoder for various tasks, including for mental health detection [23]. In particular, Ji et al. [14] trained a domain-specific language model MentalRoBERTa for mental healthcare, which introduces domain-related knowledge and outperforms basic PLMs on several mental health datasets. We exploited MentalRoBERTa as the semantic encoder. Since we require sentiment information to help us to better understand a user's state of mind, we also adopt the pre-trained language model SentiLARE [22] that achieves good performance on various sentiment analysis tasks, as the sentiment encoder to obtain the sentiment hidden representation of each sentence. Specifically, given a post $P = \{s_1, s_2, ..., s_i\}$, we have obtained i split sentences and each sentence is $s = \{[\mathrm{CLS}], x_1, x_2, ..., x_j\}$, where [CLS] is a special token for marking the start of the sentence and x_j is the j-th word. Then, we feed each sentence sequence into MentalRoBERTa and SentiLARE to get the two embeddings of the [CLS] as the corresponding sentence embeddings. We denote the process as:

$$\mathrm{H}^m = \{\mathrm{H}_1^m, \mathrm{H}_2^m, ..., \mathrm{H}_i^m\} = Mental\_Encoder(\{s_1, s_2, ..., s_i\}) \quad (1)$$

$$\mathrm{H}^s = \{\mathrm{H}_1^s, \mathrm{H}_2^s, ..., \mathrm{H}_i^s\} = Senti\_Encoder(\{s_1, s_2, ..., s_i\}) \quad (2)$$

where $Mental\_Encoder$ and $Senti\_Encoder$ denote the MentalRoBERTa encoder and SentiLARE encoder respectively; H_i denotes the i-th sentence embedding; H^m and $\mathrm{H}^s \in \mathbb{R}^{n\times d}$, n denotes the number of sentences, and d is the hidden dimension of the encoder.

3.2.2. *Sentiment-guided Transformer*

The architecture of the Sentiment-guided Transformer is illustrated in the red-dotted box Figure 1, which is composed of a stack of N blocks with their own training parameters. Each block contains a Self_Attention module and a Sentiment-guided Co_Attention module. For one social media post, there are two post embeddings generated as inputs $\{\mathrm{H}^m; \mathrm{H}^s\}$ for sentiment-guided Transformer.

Firstly, the two embeddings are fed into Self_Attention module separately. This module is the most important component of Googles' Transformer [41] that learns the sequence's representation consisting of a multi-head self-attention mechanism (MHA), a feed-forward network and two layers of normalization [42] with residual connection [43]. The self-attention mechanism can be described as the relationships between a query and a set of key–value pairs to compute an output, where query, key, value and output are vectors. For semantic embedding, given the matrix H^m, we can get a query matrix $Q \in \mathbb{R}^{n\times d}$, a key matrix $K \in \mathbb{R}^{n\times d}$ and a value matrix $V \in \mathbb{R}^{n\times d}$. The self-attention function is expressed as:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{3}$$

where n is the length of the post, d is the embedding hidden dimension, $softmax$ denotes softmax operation, $\sqrt{d}$ is a scaling factor. $Q = K = V = \mathrm{H^m}$.

To allow the model to jointly learn contextual information from different representation sub-spaces, the three matrices Q, K, V are multiplied with the W_i^Q, W_i^K, W_i^V respectively, and the MHA with h heads is used. Then, the results of all attention heads are concatenated together, and a weight matrix W^O is used to obtain the final output of the encoder:

$$MHA(Q,K,V) = Concat(head_1, ..., head_h)W^o \tag{4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

where parameter weight matrices $W_i^Q \in \mathbb{R}^{d\times d_h}$, $W_i^K \in \mathbb{R}^{d\times d_h}$, $W_i^V \in \mathbb{R}^{d\times d_h}$ and $W^o \in \mathbb{R}^{d\times d}$, $d_h = d/h$.

After Self_Attention module, we will get the semantic-level output SA^m and the sentiment-level output SA^s. In order to obtain contextual sentiment-guided sentence representation, we leverage the Sentiment-guided Co_Attention module to fuse sentiment information with the corresponding semantic information. In contrast to the Self_Attention, the inputs of Sentiment-guided Co_Attention are from two different embeddings.

We switch the query Q^s from the Self_Attention output SA^s (the sentiment-level post hidden representation), the key K^m and the value V^m from SA^m (the semantic-level post hidden representation). The calculation process is the same as the multi-head self-attention mechanism:

$$Co\_Attention(Q^s, K^m, V^m) = softmax(\frac{Q^s(K^m)^T}{\sqrt{d}})V^m \tag{6}$$

After all the Sentiment-guided Transformer blocks are computed, we obtain the final block's outputs including the semantic hidden feature representation $o^m = \{o_1^m, o_2^m, ..., o_i^m\}$ and the sentiment-guided hidden feature representation $o^s = \{o_1^s, o_2^s, ..., o_i^s\}$. Considering the different importance of each sentence to the post, we leverage Soft_Attention mechanism [44] to get semantic-level post embedding p^m and sentiment-level post embedding p^s:

$$u_i = \tanh(W_p o_i + b_p) \tag{7}$$

$$\alpha_i = \frac{\exp(u_i{}^T u_p)}{\sum_i \exp(u_i{}^T u_p)} \tag{8}$$

$$p = \sum_i \alpha_i o_i \tag{9}$$

Where W_p is a weight matrix, b_p is a bias vector, and u_p is a post context vector which is randomly initialized.

Finally, we concatenate these two embeddings $p^f = concat(p^m, p^s)$ as the fusion post embedding that retains contextual semantic and sentiment information.

3.2.3. Supervised severity-aware contrastive learning

Recent works [38, 39, 40] use supervised contrastive learning to better capture the class-specific information in order to make sentences with the same label cohesive and different labels mutually exclusive. In addition, following [45], we design a supervised severity-aware contrastive learning framework to help the model differentiate the more difficult negatives.

To ensure that there are at least two examples of a class in one batch for contrastive learning, according to [39], we copy each post embedding to produce augmented examples with the batch size will changing from N_b to $2N_b$. The traditional supervised contrastive loss is given by:

$$L_{SCL} = \sum_{i=1}^{2N_b} -\frac{1}{|J|}\sum_{j\in J}\log\frac{\exp(h_i^f \cdot h_j^f/\tau)}{\sum_{k\in I\setminus i}\exp(h_i^f \cdot h_k^f/\tau)} \tag{10}$$

$$h_i^f = L2 - Normalization(Dense\_layer(p_i^f)) \tag{11}$$

where $J = \{j : j \in I, y_j = y_i \wedge j \neq i\}$, i is the sample index, h_i^f is the L2-normalised vector of post hidden representation obtained from an encoder (dense layer in Figure 1), and τ indicates the temperature hyper-parameter.

We can see that traditional supervised contrastive learning treats both negative samples (the comparison in Figure 1) the same. However, in the depression severity detection task, the labels contain an inherent ordinal nature. For example, the sample with the label *severe* is semantically closer to the sample with the label *moderate* than the one with the label *mild* in the representation space. Therefore, to distinguish better the closely-labeled examples, we introduce supervised severity-aware contrastive loss for adaptively weighting the samples based on the relationships between the severity labels.

In order to obtain the adaptive weights for different comparison samples, following [45], we use a dual-model strategy as shown in Figure 1. The post hidden representations are fed into a weighted layer that is same as the classifier layer, and the output is optimised using cross-entropy loss L_w. Then, by using softmax function to obtain prediction probabilities, we can get the confidence scores of the sample that are classified to different classes:

$$hh_i^f = weighted\_layer(h_i^f) \tag{12}$$

$$w_i = \{w_{i,c}\}_{c=1}^C = \frac{\exp(hh_{i,c}^f)}{\sum_{c=1}^C \exp(hh_{i,c}^f)} \tag{13}$$

where C is the number of the classes, $w_{i,c}$ denotes the confidence that i-th sample is classified to class c, $hh_{i,c}^f$ is the c-th value of hh_i^f.

The w_i is taken as weighting vector to weight the pairwise similarity values, and the severity-aware contrastive loss is as follows:

$$L_{SSCL} = \sum_{i=1}^{2N_b} -\frac{1}{|J|}\sum_{j\in J}\log\frac{w_{i,y_i}\cdot\exp(h_i^f \cdot h_j^f/\tau)}{\sum_{k\in I\setminus i} w_{i,y_k}\cdot\exp(h_i^f \cdot h_k^f/\tau)} \tag{14}$$

Table 1
Datasets Statistics

Stats/Datasets	DsD [17]	DepSign [46]
Number of posts	3,553	13,387
Average number of sentences per post	4.79	9.25
Average number of tokens per post	87.91	154.55
Label	{*minimal*, *mild*, *moderate*, *severe*}	{*not depressed*, *moderate*, *severe*}
Label distribution	44.58% /19.53% /16.60% /19.19%	28.39% /62.19% /9.42%

Table 2
Some examples of the two datasets. The posts have been obfuscated and paraphrased for user privacy.

Dataset	Examples	Label
DsD	It was created by a friend of the h**d. She's in for a long, hard road after she gets done with this f**l. If anyone can help. Please do.	minimal
	I have NOONE to talk too. stress I've taken d**bt classes. Yes I've tried grounding tech**es and get frustrated after a b**ch don't work. Cannot Afford a pyscologist	mild
	Now I want to drink till I'm drunk a**n and my head sounds as though there is yelling when I'm the only one home. At least I have a drs ap**t coming up soon so I can b**g it up.	moderate
	Try and learn me**on. I thought it wasn't my scene but I got an app called H**e and it surprisingly helps. I've been close to killing myself 3 times.	severe
DepSign	Youre all tough. : Youre all s**g people, youve gone t**h so much and made it this far.	not depressed
	Somone else Feeling like 2020 will be t**e last year on earth b**e even wen your hammerd your Feeling like a moron thats depressed?	moderate
	Words can't d**e how bad I feel right now : I just w**t to fall asleep forever.	severe

Where w_{i,y_i} represents the weight between the i-th sample and its corresponding severity label y_i. We can find that the loss assign higher weights to confusable negatives based on the model's confidence scores.

We combine the training of depression severity detection and the supervised severity-aware contrastive learning task in a multi-task learning framework. The framework is jointly optimised using the following training losses: (1) the classification output's cross-entropy loss L_c; (2) the severity-aware contrastive loss L_{SSCL}; (3) the weighted layer output's cross-entropy loss L_w. We formalised the overall training loss as follows:

$$L_{overall} = \alpha(L_c + L_w) + (1-\alpha)L_{SSCL} \tag{15}$$

where α is a tunable weight parameter.

4. Experiments

4.1. Datasets

To evaluate and compare our method with related works, we conducted experiments on two publicly available depression severity detection datasets: DsD[2] [17] and DepSign[3] [46]. The statistical details are given in Table 1. In addition, we also provide some examples of these two datasets in Table 2.

Depression severity Dataset (DsD): The dataset is collected from the social media platform Reddit[4] and annotated according to the disorder annotation scheme [47], where each post is labeled with one of four severity levels: {*minimal*, *mild*, *moderate*, *severe*} following the depression rating scale by [48].

Depression signs detection Dataset (DepSign): The dataset is from shared task and also collected from Reddit. The posts belong to the following subreddits groups, "r/depression, r/mental health". The data is annotated by two domain experts according to the signs of depression, and each post is labeled with the severity level of signs of depression: {*not depressed*, *moderate*, *severe*}. Since the shared task does not make the test set publicly available, we merge the train set and valid set as the whole dataset.

4.2. Evaluation metrics

To better evaluate our model from the point of view of depression severity assessment, we choose the evaluation metrics Graded Precision (GP), Graded Recall (GR) and FScore used by [17, 49]. They altered the formulation of False Positives (FP) and False Negatives (FN). FP is changed as the proportion of the number of times the predicted severity level (l^p) is higher than the actual severity level (l^a). FN is the proportion of the number of times the predicted severity level l^p is lower than the actual severity level l^a.

$$FP = \frac{\sum_{i=1}^{N^T} I(l_i^p > l_i^a)}{N^T}, FN = \frac{\sum_{i=1}^{N^T} I(l_i^a > l_i^p)}{N^T} \tag{16}$$

where $\Delta(l_i^p, l_i^a)$ denotes the difference between predicted l^p and actual l^a depression severity level for post P_i, N^T is the size of test set.

[2] https://github.com/usmaann/Depression_Severity_Dataset
[3] https://competitions.codalab.org/competitions/36410
[4] https://www.reddit.com

The precision and recall are reformulated as GP and GR since FP and FN contain the comparison between l^p and l^a:

$$GP = \frac{TP}{TP+FP}, GR = \frac{TP}{TP+FN} \tag{17}$$

$$FScore = \frac{2GP \cdot GR}{GP+GR} \tag{18}$$

where TP is the True Positives, FScore is the F1-score calculated by GP and GR.

4.3. Implementation details

We perform pre-processing steps on the datasets to clean the texts. We use the *ekphrasis*[5] library that is a processing tool for texts from social media to correct spelling, remove hashtags and normalise some special words (URLs, emails, digits, user names, elongated words, etc.). We further use *emoji*[6] to replace emoticons and emojis with the associated words.

We use grid search to explore the parameters: the number of Sentiment-guided Transformer blocks $N \in \{1, 2, 4, 8\}$, the number of hidden dimension $H \in \{128, 256, 512, 1024\}$, the number of heads $h \in \{2, 4, 8, 16\}$, dropout $\delta \in \{0.1, 0.2, 0.3, 0.4\}$, learning rate $lr \in \{5e-4, 1e-4, 5e-5, 1e-5\}$, the optimization $O \in \{Adam, AdamW\}$, the batch size $b \in \{8, 16, 32\}$, tunable weight parameter $\alpha \in \{0.2, 0.5, 0.8\}$. The optimal hyperparameters used are $N = 4, H = 1024, h = 8, \delta = 0.2, lr = 5e{-}5, O = AdamW, b = 16, \alpha = 0.5$.

We train the model on an Nvidia Tesla V100 GPU with 16GB of memory for 100 epochs and apply early stopping with patience of 20 epochs. We use Huggingface's Transformers library[7] as pre-trained language models. For consistency, we use the same experimental settings and 10-fold cross-validation for all models. Our reported results are averaged across all folds on both datasets.

5. Analysis

5.1. Overall results

We compare our model with recent deep learning-based methods to show the overall performance. Table 3 shows the experimental results of our model and the baselines. The comparative approaches are as follows:

- **DepressionNet** [50] : a text summarization model for depression detection.
- **DepressionGCN** [17] : a recently proposed depression severity detection method combining GCN and Bi-LSTM structures.
- **BERT** [51] : a widely used basic pre-trained language model.
- **MentalBERT** and **MentalRoBERTa** [14] : two domain specific PLMs pre-trained on mental health corpus. Moreover, MentalRoBERTa is the current state-of-the-art model on this task.

[5]https://pypi.org/project/ekphrasis/
[6]https://pypi.org/project/emoji/
[7]https://huggingface.co/models

As shown in Table 3, we notice that PLMs-based models, especially mental health-related PLMs, offer a general advantage with over 85% performance and outperform DepressionGCN on both datasets, which shows the advantage of context modelling for PLMs. Compared with these baselines, our model achieves a new state-of-the-art on both datasets (87.5% / 89.1% of FScore), with 0.8% and 1.9% improvements over the MentalRoBERTa. In summary, the results prove that the proposed model achieves better performance on the depression severity detection task.

Table 3
Experimental performance comparison for different models. We highlight top-1 values in bold.

Model	DsD			DepSign		
	GP	GR	FS	GP	GR	FS
DepressionNet[50]	80.0	70.0	78.0	-	-	-
DepressionGCN[17]	**95.0**	75.0	85.0	84.2	88.3	86.2
BERT[51]	92.6	80.2	85.9	83.7	89.3	86.4
MentalBERT[14]	93.0	80.8	86.5	84.3	89.9	87.0
MentalRoBERTa[14]	94.1	80.4	86.7	84.4	**90.3**	87.2
Our model	93.1	**82.7**	**87.5**	**88.1**	90.2	**89.1**

5.2. Ablation study

In this section, we perform ablation study of our model on both DsD and DepSign datasets to illustrate the effectiveness of each module of our proposed model. Specifically, we conduct module ablation experiment, sentiment fusion ablation experiment and contrastive ablation experiment.

5.2.1. module ablation

To evaluate the contribution of different modules in our model, we perform a module ablation study. 'w/o contrastive' denotes the model without severity-aware contrastive learning module. 'w/o SGT' denotes the model without sentiment-guided Transformer module. 'w/o all' removes all modules while keeping MentalRoBERTa encoder. The results are shown in Table 4. We note that our complete model achieves the best performance while the performance degrades when components are removed. In detail, e.g., with DepSign, the performance of FScore decreases by 0.6% when contrastive learning is removed and the FScore drops by 1.5% when SGT is removed. The model without SGT consistently performs worse than the model without contrastive, which indicates the importance of sentiment information in depression detection. If the two modules are both removed, the model degrades to the basic MentalRoBERTa encoder and the FScore drops by 1.9%. The results above demonstrate that each proposed module works effectively and the performance is improved when these components are combined together.

Table 4
The results of module ablation study. We highlight top-1 values in bold.

Model	DsD			DepSign		
	GP	GR	FS	GP	GR	FS
Our model	93.1	**82.7**	**87.5**	**88.1**	90.2	**89.1**
w/o contrastive	93.2	81.8	87.1	86.8	**90.3**	88.5
w/o SGT	**94.3**	80.6	86.9	85.2	90.2	87.6
w/o all	94.1	80.4	86.7	84.4	**90.3**	87.2

Table 5
The results of sentiment fusion ablation study. We highlight top-1 values in bold.

Model	DsD			DepSign		
	GP	GR	FS	GP	GR	FS
SA(m)	94.1	80.4	86.7	84.4	90.3	87.2
SA(m)+s	92.5	**82.1**	86.9	84.3	**90.4**	87.2
SA(m)+SA(s)	92.7	81.8	86.9	86.3	90.0	88.1
SGT(m+s)	**93.2**	81.8	**87.1**	**86.8**	90.3	**88.5**
EGT(m+e)	92.7	81.7	86.8	86.5	90.1	88.3

5.2.2. *sentiment fusion ablation*

In order to demonstrate the effectiveness of our proposed sentiment fusion strategy (sentiment-guided Transformer), we compare it with other methods. 'SA(m)' means only using hidden representations (m) from MentalRoBERTa encoder combining Self_Attention module. 'SA(m)+s' denotes the concatenation that the representations from SA(m) are concatenated with a sentiment label, which is a common fusion strategy to add external features. We leverage SenticNet 6 [52] to obtain the sentiment label of each sentence and the concatenation is implemented at the classification layer. 'SA(m)+SA(s)' denotes the concatenation of semantic representations from SA(m) and sentiment representations from SA(s). 'SGT(m+s)' denotes our sentiment-guided Transformer. In addition, we also explore whether emotion information is valuable for depression severity detection, thus we fuse semantic information and emotion information using the emotion-guided Transformer module (EGT). We use the pre-trained model j-hartmann-roberta [8] as the post emotion encoder, which is trained on various emotion classification datasets and achieves good performances.

The results are shown in Table 5. We notice that adding sentiment information improves the performance of the model regardless of the fusion strategy used. In particular, our proposed sentiment-guided Transformer performs better than others since the fusion structure can better integrate semantic and sentiment hidden information, while only using label information is not sufficient. We observe that the emotion-guided Transformer also improves the performance on two datasets, by only 0.1% FScore on DsD, and the improvements are both smaller than the sentiment-guided Transformer. A possible reason is that emotion classification is a more difficult task than sentiment classification, resulting in lower accuracy of sentiment classification (66% F1-score on emotion task than 80%-90% F1-score on sentiment task), which will bring the noise to the depression detection.

5.2.3. *contrastive learning ablation*

We also conducted a contrastive learning ablation experiment. 'Base' means our proposed model without a contrastive learning module. '+CL' denotes the base model with supervised contrastive learning. '+severity-aware CL' is the base model with our supervised severity-aware contrastive learning. Table 6 shows the performance comparison of the case study. The results demonstrate that adding supervised contrastive learning improves the detection performance. In detail, on both datasets, the FScore of using severity-aware contrastive learning is 0.3% higher than that of traditional supervised learning, which proves the effectiveness of our designed severity-aware contrastive learning.

[8]https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Table 6
The results of contrastive learning ablation study. We highlight top-1 values in bold.

Model	DsD			DepSign		
	GP	GR	FS	GP	GR	FS
Base	**93.2**	81.8	87.1	86.8	90.3	88.5
+CL	92.4	82.5	87.2	86.8	**90.9**	88.8
+severity-aware CL	93.1	**82.7**	**87.5**	**88.1**	90.2	**89.1**

5.3. Ethical considerations

Although the datasets are publicly available, we also follow strict ethical protocols [53] of sharing datasets. To prevent misuse, all examples shown in this paper have been obfuscated and paraphrased according to the moderate disguise scheme [54]. Our model is not intended to be used as a psychiatric diagnosis tool but an estimate of depression level for users, which can be utilised to direct intervention and treatment.

6. Conclusion

In this paper, we propose a depression severity detection model based on sentiment-guided Transformer and severity-aware contrastive learning. The sentiment-guided Transformer can efficiently fuse semantic and sentiment information from user's posts. Then we use a severity-aware contrastive learning framework, which can fully leverage label information for capturing severity-specific features and helping model distinguish closely-labeled categories. Experimental results show that our model performs better than recent models and achieves new state-of-the-art performance on two publicly available datasets. Further analysis determines the effectiveness of all proposed modules. In future work, we would like to explore other fusion strategies to better integrate various information on the depression severity detection task.

CRediT authorship contribution statement

Tianlin Zhang: Methodology, Software, Writing - original draft. **Kailai Yang:** Writing - Review & Editing. **Sophia Ananiadou:** Supervision, Funding acquisition, Writing - Review & Editing.

Declaration of competing interest

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

Acknowledgments

This work is supported in part by funds from MRC MR/R022461/1 and the Alan Turing Institute and the National Centre for Text Mining.

References

[1] Mark Olfson, Carlos Blanco, and Steven C Marcus. Treatment of adult depression in the united states. *JAMA internal medicine*, 176(10):1482–1491, 2016.

[2] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[3] Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31, 2020.

[4] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11, 2020.

[5] Esteban A Ríssola, David E Losada, and Fabio Crestani. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, 2(2):1–31, 2021.

[6] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893, 2019.

[7] Rodrigo Martínez-Castaño, Juan C Pichel, David E Losada, and Fabio Crestani. A micromodule approach for building real-time systems with python-based models: Application to early risk detection of depression on social media. In *European Conference on Information Retrieval*, pages 801–805. Springer, 2018.

[8] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana 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*, pages 88–97, 2018.

[9] Yu-Tseng Wang, Hen-Hsen Huang, Hsin-Hsi Chen, and H Chen. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)*, 2018.

[10] Abdul Hasib Uddin, Durjoy Bapery, and Abu Shamim Mohammad Arif. Depression analysis from social media data in bangla language using long short term memory (lstm) recurrent neural network technique. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pages 1–4. IEEE, 2019.

[11] Min Yen Wu, Chih-Ya Shen, En Tzu Wang, and Arbee LP Chen. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54(2):225–244, 2020.

[12] David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. Towards preemptive detection of depression and anxiety in twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 82–89, 2020.

[13] Anca Dinu and Andreea-Codrina Moldovan. Automatic detection and classification of mental illnesses from general social media texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 358–366, 2021.

[14] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7184–7190, June 2022.

[15] Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

[16] Steven D Hollon, Michael E Thase, and John C Markowitz. Treatment and prevention of depression. *Psychological Science in the public interest*, 3(2):39–77, 2002.

[17] Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572, 2022.

[18] John Edward Desrochers and Gail M Houck. *Depression in children and adolescents: Guidelines for school practice*. National Association of School Nurses Silver Spring, MD, 2014.

[19] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[20] Marc L Molendijk, Lotte Bamelis, Arnold AP van Emmerik, Arnoud Arntz, Rimke Haringsma, and Philip Spinhoven. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behaviour Research and Therapy*, 48(1):44–51, 2010.

[21] Nirmal Varghese Babu and E Kanaga. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science*, 3(1):1–20, 2022.

[22] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, 2020.

[23] Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):1–13, 2022.

[24] Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6(1):1–12, 2018.

[25] Alina Trifan, Dave Semeraro, Justin Drake, Radek Bukowski, and José Luís Oliveira. Social media mining for postpartum depression prediction. In *Digital Personalized Health and Medicine*, pages 1391–1392. IOS Press, 2020.

[26] Antoine Briand, Hayda Almeida, and Marie-Jean Meurs. Analysis of social media posts for early detection of mental health conditions. In *Canadian Conference on Artificial Intelligence*, pages 133–143. Springer, 2018.

[27] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, Victor Carneiro, et al. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, 21(6):e12554, 2019.

[28] Shreya Ghosh and Tarique Anwar. Depression intensity estimation via social media: a deep learning approach. *IEEE Transactions on Computational Social Systems*, 8(6):1465–1474, 2021.

[29] Ivan Sekulic and Michael Strube. Adapting deep learning methods for mental health prediction on social media. *W-NUT 2019*, 14(162.2):322, 2019.

[30] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.

[31] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer, 2019.

[32] S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 331–338, 2022.

[33] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, 2019.

[34] Cristóbal Colón-Ruiz and Isabel Segura-Bedmar. Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, 110:103539, 2020.

[35] Da Yin, Tao Meng, and Kai-Wei Chang. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, 2020.

[36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[37] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, 2021.

[38] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

[39] Kailai Yang, Tianlin Zhang, and Sophia Ananiadou. A mental state knowledge–aware and contrastive network for early stress and depression detection on social media. *Information Processing & Management*, 59(4):102961, 2022.

[40] Wei-Yao Wang, Yu-Chien Tang, Wei-Wei Du, and Wen-Chih Peng. Nycu_twd@ lt-edi-acl2022: Ensemble models with vader and contrastive learning for detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 136–139, 2022.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[44] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[45] Varsha Suresh and Desmond Ong. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, 2021.

[46] S Kayalvizhi and D Thenmozhi. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*, 2022.

[47] Danielle L Mowery, Craig Bryan, and Mike Conway. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 89–98, 2015.

[48] Mark Zimmerman, Jennifer H Martinez, Diane Young, Iwona Chelminski, and Kristy Dalrymple. Severity classification on the hamilton depression rating scale. *Journal of affective disorders*, 150(2):384–388, 2013.

[49] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525, 2019.

[50] Zogan Hamad, Razzak Imran, Mohammad Shoaib Jameel, and Xu Guandong. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 133–142. ACM (Association for Computing Machinery), 2021.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[52] Erik Cambria, Yang Li, Frank Z Xing, Soujanya Poria, and Kenneth Kwok. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 105–114, 2020.

[53] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102, 2017.

[54] Amy Bruckman. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4(3):217–231, 2002.