Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/43/20240763

Detecting adolescent depression in social media: A hierarchical ensemble learning approach

Tianyu Sheng[1,4,†], **Wenzhen Cai**[1,†], **Junlin Huang**[2], **Zhaosheng Dong**[3]

[1] School of Science, Chang'an University, 710064, China
[1] College of Letters and Science, University of California Santa Barbara, California, 931, United States
[2] Institute of Technology, Wenzhou-Kean University, Wenzhou, 325060, China
[3] Science Faculty, Hong Kong Baptist University, Hong Kong, 999077, Hong Kong SAR, PRC

[4] 2295407485@qq.com
[†] Both authors contributed equally to this paper

Abstract. As digital landscapes evolve, adolescents increasingly rely on social media platforms for self-expression, leading to the vast dissemination of their mental and emotional states. Amidst the physiological and psychological transitions characteristic of adolescence, there is a heightened risk of depressive disorders. These shifts, coupled with the dynamic nature of their online expressions, create a compelling case for detecting latent signs of depression within their digital footprints. This thesis delves into the intersection of Natural Language Processing (NLP) and adolescent depression detection, introducing an innovative hierarchical ensemble model tailored for the intricate task of identifying depressive markers in adolescent social media content. This model amalgamates traditional word embeddings with state-of-the-art pretrained models, encapsulating both sentence-level and word-level representations. Empirical validation, conducted on a unique dataset centered on adolescent depression detection, indicates the model's superior efficacy over existing baselines. By offering an ensemble approach that captures the nuances of adolescent linguistic expressions, this research illuminates the potential for timely and non-intrusive interventions in adolescent mental health.

1. Introduction

In the contemporary information epoch, the ubiquity of social media platforms has entrenched itself in the lives of adolescents[1, 2]. Owing to the pervasive accessibility of smartphones coupled with the escalating internet connectivity, social media platforms have seen a surge in adoption among the youth demographic. Empirical data consistently underscores the dominant role social media plays in the quotidian lives of these adolescents. For instance, a recent study indicates that 97.

Adolescence, as a developmental stage, is punctuated by swift physiological, psychological, and sociological transitions[3]. Amid the crescendo of life stressors, hormonal vicissitudes, and the ongoing quest for self-actualization, this age cohort emerges prominently in the prevalence of depressive disorders[4, 5]. Scholarly research suggests a nexus between depressive sentiments during adolescence and a confluence of hormonal shifts, societal adaptation pressures, and the introspective journey of personal identity formation[6, 7, 8]. Yet, the quintessential rebellious temperament and aspiration for autonomy during this phase often disincentivize adolescents from seeking face-to-face counsel with parental figures or confidantes when navigating emotional quandaries[9, 10]. This reticence is bifurcated into a trepidation of misapprehension and the conviction that peers might offer a more sympathetic ear. In this milieu, social media platforms burgeon as the preferred conduit for emotive articulation[11]. Adolescents curate content that mirrors their emotional landscape, challenges, and mental state, either

DOI: 10.54254/2753-8818/43/20240763

in pursuit of validation or as a cathartic exercise in self-expression. However, the inherent virality of social media ensures that content, especially of a depressive nature, cascades rapidly. An isolated expression of melancholy can inadvertently transmute into an "emotional contagion," echoing within the wider adolescent community and potentially amplifying or even triggering depressive sentiments in its wake[12].

In the contemporary digital landscape, adolescents are increasingly utilizing social media platforms as conduits to articulate their emotions and concerns. Given this phenomenon, the early detection of depressive inclinations could act as a harbinger for prompt interventions, thus forestalling the potential deterioration of mental well-being[13]. Coinciding with the ascendancy of deep learning—particularly advancements in language models—Natural Language Processing (NLP) offers a novel lens through which this challenge can be addressed. Harnessing the capabilities of NLP to discern depressive tendencies from the plethora of adolescents'social media entries presents a scalable, automated approach. Moreover, the examination of online discourse can offer an unobtrusive mechanism to assess an adolescent's psychological well-being. A burgeoning academic inclination is evident within the NLP domain to detect manifestations of depression in social media content.

For illustration, several studies[14, 15, 13] have ventured into ascertaining users displaying depressive symptoms by scrutinizing tweet content, linguistic nuances, and sentiment analytics. Simultaneously, other research endeavors[16, 17] have probed into the correlations between users'digital behaviors—encompassing metrics like posting cadence and interaction patterns—and manifestations of depression. Notwithstanding the depth of these investigations, they predominantly focus on the broad spectrum of internet users, with a conspicuous lacuna in research specifically tailored to the adolescent segment. Given the idiosyncratic and dynamic nature of adolescents'linguistic expressions on social platforms—and recognizing that their emotional trajectories are punctuated by variances intrinsic to their developmental phase—the task of unearthing the latent depressive markers within their digital footprints is indeed intricate.

Addressing this exigent challenge, this paper introduces a pioneering deep learning methodology tailored for the detection of depressive tendencies evident within adolescent social media content. Specifically, motivated by the success of ensemble learning[18, 19, 20], our proposed algorithm employs a hierarchical ensemble model that amalgamates multiple embedding and encoding components. Given a sequence from a social post, our system concurrently embeds it within sentence-level and word-level representational spaces. This is followed by a Siamese architecture, amalgamating three distinct encoders, which processes the aforementioned embedding matrices, resulting in preliminary logits. Subsequent to this, a dynamic weighting mechanism is employed at the secondary ensemble layer to yield the ultimate detection outcomes. Extensive empirical validation, undertaken on a genuine dataset pertaining to adolescent depression detection, attests to the superior performance of our methodology in this domain.

Overall, the contribution of this paper are three-fold:

(i) **Novel Hierarchical Ensemble Approach.** This study introduces a unique hierarchical ensemble model that combines traditional word embeddings (like GloVe) and state-of-the-art pretrained embeddings (like BERT). The two-layer ensemble architecture effectively leverages the strengths of both embedding methods, leading to improved performance in the depression detection task.

(ii) **Comprehensive Evaluation with Strong Baselines.** The proposed method was extensively tested against several strong baselines, including both traditional word embeddings and newer pretrained embeddings. Results consistently indicated the superiority in adolescent depression detection.

(iii) **Practical Implications** By achieving high scores on metrics like Micro F1, Macro F1, and AUC, the proposed model showcases its potential as a reliable tool for online adolescent depression detection. Given the increasing importance of mental health and the prevalence of online communication, such tools can have significant real-world implications.

The remainder of this study is methodically delineated as follows: Section 2 delves into a

DOI: 10.54254/2753-8818/43/20240763

comprehensive review of extant literature focusing on the detection of adolescent depression within the purview of NLP. Section 3 provides an in-depth exposition of our innovative hierarchical ensemble technique. In Section 4, we elucidate the dataset utilized, the experimental framework, and the derived results. Lastly, Section 5 proffers a summative discussion of our study.

2. Literate review

Social media platforms provide a unique avenue to study users'mental health, given that people often express their emotions, opinions, and feelings in this space. Several studies have explored various techniques and methodologies to identify depression tendencies in social media data, particularly from textual content.

2.1. Traditional machine learning approaches

Machine learning has been a cornerstone in this field. Li et al.[21] investigated linguistic features, using algorithms such as Multilayer Perceptron or Logistic Regression, and more, to identify posts with and without depression stigma. Their findings underline the utility of linguistic analysis in stigma detection. Similarly, Al Asad et al.[22] employed Natural Language Processing (NLP) combined with SVM and Na¨ıve Bayes algorithms, demonstrating the potential of these methods to efficiently detect depression from social media. Tadesse et al.[23] used NLP techniques with SVM achieving a high accuracy of 80%.

2.2. Deep learning methods

Deep learning has prompted several studies to investigate its efficacy in depression detection. Yang et al.[24] introduced KC-Net, a knowledgeaware network that utilizes GRU models for mental state detection, achieving promising results. On similar lines, Zogan et al.[25] presented a hierarchical deep-learning network for depression detection. Additionally, Amanat et al.[15] proposed an LSTM model, emphasizing its potential in foreseeing depressive tendencies. The Transformers model by Malviya et al.[26] further showcased the potential of deep learning with a high accuracy rate of 0.98. Cong et al.[27] combined deep learning with XGBoost to handle imbalanced social media data, achieving significant results.

2.3. Multimodal and hybrid approaches

Several studies have begun integrating multiple data types or models for enhanced accuracy. Gui et al.[28] detected depression based on both texts and images, utilizing a novel reinforcement learning model. Chiu et al.[29] and Cheng and Chen[30] also adopted multimodal systems, with the latter focusing on time-aware LSTM for irregular data intervals. The need for models to be both accurate and interpretable has driven research into attention mechanisms and explainability. Zogan et al.[31] introduced a hierarchical attention model for the multiple-aspect depression detection task, which also aimed to explain model predictions.

Adarsh et al.[32] went a step further, presenting an ensemble model combining clustering (KNN) and classification (SVM) methods, introducing label correction with NMT and intrinsic explainability methods to refine results. Similarly, Burdisso et al.[33] proposed SS3, a novel classifier targeting early risk detection challenges.

2.4. Lexical and language-centric approaches

A focus on the language employed on social media platforms has emerged as a vital area. Cha et al.[34] introduced a novel post-classifier using multiple languages and constructed a verified depression lexicon. Guo et al.[35] leveraged a depression lexicon based on domain knowledge for efficient feature extraction, whereas Arag´on et al.[36] proposed the Bag of Sub-Emotions representation for documents.

The endeavor to detect depression hidden in social media data is an emergent and rapidly expanding area of scholarly investigation, encompassing a diverse array of methodologies from foundational

DOI: 10.54254/2753-8818/43/20240763

machine learning techniques to avant-garde deep learning paradigms. Notwithstanding this proliferation, dedicated research centering on the adolescent demographic remains conspicuously scant. Our study represents a pioneering effort in leveraging sophisticated deep-learning techniques to discern depressive tendencies among adolescents within the social media milieu.

3. Methodolgy

In this research, we introduce a hierarchical ensemble learning technique tailored specifically for detecting depressive tendencies within the adolescent cohort. This method is underpinned by a two-pronged embedding schema for textual representation and an ensemble of primary classifiers. The secondlevel classification leverages a weighted voting strategy, as visualized in Fig 1.

For text representation, we employ a duality of embeddings: GloVe-based word embedding[37] and BERT-based sentence embedding[38]. Following this representation phase, the dual vectors are processed by an ensemble classifier. This ensemble classifier integrates three quintessential base learners: the Dense Neural Network (DNN), the Text-specific Convolutional Neural Network (TextCNN)[39], and the Long Short-Term Memory (LSTM) network[40]. Complementing this, a weighting module discerns the significance attributed to each base learner.

Post the inaugural ensemble phase, resultant logit vectors from both the GloVe and BERT embeddings are obtained. A subsequent voting mechanism facilitates the second-level ensemble, culminating in the final classification. A comprehensive schematic representation of the proposed technique is depicted in Fig 1. The ensuing subsections provide a meticulous breakdown of the dual embedding process and the stratified ensemble phases.

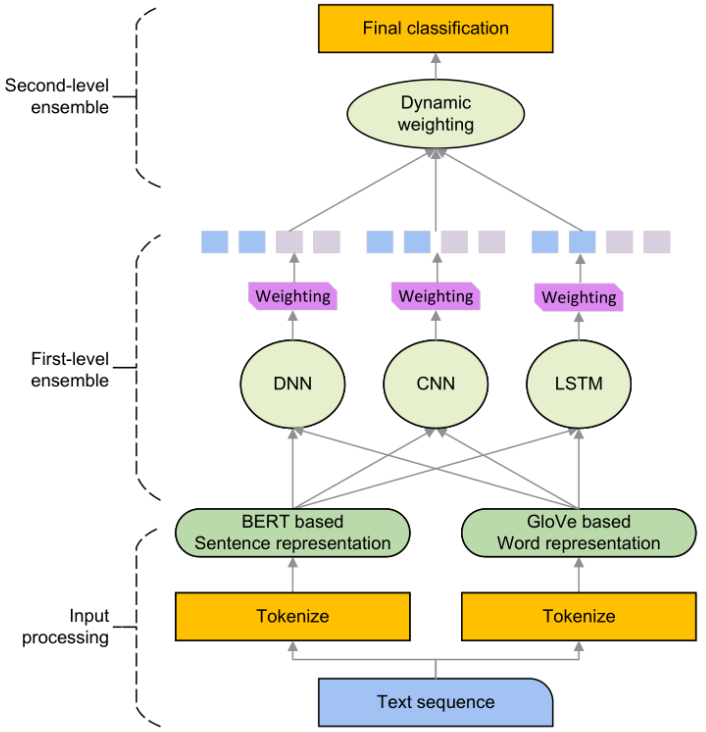


Figure 1: The architecture of the hierarchical ensemble learning technique for detecting depressive tendencies.

DOI: 10.54254/2753-8818/43/20240763

3.1. Dual embedding

Consider a given text T represented as $T = (w_1, w_2, w_3, \ldots, w_x)$. The initial step involves tokenizing this text based on the vocabularies predefined by GloVe and BERT. Following tokenization, the sequences are encoded through their respective embedding layers to generate textual representations. This process can be mathematically articulated as:

$$\mathbf{E}_{\text{glove}} = \text{GloVeEmb}(T) \in \mathbb{R}^{l \times m},$$
$$\mathbf{E}_{\text{bert}} = \text{BertEmb}(T) \in \mathbb{R}^{l \times n},$$

where l denotes the predetermined sequence length, while m and n indicate the dimensions of the respective embeddings. For subsequent formulations and discussions, we will use $\mathbf{E} \in \mathbb{R}^{l \times d}$ as a general representation of the embedding vector.

3.2. First–level ensemble

3.2.1. DNN model The architecture of the DNN module within our proposed model is detailed in Fig 2. Given an input embedding matrix, $\mathbf{E} \in \mathbb{R}^{l \times d}$, this matrix undergoes transformations across several hidden layers, each connected using a fully connected architecture. This transformation can be formally represented as:

$$f_n = \text{FeedForward}(f_{n-1}), \quad \text{where} \quad f_0 = v.$$

The output from the hidden layers yields $f_n \in \mathbb{R}^{l \times n}$. In this configuration, the initial embedding dimension, d, is condensed to n, which corresponds to the quantity of neural units present in the terminal hidden layer.

To manage disparities in sequence lengths and to extract the maximum value from the neural unit's output, f_n is subsequently processed through a global max pooling layer:

$$V_{\text{dnn}} = \text{GlobalMaxPolling}(f_n).$$

Post this operation, the dimensionality of f_n diminishes to $V_{\text{dnn}} \in \mathbb{R}^n$. This results in a one-dimensional vector possessing a consistent length.

3.2.2. TextCNN module The word embedding matrix, $\mathbf{E}$, initially undergoes a cyclic process that includes a CONv_1D layer (one-dimensional convolution layer) followed by a max pooling layer. The CONv 1D is employed to extract semantic features from the sequences with a sliding convolution kernel. The subsequent one–dimensional max pooling aims to reduce the parameter size of the base learner and ensure a consistent-length input for successive stages, given the variable output length from the convolution layer. Filters and kernels are utilized to dictate the convolution layer's output shape, as represented by:

$$\begin{cases} MP_0 = \mathbf{E}, \\ C_n = \text{Conv1d}(MP_{n-1}), \\ MP_n = \text{MaxPolling1d}(C_n). \end{cases}$$

Subsequent to the convolution and pooling operations, we flatten the twodimensional matrix MP_n into a one-dimensional vector. Batch normalization is then applied to this vector to ensure a consistent distribution. Lastly, this normalized vector is processed through feedforward and softmax layers to yield a probability vector. The associated procedures are articulated in the subsequent equations:

$$\begin{cases} V_{\text{normal}} = \text{BatchNormalization}(\text{Flatten}(mp_n)), \\ V_{\text{cnn}} = \text{FeedForward}(V_{\text{normal}}). \end{cases}$$

DOI: 10.54254/2753-8818/43/20240763

3.3. Bi-LSTM model

Both the forward LSTM and the backward LSTM have analogous structures. The hidden vector, $\overrightarrow{h}$, is updated by the current input and the preceding hidden vector. The following equations illustrate how the Bi-LSTM processes:

$$\overrightarrow{h} = \mathrm{LSTM\_forward}(\mathbf{E}),$$
$$\overleftarrow{h} = \mathrm{LSTM\_backward}(\mathbf{E}).$$

Upon completing the sequence processing using the Bi-LSTM, two semantic encoding vectors are derived, $\overrightarrow{h}$ and $\overleftarrow{h}$. These vectors are then concatenated to form the final representation, h:

$$h = \mathrm{concatenate}(\overrightarrow{h}, \overleftarrow{h}).$$

Subsequently, h is processed through a feedforward layer for classification:

$$V_{\mathrm{lstm}} = \mathrm{FeedForward}(h).$$

3.3.1. Weighing module Upon obtaining the three parallel vectors, V_{dnn}, V_{cnn}, and V_{lstm}, they are individually passed through a softmax layer, resulting in three distinct logits vectors. Each of these vectors is then weighted by trainable parameters to gauge the importance of the outputs from each base learner. This can be mathematically represented as:

$$P_* = \mathrm{softmax}(\mathrm{V}_*),$$

where V_* denotes the final output from each base learner. Then, the weighted sum of the logits vectors is calculated as:

$$P_* = \mathbf{W_1} \cdot P_{\mathrm{dnn}} + \mathrm{W}_2 \cdot P_{\mathrm{cnn}} + \mathbf{W_3} \cdot P_{\mathrm{lstm}},$$

Here, P_* represents the logits vector corresponding to each base learner's output.

3.4. Second–level ensemble

After the first-level ensemble, we derive two ensemble learning models utilizing BERT and GloVe embeddings. Using these models, we can swiftly compute the prediction logits for test samples. A further second-level ensemble is then executed, combining these prediction logits to optimize performance. Notably, the second-level ensemble is a non-trainable process.

From the first-level ensemble, we represent the logits of predicted samples acquired from BERT and GloVe embeddings as $\mathcal{P}_B$ and $\mathcal{P}_G$, respectively. Let:

$$\mathcal{P}_B = \{P_B^1, P_B^2, \cdots, P_B^t\},$$
$$\mathcal{P}_G = \{P_G^1, P_G^2, \cdots, P_G^t\}.$$

Given $\mathcal{P}_B$ and $\mathcal{P}_G$, alongside their corresponding true labels $\mathcal{Y}_{\mathrm{true}}$, our objective is to dynamically adjust the weights of these logits to attain optimal classification outcomes.

Initially, we set equal weights to the two models: $w = [0.5, 0.5]$. Given the potential for a large dataset, we divide the logits and $\mathcal{Y} + \mathrm{true}$ into multiple batches, with N signifying the number of these batches. For each batch, the category which has the highest probability score is chosen as the predicted target:

$$\mathrm{pred}_i = \mathrm{argmax}(P_*^i).$$

Subsequently, we can compute the accuracy for the current batches from both models, represented as acc_B and acc_G. The weight adjustment is based on these accuracies:

$$w_1 = (a - \alpha) \times w_1 + \alpha \times \frac{\mathrm{acc}_B}{\mathrm{acc}_B + \mathrm{acc}_G},$$
$$w_2 = 1 - w_1.$$

DOI: 10.54254/2753-8818/43/20240763

Where α stands for the update rate, guiding the pace of the weight update. From these, the final weights to combine the two sets of logits are determined:

$$\text{combined logits} = w_1 \times \mathcal{P}_B + w_2 \times \mathcal{P}_G.$$

With these steps, the second-level ensemble provides dynamically adjusted weights and amalgamated logits for the concluding classification.

4. Experiment

4.1. Dataset

We used an open-access dataset sourced from Kaggle[41]. This dataset, presented in Excel format, encompasses approximately 7,489 records derived from various digital sources, notably social media posts and Facebook comments. A stringent criterion was applied in the selection process to ensure that all posts are from native English speakers, with ages restricted to the 15-17 range. A comprehensive overview of this dataset's metadata can be found in Table 1.

Table 1: Metadata of the Adolescent Depression Detection Dataset

Item	**Description**
Total Samples	7,489
Class Categories	2 (Label 0 for posts indicative of non-depression; Label 1 for posts suggestive of depression)
Class Distribution	6,259 samples with Label 0; 1,227 samples with Label 1
Average Word Count per Sample	70.35
Average Sentence Count per Sample	1.47
Language	English

Table 1 underscores the binary nature of this text classification task, centered around depression detection. However, the stark imbalance in class distribution introduces a layer of complexity, making accurate depression detection more challenging.

4.2. Baselines

To benchmark the efficacy of our proposed hierarchical ensemble method on the depression detection task, we selected several state-of-the-art models renowned in the realm of text classification. These models span both advanced pretrained language representations and potent word-embedding models.

The pretrained models encompass:

(i) **BERT (Bidirectional Encoder Representations from Transformers)**[38]: This model pioneered a seismic shift in the NLP paradigm through its introduction of a bidirectional transformer framework. While it's pretrained on expansive corpora, it is subsequently fine-tuned for distinct tasks.

(ii) **RoBERTa**[42]: As an enhancement of BERT, RoBERTa refines the pretraining process by incorporating prolonged training, magnified batches, and augmented data.

(iii) **XLNet**[43]: A fusion of Transformer-XL and BERT, XLNet's prowess surpasses BERT's by adopting a permutation-centric training approach, eliminating the need for masked tokens.

Word-embedding models include:

DOI: 10.54254/2753-8818/43/20240763

(i) **HAN (Hierarchical Attention Network)**[44]: HAN employs a hierarchical schema reflecting the inherent hierarchical structure of documents. It crafts representations of sentences initially and scales up to the entire document.

(ii) **RCNN (Recurrent Convolutional Neural Network)**[45]: By leveraging RNNs, RCNN captures the contextual nuances from preceding words and subsequently deploys convolutional layers for high-level feature abstraction.

(iii) **TextCNN (Text Convolutional Neural Network)**[39]: Predicated on using convolutional neural networks for text, TextCNN employs diverse filters on localized word clusters (n-grams), pinpointing and harnessing pivotal local textual features.

(iv) **BI-GRU-ATTENTION**: This method applies the attention mechanism to the outputs of a bidirectional GRU. This weighs the significance of individual terms in the sequence, thereby furnishing a weighted sequence representation tailored for classification tasks.

(v) **S2SAN (sentence-to-sentence attention)**[46]: Specifically devised for online user-generated content, this is a sentence-centric attention network. It emphasizes both sentence representation and the intricate extraction of relationships between sentences.

For our research, with the pretrained models, we integrated a Linear layer followed by a softmax layer post the BERT architecture to facilitate classification. On the other hand, for the word-embedding models, we opted for pretrained 300-dimensional vectors to serve as the primary embedding weight.

4.3. Evaluation metrics

The intrinsic nature of the depression detection dataset, characterized by class imbalance, necessitates the utilization of evaluation metrics that are adept at handling imbalances. Accordingly, we adopted Micro-F1, Macro-F1, and AUC scores as our primary evaluation metrics.

- **Micro-F1**: This metric furnishes a holistic view of classification performance. It does so by incorporating TP (True Positives), FP (False Positives), and FN (False Negatives) across all categories. The collective Precision and Recall are then used to compute:

$$\text{Micro-F1} = 2 \times \frac{\text{Aggregate Precision} \times \text{Aggregate Recall}}{\text{Aggregate Precision} + \text{Aggregate Recall}}$$

- **Macro-F1**: Unlike Micro-F1 which considers the aggregate, Macro-F1 computes the F1 score independently for each class and then averages them:

$$\text{F1} = 2 \times \frac{\text{P} \times \text{R}}{\text{P} + \text{R}}$$

where

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

and

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **AUC (Area Under the ROC Curve)**: AUC is especially pertinent in tasks with class imbalances. It quantifies the performance of the classification algorithm to discriminate between the positive and negative classes across varying thresholds. The metric necessitates the calculation of the True Positive Rate (TPR) and False Positive Rate (FPR):

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

DOI: 10.54254/2753-8818/43/20240763

and

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

The AUC score is the integral of the ROC curve plotted using FPR against TPR.

The choice of these metrics ensures that our evaluation not only assesses the overall classification performance but also critically examines the model's proficiency in classifying minority class instances.

4.4. Performance evaluation

For the given dataset, we reserved 70% of the samples for training. The remaining data was evenly divided into validation and test sets, each constituting 15% of the total samples. Given that the dataset isn't exceedingly large, repeating the experiment under varying conditions helps to bolster the reliability of our results. Therefore, to counteract potential overfitting and to enhance the credibility of our results, we ran the experiment five times. For each iteration, we employed a unique random seed to shuffle and split the dataset.

Table 2: Performance on validation set

	exp1			exp2			exp3			exp4			exp5		
	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC
BERT	0.946	0.904	0.957	0.942	0.885	0.953	0.938	0.880	0.948	0.921	0.859	0.928	0.940	0.887	0.933
RoBERTa	0.967	0.939	0.983	0.960	0.921	0.989	0.955	0.916	0.980	0.946	0.904	0.981	0.962	0.928	0.977
XLNet	0.963	0.933	0.986	0.956	0.913	0.991	0.955	0.911	0.983	0.957	0.923	0.984	0.965	0.934	0.982
HAN	0.934	0.872	0.942	0.924	0.856	0.927	0.926	0.851	0.919	0.931	0.862	0.923	0.930	0.867	0.931
RCNN	0.932	0.871	0.929	0.920	0.853	0.929	0.919	0.853	0.945	0.921	0.851	0.921	0.917	0.853	0.933
TextCNN	0.880	0.740	0.833	0.877	0.711	0.827	0.895	0.773	0.845	0.888	0.768	0.852	0.872	0.758	0.831
BI-GRU-ATTENTION	0.921	0.837	0.895	0.906	0.818	0.900	0.913	0.829	0.909	0.921	0.837	0.917	0.907	0.816	0.873
S2SAN	0.927	0.850	0.901	0.918	0.837	0.931	0.932	0.869	0.941	0.926	0.846	0.913	0.921	0.845	0.904
Our method	0.968	0.940	0.989	0.964	0.931	0.991	0.970	0.943	0.992	0.966	0.936	0.992	0.974	0.952	0.989

The overall performance on the validation set, as per Table 2. Our proposed ensemble approach proved its mettle by outshining all other baseline models in the experiments. Achieving a commendable average micro F1 of 0.968 and macro F1 of 0.940, along with an AUC of 0.990, the results validate the strength of the ensemble mechanism, which leverages the complementary nature of various models to produce an enhanced outcome.

Among the baselines, pretrained models, namely RoBERTa and XLNet, demonstrated their prowess in text classification tasks, marking closely behind our ensemble approach. The performance of these models underscores the benefits of large-scale pretraining combined with fine-tuning for specific tasks. RoBERTa and XLNet, having F1 and AUC scores over 0.960, 0.921, and 0.982, respectively, emphasize the evolutionary advancement over the original BERT. This can be attributed to the refined pretraining mechanisms, longer training, and usage of a more significant amount of data. BERT, while not at par with its successors RoBERTa and XLNet, still stands tall against word embedding-based models, which is indicative of the paradigm shift introduced by transformer-based models in the realm of NLP.

Diving into word-embedding models, it's discernible that sophistication in network architecture plays a pivotal role in achieving better results. Hierarchical Attention Network (HAN), Recurrent Convolutional Neural Network (RCNN), and S2SAN—models that incorporate intricate architectures and attention mechanisms—overshadow simpler architectures, underscoring the importance of capturing contextual and hierarchical patterns in text data. TextCNN's relatively underwhelming performance, with average scores of 0.882, 0.750, and 0.838 for Micro F1, Macro F1, and AUC respectively, suggests that simpler convolutional techniques might not be the most adept for the intricacies of depression detection in social media text, which often requires a deeper understanding of context.

The performance on the test set of all five groups of experiments is listed in Table 3. An intriguing observation is the consistency in rankings of model performance across both validation and test sets. The proposed model's exceptional performance, boasting metrics like micro F1 of 0.972, macro F1 of 0.948, and AUC of 0.991, exemplifies its capability. It's evident that the ensemble approach, which combines the strengths of different models, offers a notable advantage.

DOI: 10.54254/2753-8818/43/20240763

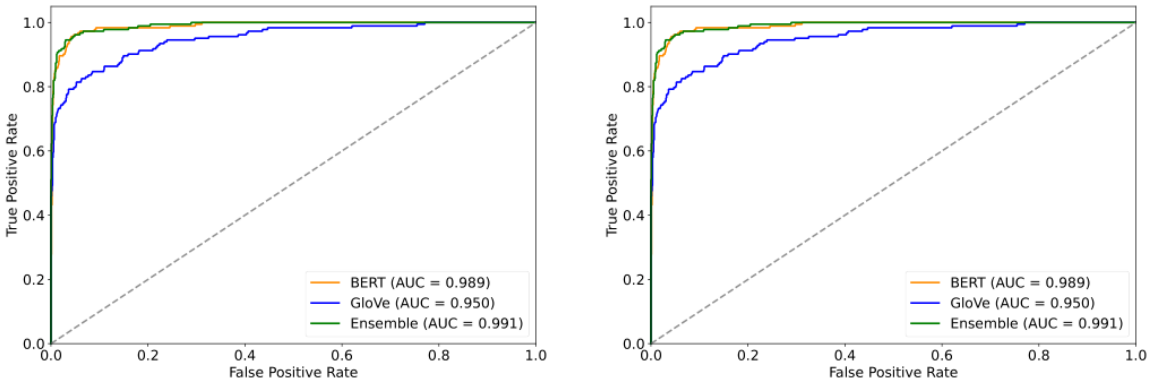
Table 3: Performance on test set

	exp1			exp2			exp3			exp4			exp5		
	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC
BERT	0.943	0.898	0.956	0.939	0.882	0.930	0.954	0.914	0.958	0.927	0.871	0.920	0.950	0.906	0.946
RoBERTa	0.953	0.913	0.976	0.953	0.908	0.964	0.970	0.944	0.987	0.952	0.915	0.985	0.961	0.926	0.979
XLNet	0.957	0.924	0.983	0.959	0.920	0.971	0.969	0.939	0.993	0.959	0.927	0.988	0.964	0.933	0.982
HAN	0.922	0.852	0.933	0.923	0.850	0.916	0.924	0.863	0.940	0.908	0.806	0.901	0.930	0.863	0.935
RCNN	0.926	0.858	0.932	0.934	0.874	0.944	0.918	0.841	0.942	0.916	0.833	0.918	0.935	0.872	0.936
TextCNN	0.878	0.739	0.829	0.880	0.713	0.840	0.866	0.703	0.818	0.870	0.727	0.816	0.878	0.760	0.848
BI-GRU-ATTENTION	0.906	0.809	0.908	0.915	0.836	0.888	0.911	0.831	0.891	0.906	0.799	0.885	0.915	0.832	0.909
S2SAN	0.927	0.865	0.936	0.928	0.862	0.939	0.921	0.849	0.944	0.918	0.845	0.922	0.921	0.854	0.945
Our method	0.971	0.945	0.991	0.971	0.947	0.995	0.974	0.952	0.990	0.967	0.938	0.991	0.979	0.960	0.989

A critical aspect of model evaluation, especially in sensitive domains like depression detection, is the model's robustness. The consistent performance across different experimental iterations, evidenced by low standard deviation values on all metrics (all below 0.01), is an encouraging sign.

4.5. Abation study

4.5.1. Performance before and after second-level ensemble In the previous subsection, we utilized multiple experiments with robust baselines to demonstrate the effectiveness of our proposed hierarchical ensemble methods. To delve deeper into the workings of each module within the hierarchical ensemble method, particularly in the depression detection task, we conducted ablation studies. Initially, we presented both the final performance and the performance before the implementation of the secondlevel ensemble. These results were yielded by models using only the GloVe embedding, as well as the BERT embedding, as inputs. The corresponding ROC curve is visualized, with results displayed in Fig 2 and Table 4. Only the ROC curves from the first group of experiments (exp1) are displayed, as other groups exhibited a similar distribution.



(a) ROC curve of the validation set in exp1

(b) ROC curve of the test set in exp1

Figure 2: Receiver Operating Characteristic

Table 4: Average performance within the second-level ensemble

Model	Micro F1	Macro F1	AUC
BERT embedding only	0.965 ± 0.003	0.935 ± 0.006	0.989 ± 0.004
GLoVe embedding only	0.933 ± 0.005	0.873 ± 0.006	0.950 ± 0.005
Combined and ensembled	0.970 ± 0.004	0.944 ± 0.009	0.991 ± 0.002

DOI: 10.54254/2753-8818/43/20240763

Table 4 lists only the average performance scores and their standard deviations across all ten groups of experiments, using both validation and test sets. The first-level ensemble performance using BERT embedding surpassed that of the GloVe embedding. This observation aligns with previous experimental results. While the GloVe embedding was less effective compared to its counterparts, its output did enhance the overall performance when combined and ensembled with the BERT embedding, with all three metrics exhibiting varying degrees of improvement.

4.5.2. Ablation study within the first-Level ensemble We further conducted the ablation study within the first-level ensemble process. To reveal the importance of each module, we removed the DNN, CNN, LSTM and the weighting module in turn and conducted the depression detection using the same datasets. The results are listed in Table 5.

Table 5: Average performance of within first-level ensemble

	BERT module			GloVe module		
	Micro-F1	macro-F1	AUC	Micro-F1	macro-F1	AUC
full model	0.965 ± 0.003	0.935 ± 0.006	0.989 ± 0.004	0.933 ± 0.005	0.873 ± 0.006	0.950 ± 0.005
DNN removed	0.967 ± 0.004	0.939 ± 0.008	0.988 ± 0.005	0.934 ± 0.005	0.866 ± 0.012	0.946 ± 0.005
CNN removed	0.958 ± 0.007	0.921 ± 0.013	0.980 ± 0.007	0.931 ± 0.006	0.867 ± 0.011	0.945 ± 0.005
LSTM removed	0.96 ± 0.005	0.926 ± 0.009	0.984 ± 0.006	0.930 ± 0.004	0.867 ± 0.008	0.945 ± 0.005
weighting module removed	0.965 ± 0.005	0.935 ± 0.011	0.986 ± 0.006	0.927 ± 0.005	0.862 ± 0.01	0.945 ± 0.005

The analysis of the ablation study in the context of the first-level ensemble elucidates the critical importance of each individual module in the depression detection task. Through these results, we get an understanding of how the ensemble's constituents—DNN, CNN, LSTM, and the weighting module—interplay with one another and the resulting impact on performance. The removal of the DNN module resulted in only a minuscule effect on the model's performance. Intriguingly, the Micro-F1 scores observed a minor increase for both BERT and GloVe modules. This is indicative of the limited or potentially redundant contribution the DNN module offers in the ensemble, suggesting its role is either being overshadowed by more effective architectures or is not essential for this specific task. Both CNN and LSTM are known to be effective for capturing spatial and temporal patterns in the data, respectively. The fact that their removal impacts the performance of the BERT module significantly suggests they are paramount in grasping the intricate contextual patterns in BERT embeddings. On the other hand, the GloVe module's lesser impact when LSTM or CNN is removed might indicate that traditional embeddings like GloVe do not benefit as much from these architectures, at least in the context of depression detection. The weighting module, designed to assign significance to various features, displays a differential impact on BERT and GloVe. While its removal hardly affects the BERT module, it significantly impacts GloVe's performance.

4.6. Discussion

The hierarchical ensemble approach adopted in this study showcases the merits of combining pretrained and word embedding methods for the critical problem of depression detection. By consistently outperforming the baseline models, it emphasizes the potential of layered ensemble approaches in capturing intricate patterns. The near parity in performance between models like RoBERTa, XLNet, and our proposed method underscores the potency of pretrained models in such tasks. Their innate capability, due to extensive pretraining on vast corpora, makes them formidable tools for depression detection. While word embeddings like GloVe might not exhibit strength independently, their value in augmenting performance, when combined with stronger embeddings like BERT, cannot be overlooked. The uplift in performance with their inclusion underlines the benefits of diverse feature representation.

The ablation study yields several key takeaways. The DNN's limited influence suggests potential redundancy in this setup. The BERT architecture's pronounced sensitivity to the removal of CNN and

DOI: 10.54254/2753-8818/43/20240763

LSTM highlights their pivotal role in this context. The differential impact of the weighting module on BERT and GloVe further cements the notion that while some embeddings might be self-sufficient, traditional ones can still benefit from added modulation.

In summation, the hierarchical ensemble method underscores the efficacy of diverse feature representations and ensemble strategies. The research points towards a future where collaborative ensemble methods are the norm in tasks like depression detection.

5. Conculsion

The research presented a pioneering hierarchical ensemble approach for detecting adolescent depression. Rigorous experiments and ablation studies showcased its superiority over contemporary baselines, especially in its adeptness at leveraging diverse embeddings. Theoretically, the study emphasizes the *synergistic power* of merging different word representations. While pretrained embeddings consistently yield robust performance owing to their comprehensive training, traditional embeddings can augment the feature space. The detailed findings from the ablation study highlight the *differential impact* of various network architectures depending on the nature of the embeddings. These insights are invaluable for the design of subsequent ensemble models. On a practical note, online depression detection is crucial for proactive interventions. Given its high F1 and AUC scores, the introduced model holds promise for integration into online platforms, enabling early depression detection. This hierarchical ensemble model serves as a paradigm for other NLP endeavors where the *confluence* of embeddings can be advantageous.

However, it's pertinent to mention that this study was grounded on a specific dataset endowed with human-annotated depression labels. Consequently, the performance of the model might exhibit variation when tested on more heterogeneous datasets. The ostensibly redundant role of the DNN module prompts concerns about the model's efficiency and potential overelaboration. Probing the model's prowess in other mental health realms can bolster its versatility. Streamlining the model, by assessing the indispensability of individual modules, could enhance its efficiency without curtailing its performance.

Acknowledgement

Tianyu Sheng and Wenzhen Cai contributed equally to this work and should be considered co-first authors.

References

[1] G.S.O'Keeffe, K.Clarke-Pearson, et al., The impact of social media on children, adolescents, and families, Pediatrics 127 (2011) 800–804.

[2] N.Caner, Y.S.Efe, ¨O.Ba¸sda¸s, The contribution of social media addic- tion to adolescent life: Social appearance anxiety, Current Psychology 41 (2022) 8424–8433.

[3] J.M.Griffith, H.M.Clark, D.A.Haraden, J.F.Young, B.L.Hankin, Affective development from middle childhood to late adolescence: Tra- jectories of mean-level change in negative and positive affect, Journal of youth and adolescence 50 (2021) 1550–1563.

[4] S.Nolen-Hoeksema, J.S.Girgus, The emergence of gender differences in depression during adolescence., Psychological bulletin 115 (1994) 424.

[5] H.M.Boynton, The healthy group: A mind–body–spirit approach for treating anxiety and depression in youth, Journal of Religion & Spiri- tuality in Social Work: Social Thought 33 (2014) 236–253.

[6] G.D.Rosenblum, M.Lewis, Emotional development in adolescence, Blackwell handbook of adolescence (2006) 269–289.

[7] T.Brezina, Adapting to strain: An examination of delinquent coping responses, Criminology 34 (1996) 39–60.

[8] L.Raemen, L.Claes, M.Verschueren, L.Van Oudenhove, S.Vandek- erkhof, I.Triangle, K.Luyckx, Personal identity, somatic symptoms, and symptom-related thoughts, feelings, and behaviors: Exploring asso- ciations and mechanisms in adolescents and emerging adults, Self and identity 22 (2023) 155–180.

[9] J.Wolak, D.Finkelhor, K.Mitchell, Internet-initiated sex crimes against minors: Implications for prevention based on findings from a national study, Journal of adolescent health 35 (2004) 424–e11.

[10] H.Williamson, D.Harcourt, E.Halliwell, H.Frith, M.Wallace, Ado- lescents' and parents' experiences of managing the psychosocial impact of appearance change during cancer treatment, Journal of Pediatric Oncology Nursing 27 (2010) 168–175.

DOI: 10.54254/2753-8818/43/20240763

[11] S.F.Waterloo, S.E.Baumgartner, J.Peter, P.M.Valkenburg, Norms of online expressions of emotion: Comparing Facebook, Twitter, Insta- gram, and Whatsapp, New media & society 20 (2018) 1813–1831.
[12] J.Nesi, S.Choukas-Bradley, M.J.Prinstein, Transformation of adoles- cent peer relations in the social media context: Part 2—application to peer group processes and future directions for research, Clinical child and family psychology review 21 (2018) 295–319.
[13] F.M.Shah, F.Ahmed, S.K.S.Joy, S.Ahmed, S.Sadek, R.Shil, M.H.Kabir, Early depression detection from social network using deep learning techniques, in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 823–826.
[14] A.H.Orabi, P.Buddhitha, M.H.Orabi, 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, 2018, pp. 88–97.
[15] A.Amanat, M.Rizwan, A.R.Javed, M.Abdelhaq, R.Alsaqour, S.Pandya, M.Uddin, Deep learning for depression detection from tex- tual data, Electronics 11 (2022) 676.
[16] M.Y.Wu, C.-Y.Shen, E.T.Wang, A.L.Chen, A deep architecture for depression detection using posting, behavior, and living environment data, Journal of Intelligent Information Systems 54 (2020) 225–244.
[17] S.Choudhary, N.Thomas, J.Ellenberger, G.Srinivasan, R.Cohen, et al., A machine learning approach for detecting digital behavioral pat- terns of depression using nonintrusive smartphone data (complementary path to patient health questionnaire-9 assessment): Prospective obser- vational study, JMIR Formative Research 6 (2022) e37736.
[18] D.Liang, B.Yi, Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification, Information Sciences 547 (2021) 271–288.
[19] A.Mohammed, R.Kora, An effective ensemble deep learning framework for text classification, Journal of King Saud University-Computer and Information Sciences 34 (2022) 8825–8837.
[20] J.Hou, P.Wang, Assemble the shallow or integrate a deep? toward a lightweight solution for glyph-aware Chinese text classification, Plos one 18 (2023) e0289204.
[21] A.Li, D.Jiao, T.Zhu, Detecting depression stigma on social media: A linguistic analysis, Journal of affective disorders 232 (2018) 358–362.
[22] N.Al Asad, M.A.M.Pranto, S.Afreen, M.M.Islam, Depression detec- tion by analyzing social media posts of user, in: 2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON), IEEE, 2019, pp. 13–17.
[23] M.M.Tadesse, H.Lin, B.Xu, L.Yang, Detection of depression-related posts in reddit social media forum, Ieee Access 7 (2019) 44883–44893.
[24] K.Yang, T.Zhang, S.Ananiadou, A mental state knowledge–aware and contrastive network for early stress and depression detection on social media, Information Processing & Management 59 (2022) 102961.
[25] H.Zogan, I.Razzak, S.Jameel, G.Xu, Depressionnet: learning multi- modalities with user post summarization for depression detection on social media, in: proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 133–142.
[26] K.Malviya, B.Roy, S.Saritha, A transformers approach to detect de- pression in social media, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 2021, pp. 718–723.
[27] Q.Cong, Z.Feng, F.Li, Y.Xiang, G.Rao, C.Tao, XA-BiLSTM: a deep learning approach for depression detection in imbalanced data, in: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 2018, pp. 1624–1627.
[28] T.Gui, L.Zhu, Q.Zhang, M.Peng, X.Zhou, K.Ding, Z.Chen, Co- operative multimodal approach to depression detection in twitter, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 110–117.
[29] C.Y.Chiu, H.Y.Lane, J.L.Koh, A.L.Chen, Multimodal depression detection on Instagram considering time interval of posts, Journal of Intelligent Information Systems 56 (2021) 25–47.
[30] J.C.Cheng, A.L.Chen, Multimodal time-aware attention networks for depression detection, Journal of Intelligent Information Systems 59 (2022) 319–339.
[31] H.Zogan, I.Razzak, X.Wang, S.Jameel, G.Xu, Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media, World Wide Web 25 (2022) 281–304.
[32] V.Adarsh, P.A.Kumar, V.Lavanya, G.Gangadharan, Fair and ex- plainable depression detection in social media, Information Processing & Management 60 (2023) 103168.
[33] S.G.Burdisso, M.Errecalde, M.Montes-y G´omez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.
[34] J.Cha, S.Kim, E.Park, A lexicon-based approach to examine de- pression detection in social media: the case of Twitter and university community, Humanities and Social Sciences Communications 9 (2022) 1–10.
[35] Z.Guo, N.Ding, M.Zhai, Z.Zhang, Z.Li, Leveraging domain knowl- edge to improve depression detection on Chinese social media, IEEE Transactions on Computational Social Systems (2023).
[36] M.E.Arag´on, A.P.L.Monroy, L.C.Gonz´alez-Gurrola, M.Montes, Detecting depression in social media using fine-grained emotions, 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), 2019, pp. 1481–1486.
[37] J.Pennington, R.Socher, C.D.Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014

DOI: 10.54254/2753-8818/43/20240763

conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[38] J.D.M.-W.C.Kenton, L.K.Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[39] Y.Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguis- tics, Doha, Qatar, 2014, pp. 1746–1751.
[40] S.Hochreiter, J.Schmidhuber, Long short-term memory, Neural com- putation 9 (1997) 1735–1780.
[41] Kaggle, Student-depression-text, 2023.URL: https://www.kaggle. com/datasets/nidhiy07/student-depression-text, accessed: 2023- 08-21.
[42] Y.Liu, M.Ott, N.Goyal, J.Du, M.Joshi, D.Chen, O.Levy, M.Lewis, L.Zettlemoyer, V.Stoyanov, RoBERTa: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907. 11692 (2019).
[43] Z.Yang, Z.Dai, Y.Yang, J.Carbonell, R.R.Salakhutdinov, Q.V. Le, XLNet: Generalized autoregressive pretraining for language under- standing, Advances in neural information processing systems 32 (2019).
[44] Z.Yang, D.Yang, C.Dyer, X.He, A.Smola, E.Hovy, Hierarchi- cal 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, 2016, pp. 1480–1489.
[45] S.Lai, L.Xu, K.Liu, J.Zhao, Recurrent convolutional neural net- works for text classification, in: Proceedings of the AAAI conference on artificial intelligence, volume 29, 2015.
[46] P.Wang, J.Li, J.Hou, S2SAN: A sentence-to-sentence attention net- work for sentiment analysis of online reviews, Decision Support Systems 149 (2021) 113603.
[47] IOP Publishing is to grateful Mark A Caprio, Center for Theoretical Physics, Yale University, for permission to include the `iopart-num` BIBTEXpackage (version 2.0, December 21, 2006) with this documentation. Updates and new releases of `iopart-num` can be found on `www.ctan.org` (CTAN).