Ensemble Hybrid Learning Methods for Automated Depression Detection

Luna Ansari, Shaoxiong Ji, Qian Chen, and Erik Cambria, *Fellow, IEEE*

***Abstract*— Changes in human lifestyle have led to an increase in the number of people suffering from depression over the past century. Although in recent years, rates of diagnosing mental illness have improved, many cases remain undetected. Automated detection methods can help identify depressed or individuals at risk. An understanding of depression detection requires effective feature representation and analysis of language use. In this article, text classifiers are trained for depression detection. The key objective is to improve depression detection performance by examining and comparing two sets of methods: hybrid and ensemble. The results show that ensemble models outperform the hybrid model classification results. The strength and effectiveness of the combined features demonstrate that better performance can be achieved by multiple feature combinations and proper feature selection.**

***Index Terms*— Deep neural networks, depression detection, ensemble methods, sentiment lexicon.**

I. Introduction

DRASTIC changes in the human lifestyle in modern society have led to an increase in the number of people suffering from depression. Depression is known to be "a disease of modernity" [1], and it has been predicted that by 2030, one of the three causes of illness will be depression [2]. The social stigma surrounding depression and the high rate of misdiagnosis has led to a lack of access to proper diagnosis and care [3]. Serve mental disorders but without effective intervention could turn to suicidal ideation [4]. Therefore, the timely detection of depression can be highly beneficial for individuals and society.

Depression symptoms can be reflected in various human activities and behaviors and different degrees [5]. One of the sources, which can help identify depression symptoms in individuals, is the use of language [6]. Cognitive and linguistic studies have numerously shown that people with depression use language features differently [6]. For example, they tend to use more first-person singular pronouns (I, me, or we) and more negatively valenced words [7].

Online social content is one source for automatic mental disorder detection as it is one of the platforms through which users communicate. In recent years, social networking platforms have been widely applied to study users' behavior and have inspired various researchers to introduce new forms of health care solutions [8], [9]. Furthermore, the stigma surrounding depression can make individuals less willing to seek professional assistance, and they turn to less traditional sources such as social media. Social media can be an essential source of information about individuals' opinions and feelings in the study of depression [10]. More specifically, research has addressed depression detection at various levels of granularity and approached it from different standpoints. Several social network sites (SNSs), such as Reddit, Twitter, Facebook, and Weibo, have been utilized for research about depression and other mental state disorders, such as postpartum depression [11] and posttraumatic stress disorder (PTSD) [12].

Earlier approaches to depression detection have primarily taken a bottom-up approach to learn and apply deep learning (DL) and machine learning (ML) methods. While such subsymbolic artificial intelligence (AI) methods can provide valuable insights about word frequencies and statistical correlations, they are not sufficient to analyze narrative and understanding of dialog systems in sentiment analysis [13]. Although there has been advancement with natural language processing (NLP) methods using DL methods, the predictive power of such approaches is limited mainly because DL methods learn better from large sets of data. Besides, communication entails a broader range of contributors, including understanding the world, social norms, and cultural awareness.

To address these challenges, recently, research in depression detection has taken top-down approaches to learn by applying symbolic AI methods such as logical reasoning. In particular, the hybrid combination of subsymbolic approaches with symbolic methods has been shown to induce more meaningful patterns in natural language texts [13]. Hence, it is vital to integrate symbolic approaches to learning with subsymbolic approaches in tackling the task of automated depression detection.

In addition to hybrid methods, another set of approaches that yields high accuracy are ensemble methods in which several learning methods are combined [14]. Ensemble methods have frequently achieved high performance in solving various predictive problem areas [14].

The current study builds on these recent advancements to extend existing knowledge on automated depression detection. Our study contributes to the literature by improving the performance of depression detection as a text classification

Manuscript received 26 October 2021; revised 21 January 2022; accepted 10 February 2022. Date of publication 14 March 2022; date of current version 31 January 2023. *(Corresponding author: Shaoxiong Ji.)*

Luna Ansari and Shaoxiong Ji are with the Department of Computer Science, Aalto University, 02150 Espoo, Finland (e-mail: luna.ansari@aalto.fi; shaoxiong.ji@aalto.fi).

Qian Chen and Erik Cambria are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: chen1028@e.ntu.edu.sg; cambria@ntu.edu.sg).

Digital Object Identifier 10.1109/TCSS.2022.3154442

task. In principle, this is an experimental study to show how the hybrid methods (symbolic and subsymbolic) and ensemble methods can improve performance for depression detection. For this purpose, eight folds of experiments on each of three separate datasets are conducted. The experiments with hybrid methods utilize different sentiment lexicons and apply logistic regression (LR) for text classification. The experiments with the ensemble methods are conducted by combining DL approaches and lexicon-based models. In this set of experiments, the DL methods include two recurrent neural networks (RNNs), i.e., long short-term memory (LSTM) and AttentionLSTM. By doing so, this study contributes to the literature on automated depression detection by applying hybrid methods.

This article is structured as follows. Section II reviews existing research on NLP techniques for text classification and work related to automated depression detection. The learning methods are discussed in Section III. In Section IV, the conducted experiments, an overview of the data, and exploratory data analysis are described and achieved results for classification are presented. Finally, a comparison between these models, concluding remarks, limitations, and further research directions is discussed.

II. Related Work

Research has found valuable insights into the evaluation of textual data and various NLP techniques to capture users' linguistic tendencies. The automated examination of the relationship between language and mental states has been accomplished by building classifier models using a range of feature extraction methods.

While some studies have focused on statistical approaches on the extraction of individual features such as N-grams [15], linguistic inquiry and word count (LIWC) [16], and bag of words (BOWs) [17], [18], other studies have tried to compare the effect of single features with different ML approaches. Recent research has explored how the combination of individual features can improve classification accuracy. These studies build on approaches such as term frequency-inverse document frequency (TF-IDF) + linear discriminant analysis (LDA) [19] TF-IDF and N-gram + LIWC [20]. In a recent review, Calvo *et al.* [21] provided a taxonomy of the NLP techniques and computational methods to detect various mental health issues.

Recent deep neural networks have also been applied to depression detection and mental healthcare. For example, Orabi *et al.* [22] applied word embeddings to improve the performance of two subtasks: model generalization capability on a Bell Lets Talk depression detection on the CLPsych2015 dataset. They compared the performance of convolutional neural network (CNN)-based models and RNN-based models and found that CNN-based models perform better. CNN-based models combined with optimized embeddings had higher generalization power [22]. Benton *et al.* [15] investigated the effectiveness of multitask learning (MTL) models on a small dataset. The authors predicted a set of conditions to predict mental states by combining feedforward multilayer perceptron and MTL. In another study, Nguyen *et al.* [23] transformed text into high-dimensional space and applied topic modeling to derive topics and moods of texts from the LiveJournal social networking service. In addition, Maupomé and Meurs [24] combined an unsupervised topic extraction algorithm with a multilayer perceptron by utilizing unigram, bigram, and trigram frequency to extract 30 topics. They found that a limited number of data points hindered neural networks [24].

One of the main challenges when designing natural language analysis is identifying the relevant set of features. Various methods have been applied to extract a relevant set of features within text classification literature. These methods include "BOWs," "bag of phrases," "bag of n-grams," "WordNet-based word generalizations," and "word embedding" techniques [25]. In a recent review study, Misra [25] presented an overview of language feature specification for ML- and DL-based text analysis and highlighted that such approaches have made it possible to reuse feature specification in semantically similar contexts [25]. In this study, Misra [25] considered different levels of analysis, including words, phrases, sentences, paragraphs, documents, and corpus level, which can accordingly be utilized to extract linguistic, semantic, and statistical types of features from textual data. Specifically, four main approaches to identifying features in the text classification context are discussed [25]. The first features can be extracted at the document level by applying lexico-syntactic patterns. These patterns can be captured as part of speech (POS) tag patterns, i.e., when a sequence of words matches a specific expression. Second, semantic similarity and relatedness-based features are the other set of features extracted at phrase, sentence, or document level. These features can be extracted by utilizing vector space modeling, topic modeling, neural embedding, and latent semantic analysis, which can be conducted. Semantic features can also be extracted based on ontological relationships between concepts across various text corpora. The fourth approach to feature extraction uses statistical analysis and feature engineering to find statistical features of texts.

More specifically, research on mental state detection techniques has built on various features varying from linguistic cues and statistical features to user posting patterns to build classification models. Such features include stress-related use of language, timing and frequency of posts, the sentiment of posts, and value contrast, i.e., the polarity of posts shifting between positive and negative sentiments. For example, Stankevich *et al.* [26] applied bag-of-words and word embeddings and found that while embedding features obtained higher recall scores, the TF-IDF model with morphological features obtained higher performance of 63% F1-score and higher accuracy and precision scores than embedding features. Shen *et al.* [27] extracted six depression-related linguistic features describing online social behavior and clinical depression criteria. A labeled Twitter dataset provided a multimodal depressive dictionary learning model that reached an 85% F1-score. In another setting, Tsugawa *et al.* [28] conducted an online questionnaire to measure the degree of Twitter users' depression built up multiple features and conducted topic modeling to predict users' mental states by looking into the history of the users' online activity.

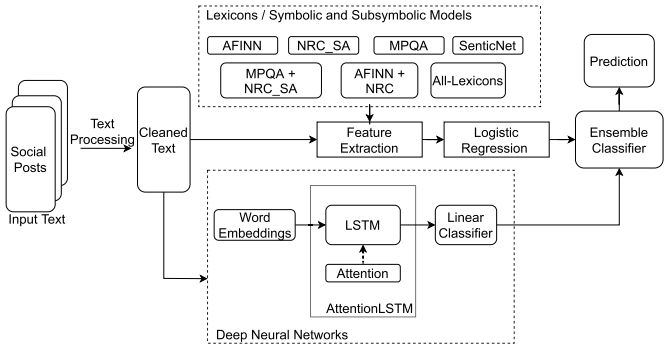


Fig. 1. Flowchart of the proposed ensemble model.

Although these studies have provided valuable insights into the utilization of textual data in mental state detection, we know less about how the combination of symbolic and subsymbolic AI can enhance sentiment classification tasks in this context.

III. Methods

NLP technologies are widely applied to solve various problems across various domains, such as text summarization, language translation, and sentiment analysis. Earlier attempts to solve these problems relied on rule-based methods and probabilistic methods, such as the hidden Markov model [29], which required much data engineering. More recently, NLP methods have relied much more on DL [30], [31]. With the rise of powerful computing systems, it has become possible to train end-to-end systems as ML helps address various problems from fields, such as image recognition [32], speech recognition [33], and NLP [34].

Early approaches to text classification relied on representing documents through a bag-of-words representation and applying ML methods in which the words were not processed sequentially [31]. In recent years, text classification has been conducted by considering the sequential nature of data using LSTM [35] neural networks, as these models can require fewer training samples due to their reliance on word embedding.

As methods, such as RNNs, LSTM, and attention-based models, have transformed speech and NLP, this study draws from these models to analyze and classify texts by detecting the fragments containing sentiments. The flowchart of the proposed ensemble model is shown in Fig. 1.

A. LSTM Networks

RNNs is a class of neural network architectures where the output of the network for the current token in a sequence depends on the output of hidden state for the previous tokens [34]. LSTM is a popular model for text classification problems in the NLP field. LSTM and gated cells LSTM belong to a special kind of RNN. Despite RNN's theoretical ability to retain information in time, it has difficulty in handling "long-term dependencies" in practice [34]. LSTM cells have the same chain-like network structure, but each unit contains four different subnetworks—the cell state, input gate, forget gate, and output gate. The forget gate and input gate work as a filter to determine which information to remove from the cell state and which information to add to it. The output gate determines how the cell state and current token should be combined to produce the output, which is the output to the sigmoid function [34].

B. Attention Mechanism

The attention mechanism works by considering a part of the sentence and allows the learning algorithm to focus more on certain parts of the input while focusing less on the rest of the input [36]. It considers all subtexts and contexts as its input and outputs the weighted arithmetic mean of these subtexts. Attention is the primary building block of the transformer model, a new family of neural network architecture. The transformer-based model is a model where the transformer is a block of self-attention combined with feedforward networks. It consists of an encoder responsible for encoding the input text and passing it to a decoder that produces the output [37]. The transformer model takes in a sentence and its context as input. The context is the string that contains the preceding sentence, the sentence itself, and the following sentence. The attention unit looks at all parts of the sentence and drives similar parts. Retrieving such similarity makes it possible to attend to longer steps behind a sequence. The attention mechanism can prioritize the relationship between very distant items/words in a sequence to detect relationships between very close words.

C. Sentiment Lexicon: Hybrid Learning Approach

One of the leading sentiment lexicons used in this study to conduct the experiments is the SenticNet sentiment lexicon. This lexicon acts as the source of sentiment information. It represents every sentence in training set in terms of several features, including polarity value, polarity label, sensitivity, introspection, temper, and attitude. For each word in a sentence, SenticNet allocates a specific value regarding these features. The features of SenticNet are engineered and driven based on a hybrid ML (HML) approach. By definition, an HML model is a combination of two or more techniques. Specifically, the first component performs the data preprocessing task, and the second constructs the classification or prediction model based on the output from the first component [38].

The SenticNet sentiment lexicon integrates the symbolic and subsymbolic AI tools [13]. In particular, it employs DL as subsymbolic AI for recognizing patterns in texts, and it uses logical reasoning as symbolic AI for deconstructing multiword expressions into primitives [13]. While the subsymbolic AI takes a bottom-up learning approach to NLP, the symbolic AI processes the information in a top-down manner. The ensemble combination of DL (subsymbolic) and logical reasoning (symbolic) tools is superior to any of the approaches being conducted alone [13].

D. Logistic Regression

LR is a supervised ML method for training binary classifiers. The predictor h is based on a linear map $h(x) = \mathbf{w}^T\mathbf{x}$, for which the training process finds the optimal weight vector $\mathbf{w}$.

TABLE I
SUMMARY OF DATASETS

Dataset	# of Positive Samples	# of Negative Samples
CLPsych	327+150	246+150
Reddit	1,200	641
eRisk	770	3,728

The predicted label $\hat{y}_i \in \{-1, 1\}$ is decided by the rule $\hat{y}_i = 1$ if $h(\mathbf{x}) \geq 0$ and $\hat{y}_i = -1$ if $h(x) < 0$. Loss is measured by the function $\epsilon(w \mid X) = (1/N) \sum_{i=1}^{N} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}))$. The optimal weight vector can be found using, e.g., gradient descent.

E. Ensemble Methods

Ensemble methods are another approach adopted in this study. Such approaches have successfully set record performance on challenging datasets and are frequently among the top winners of data science competitions. The purpose of using the ensemble approach can vary, and it is mainly applied for three purposes: decreasing bias in the model (boosting), decreasing variance in the model (bagging), and enhancing predictions (stacking) [39].

In the current study, the ensemble method is used for bagging to combine the predictive results. More specifically, classification results from LSTM and LR models are averaged. Bagging—also known as bootstrapping—is a strategy to decrease estimation variance of single models by averaging the multiple estimates together [39]. Due to parallel learning, the classifiers are trained separately so that each model emphasizes different features. The combination of the two baseline estimators, i.e., LSTM and LR, can achieve higher accuracy, can be less sensitive to variance in training dataset, and can reduce the variance overall [14].

IV. EXPERIMENTS

Separate experiments were performed for the tasks of depression detection using three public datasets. The PyTorch framework[1] and Scikit-learn library[2] were used.

A. Datasets

Table I summarizes the datasets used in the experiments of this paper.

1) CLPsych 2015 Shared Task: The Computational Linguistics and Clinical Psychology (CLPsych) was initiated in 2014 to promote collaboration between psychologists and computer scientists [40]. Specifically, "shared tasks" were defined to study and compare different methods on the same prediction problems. This dataset contains user-generated posts from users with depression or (Depression and PTSD on Twitter) PTSD on Twitter.[3] Specifically, there are three binary classification subtasks, i.e., 1) depression versus control; 2) PTSD versus control; and 3) depression versus PTSD. The train partition consisted of 327 depression users, 246 PTSD users, and for each an age- and gender-matched control user,

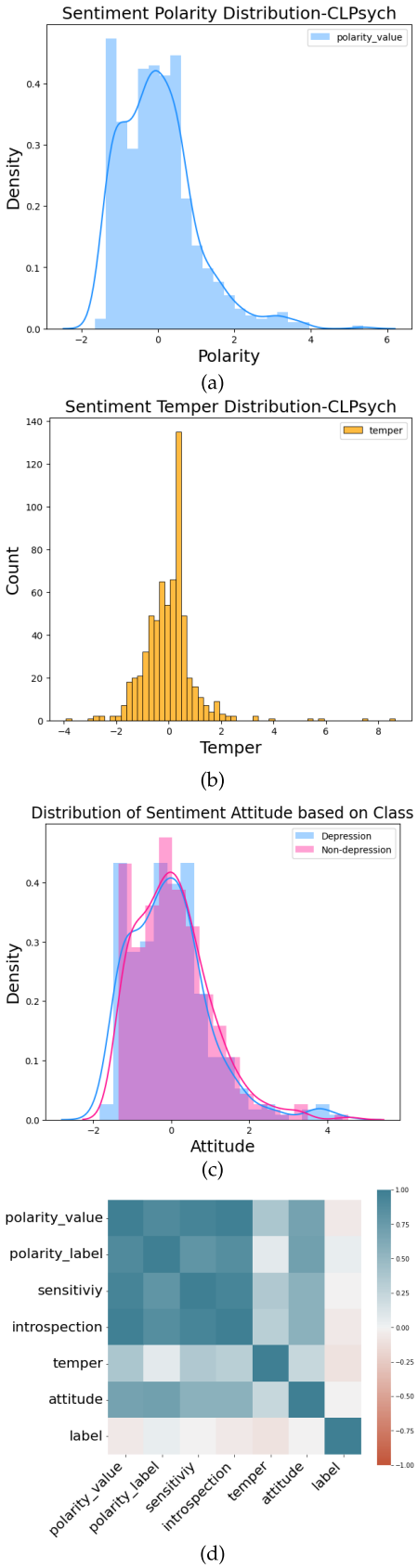


Fig. 2. Sentiment feature visualization in the CLPsych dataset. (a) Sentiment polarity value distribution. (b) Sentiment temper distribution. (c) Distribution of attitude. (d) Correlation heatmap of SenticNet features.

[1]https://pytorch.org
[2]https://scikit-learn.org/
[3]This dataset is available by request via http://www.cs.jhu.edu/~mdredze/datasets/clpsych_shared_task_2015/

for a total of 1146 users. The test data contained 150 depression users, 150 PTSD users, and age- and gender-matched control for each, for a total of 600 users. However, the actual number of users in the training and testing sets is 1711 due to an unknown data missing issue.

2) Reddit: Reddit social media collection contains posts from depressed and nondepressed users. The dataset contained 1841 users (1200 positives and 641 negatives) [41]. Reddit as a social media platform allows for the anonymity of the users, and it is widely used for discussion about stigmatic topics [8]. Reddit data have been used to study the posts specifically from Reddit users who wrote about mental health issues and who had proceeded to post topics about suicidal ideation [42]. The data were concatenated, randomly shuffled, and split into train and test sets with an 80:20 split rate. The final data frame consisted of one column of text comments and another column of labels for the corresponding comments. Each comment was labeled with 1 or 0 for depression or nondepression, respectively.

3) eRisk Dataset: This dataset is collected from the eRisk (Early Risk Prediction) forum [43]. The eRisk is a public competition platform that facilitates multidisciplinary research and creates reusable datasets and benchmarks for assessing early risk detection technologies in health and safety problem areas. The eRisk 2018 dataset was initially developed to detect early signs of depression. The eRisk collection contains posts from depressed and nondepressed 4498 users, where 3728 users belong to the nondepressed and 770 belong to the depressed class. The data were concatenated, randomly shuffled, and split into train and test sets with an 80:20 split rate.

B. Data Preprocessing

In this study, NLP tools are applied to preprocess the datasets before proceeding to the training step. First, tokenization is used to split the posts into individual tokens. Second, the punctuation and stop words are removed, and stemming is applied to reduce words' length and set them to their root form. These steps make it possible for the learning algorithm to group similar words. The datasets were filtered only to include actual comments. The comments were converted to lowercase letters. Irrelevant text as subreddit and user mentions and extra whitespace token was removed. Comments are not trimmed with regard to length, as short comments are particularly relevant to the depression identification task. In particular, research has numerously suggested that depression is correlated with the use of first-person pronouns [44].

C. Sentiment Features

We utilize four types of sentiment lexicon, i.e., AFINN [45], NRC(NRC_SA) [46], MPQA [47], and SenticNet [13]. Exploratory data analysis is conducted as it provides insights and facilitates the interpretation of results at the later stages. In the following, these analyses are provided in several figures. Fig. 2(a) visualizes "polarity value" as one of the features from the SenticNet lexicon. The distribution is skewed to the right.

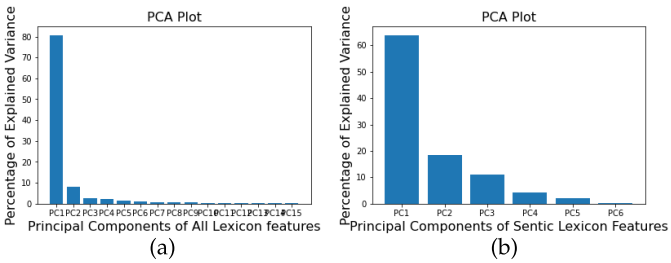


Fig. 3. Principal components of lexicon features. (a) All lexicons. (b) SenticNet.

Fig. 2(b) highlights the distribution of "temper" (a characteristic state of feeling) feature from SenticNet lexicon. Fig. 2(c) highlights the variance of class distribution with regard to the "attitude feature" (a complex mental state involving feelings or way of thinking) from the SenticNet lexicon. The distributions are almost similar in this case. Fig. 2(d) shows the correlation between the features from SenticNet. The green color signifies a correlation close to 1, and the red color signifies a correlation close to -1.

Principal component analysis (PCA) as a tool for an exploratory data analysis is used to reduce the original features into fewer orthogonal variables by applying a change of variable space. PCA is applied because collinearity is spotted among SenticNet and all lexicon features. PCA is particularly useful in processing data where multicollinearity exists between the variables. Fig. 3(a) shows that for all lexicon features, almost all the variation is along with the first principal component (PC). Therefore, a 2-D graph using PC1 and PC2 can be informative for representing the original data. Table II helps to identify which features have the largest effect on the first PC. Evaluating the loadings can help characterize each component in terms of the features. Table II shows that for all lexicon, the two features—strong subjectivity and MPQA-negative—have more explanatory power as they have the highest loading scores (0.2727, 0.2695). For SenticNet in Table II, the features' polarity value and sensitivity have the highest loading scores. For both lexicons, the data points after PCA make up one cluster, suggesting that they correlated with each other [Fig. 4(a) and (b)] as the separation of data points is not very rigid, and this suggests that data points are not very different from each other. For SenticNet, the loading values are very similar (Table II); thus, many features play a role in grouping the comments into one cluster, rather than just one or two features. The first two components describe 0.8067519 and 0.08012761 of variation, as shown in Table III. For SenticNet, both components 1 and 2 describe, respectively, 0.63942615 and 0.18486088 of the variation (Table III) and Fig. 3(b) shows that the first PC describes much variation in the data.

D. Baselines and Settings

Four main models are created to serve as a baseline for this study, and two folds of comparisons were implemented. The experiments are each conducted across three different datasets. The first fold of experiments builds on DL models.

TABLE II
FIRST PC FEATURE LOADINGS

All lexicon	strong-subjectivty	MPQA-negative	NRC-negative	NRC-positive
	0.272753	0.269599	0.268983	0.265303
SenticNet lexicon	**polarity-value**	**sensitivity**	**introspection**	**polarity-label**
	0.487506	0.455703	0.455426	0.445687

TABLE III
PCA COMPONENT EXPLAINED VARIATIONS

Components	c1	c2	c3	c4
All lexicon	0.8067519	0.08012761	0.02700621	0.02136788
SenticNet lexicon	0.63942615	0.18486088	0.10877779	0.04186599

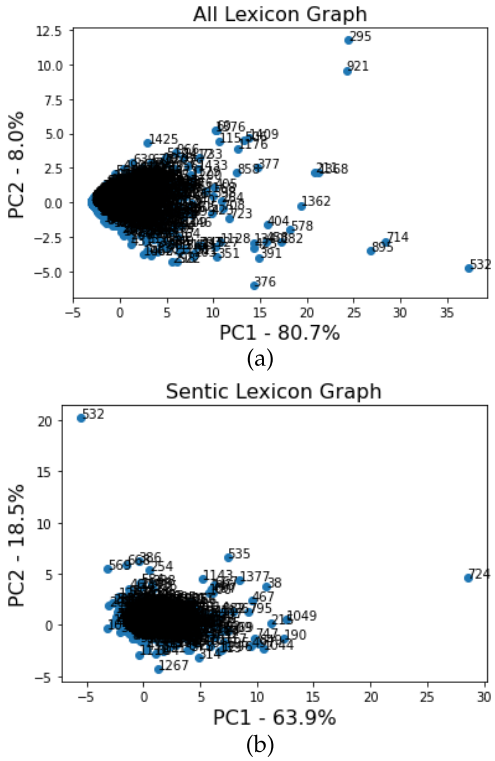


Fig. 4. Lexicon feature visualization with dimension reduction. (a) All lexicons. (b) SenticNet.

Specifically, three architectures are applied for this purpose. These models utilize raw text represented as word embeddings.

1) LSTM: The LSTM network takes the input sentence, which has a shape of (batch size, length of sequences) and returns the output of the linear layer containing logits for the positive and negative class, which receives its input as the final hidden state of the LSTM. The final output layer has the shape of (batch size, output size). For this model, a dropout probability of 0.2, a weight decay of $1e^{-2}$, a learning rate of $2e^{-2}$, a batch size of 32, and five epochs were used.

2) AttentionLSTM: The AttentionLSTM has the batch size of which is the same as the batch size of the data returned by the TorchText BucketIterator. The output is (positive, negative), and the hidden layer size is the same as the size of the hidden state of the LSTM. The vocabulary size is the number of unique words, and the embedding dimension is obtained from pretrained GloVe word embeddings. For this model, a dropout probability of 0.2, a weight decay of $1e^{-2}$, the learning rate of $2e^{-2}$, a batch size of 32, and five epochs were used.

3) Hybrid Lexicon-Based LR: The second fold of experiments builds on the lexicon-based comparisons. In particular, the preobtained SenticNet lexicon combined with LR is compared with two other lexicons. For these experiments, an LR model was developed using Python's scikit-learn's LogisticRegression with the inverse of the regularization strength set to 1.

E. Evaluation Metrics

To evaluate the classification methods described above, several evaluation metrics were applied. These metrics consist of precision, recall, F1 score, and accuracy. Accuracy is defined as the sum of the correct predictions divided by the sum of all predictions made on a dataset. However, for an imbalanced dataset, accuracy as a performance measure is not reflective of models' performance [48] because the number of data points from the majority class (nondepression) will largely outnumber the number of data points in the minority class. Consequently, even for models with low performance, high accuracy can be achieved depending on the class imbalance. Precision quantifies the number of correct positive predictions, and recall quantifies the number of correct positive predictions made of all positive predictions that could have been made [48].

While precision only considers the correct positive prediction, recall also provides information about the missed positive predictions [49]. For imbalance dataset, recall is specifically helpful in highlighting coverage of the minority class. However, neither precision nor recall alone can provide a complete guide to model performance [49]. F1 combines the two and is the metric variant most often used when learning from an imbalanced dataset. In this study, all four metrics are reported. The comparison of models' performance has been mainly based on F1 score and accuracy.

F. Results

The task in this study has been to detect the depression of the users across three datasets. The aim of combining these distinct NLP methods has been to find out what combination of models and features best enhance the performance accuracy of depression detection. In this part, the results and performance metrics achieved by the experiments as mentioned above will be discussed.

Four classifiers are applied in the experiments, three of which are based on artificial neural networks. These models consist of LR and RNN-based neural networks, such as LSTM and AttentionLSTM. The variety of DL methods helped to build up the sophistication of the models gradually. For the implementation of these classifiers, Pytorch and Scikit-learn libraries from Python language are utilized, and fivefold cross validation is used to verify the results.

Table IV shows the evaluation metrics' result of four classification models with eight lexicon-based features. Table V provides the combined results of LSTM and all-lexicon LR

TABLE IV
RESULT COMPARISON BETWEEN DL AND LEXICON-BASED METHODS ON THE CLPSYCH DATASET

Category	Model	Precision	Recall	F1 Score	Accuracy
Deep Learning	AttentionLSTM	**0.6481**	**0.6467**	**0.6469**	**0.6466**
	LSTM	0.5533	0.5940	0.5744	0.5222
Lexicon	All-Lexicons	0.6436	0.6433	0.6431	0.6433
	SenticNet	0.6439	0.6433	0.6430	0.6433
	AFINN + NRC	0.6439	0.6433	0.6430	0.5566
	MPQA + NRC_SA	0.6439	0.6433	0.6430	0.6300
	NRC	0.6439	0.6433	0.6430	0.5666
	MPQA	0.6439	0.6433	0.6430	0.5333
	AFINN	0.6439	0.6433	0.6430	0.5900
	NRC_SA	0.6439	0.6433	0.6430	0.5966

TABLE V
RESULT COMPARISON WITH ENSEMBLE METHODS ON THREE DATASETS

Dataset	Model	Precision	Recall	F1 Score	Accuracy
Reddit	LSTM	0.5333	0.5117	0.5512	0.5238
	All-lexicon LR	0.7401	0.7488	0.7281	0.7487
	Ensemble	**0.8115**	**0.7512**	**0.7701**	**0.7512**
eRisk	LSTM	0.5868	0.5133	0.5476	0.5133
	All-lexicon LR	0.6736	0.5027	0.5757	0.7513
	Ensemble	**0.8005**	**0.7455**	**0.7655**	**0.7555**
CLPsych	LSTM	0.5533	0.5940	0.5744	0.5222
	All-lexicon LR	0.6436	0.6433	0.6431	0.6433
	Ensemble	**0.6550**	**0.6500**	**0.6509**	**0.6500**

classifier using bagging-based ensemble method. The combined output has higher accuracy for all three datasets than any of the LSTM or all-lexicon LR. Overall, the best accuracy is achieved with the ensemble models, especially with the Reddit dataset, resulting in 75% accuracy and 0.77 F1 scores.

This trend is generally followed by the hybrid lexicon-based models that perform second best after ensemble models, for example, in the case of Reddit dataset and accuracy of 74% and 0.72 F1 scores for the hybrid all-lexicon model. Besides, among the hybrid lexicon-based models, we can see that single-lexicon LR and bi-lexicon LR models, in general, do not perform as well as all-lexicon and Sentic LR models. In particular, this contrast can be observed in the CLPsych dataset where single- and bi-lexicon LR models have an accuracy of less than 60%, but the all-lexicon and SenticNet LR models have an accuracy of 64%. The best feature among the single feature sets is NRC_SA, which scores the highest F1 with two out of the three datasets.

DL-based models rely on word-based embedding representation of textual data. Texts produced by users on social media can be unique to the users and have high variance as users can use repetition in letters or words and apply emojis. For this reason, word embeddings may not fully capture and represent nuances of data in a social media text. In comparison, hybrid models are based on sentiment lexicon features to represent textual data, and thus, word embeddings can be one driving source of performance difference among hybrid and DL models.

V. CONCLUSION AND FUTURE DIRECTIONS

This study is conducted with the goal of identifying depression in three social media datasets. As a result, various text classification methods have studied and characterized a connection between language usage and depression. Different sentiment lexicons were used and combined with DL pipelines. The effect of single-lexicon features and combined lexicons was examined. The combined set of features (using all-lexicon features) is demonstrated in DL-based and LR models. Overall, the ensemble models outperform the hybrid lexicon- and DL-based classification models. The strength and effectiveness of ensemble models are demonstrated with the classifier reaching 75% accuracy and 0.77 F1 scores achieving the highest performance degree for detecting the presence of depression in the Reddit social media dataset in this study. In addition, the results show that models that utilized several lexical features outperformed the models based on single-lexicon feature sets. Although major recent breakthroughs in subsymbolic AI are DL-based complex black-box models, the results of this study highlight that by utilizing sentiment lexicon features, the classical models—such as LR—can yield high performance as well.

Although this study shows that the applied feature set improves the classification performance, the absolute value of the evaluation metrics indicates that this task can be further explored and improved. DL architectures were applied in this study. Experiments could be extended with other models for text classification, such as CNNs and transformer-based pretrained language models. Future work can extend possibilities for improvement, which lies in utilizing features such as POS tags and other methods of handling imbalanced datasets.

VI. SOCIAL IMPACT

The study of social media-based mental health assessment holds ethical questions to be addressed. From a mental health point of view, misclassification can impact mental health indicators and assessment. Therefore, the features should be integrated into mental health systems responsibly. The models may help social workers find potential individuals in need of early prevention. However, the model predictions are not psychiatric diagnoses. We recommend anyone who suffers from mental health issues to call the local mental health helpline and seek professional help if possible.

In addition, maintaining transparency about what features are driven and by whom is critical. As the social stigma surrounding mental illness prevents affected individuals from seeking professional assistance, data protection and ownership frameworks are needed to ensure that users are not harmed. As it is helpful for individuals suffering from mental health disorders to seek timely help, NLP can be employed in different ways to accommodate this process. For example, future studies can investigate the relationship between users' personalities and depression-related indicators.

Data privacy is an important issue, and we try to minimize the privacy impact when using social posts for model training. The datasets applied in this study are publicly available. They contain anonymous posts that are manifestly available to the public. We have not attempted to identify the anonymous users or interact with any anonymous users. The collected data are stored securely with password protection. There might also

be some bias, fairness, uncertainty, and interpretability issues during the data collection and model training. Evaluation of those issues is essential in future research.

ACKNOWLEDGMENT

The authors acknowledge the computational resources provided by the Aalto Science-IT Project and CSC–IT Center for Science, Finland.

REFERENCES

[1] B. H. Hidaka, "Depression as a disease of modernity: Explanations for increasing prevalence," *J. Affect. Disorders*, vol. 140, no. 3, pp. 205–214, Nov. 2012.

[2] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, p. e442, Nov. 2006.

[3] S. Rodrigues *et al.*, "Impact of stigma on veteran treatment seeking for depression," *Amer. J. Psychiatric Rehabil.*, vol. 17, no. 2, pp. 128–146, Apr. 2014.

[4] S. Ji, C. P. Yu, S.-F. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, pp. 1–10, Sep. 2018.

[5] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Arch. Gen. Psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[6] S. H. Hosseini-Saravani, S. Besharati, H. Calvo, and A. Gelbukh, "Depression detection in social media using a psychoanalytical technique for feature extraction and a cognitive based classifier," in *Proc. Mex. Int. Conf. Artif. Intell.* Springer, 2020, pp. 282–292. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-60887-3_25

[7] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cogn. Emotion*, vol. 18, no. 8, pp. 1121–1133, Dec. 2004.

[8] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

[9] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Comput. Appl.*, pp. 1–11, Jun. 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00521-021-06208-y

[10] M. Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health," in *Proc. Int. AAAI Conf. Web Social Media*, 2011, vol. 5, no. 1, pp. 265–272.

[11] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared Facebook data," in *Proc. 17th ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2014, pp. 626–638.

[12] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, "Forecasting the onset and course of mental illness with Twitter data," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Dec. 2017.

[13] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 105–114.

[14] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2014.

[15] A. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health using social media text," 2017, *arXiv:1712.03538*.

[16] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, 2015, pp. 1–10.

[17] S. Paul, S. K. Jandhyala, and T. Basu, "Early detection of signs of anorexia and depression over social media using effective machine learning frameworks," in *Proc. CLEF, Working Notes*, 2018, pp. 1–15.

[18] M. Nadeem, "Identifying depression on Twitter," 2016, *arXiv:1607.07384*.

[19] Y. Tyshchenko, "Depression and anxiety detection from blog posts data," Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia, Tech. Rep. 53001016, 2018.

[20] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP," in *Proc. 1st Int. Workshop Lang. Cogn. Comput. Models*, 2018, pp. 11–21.

[21] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Lang. Eng.*, vol. 23, no. 5, pp. 649–685, 2017.

[22] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard Clinic*, 2018, pp. 88–97.

[23] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 217–226, Jul./Sep. 2014.

[24] D. Maupomé and M.-J. Meurs, "Using topic extraction on social media content for the early detection of depression," in *Proc. CLEF, Working Notes*, vol. 2125, 2018, pp. 1–5.

[25] J. Misra, "AutoNLP: NLP feature recommendations for text analytics applications," 2020, *arXiv:2002.03056*.

[26] M. Stankevich, V. Isakov, D. Devyatkin, and I. Smirnov, "Feature engineering for depression detection in social media," in *Proc. ICPRAM*, 2018, pp. 426–431.

[27] T. Shen *et al.*, "Cross-domain depression detection via harvesting social media," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1611–1617.

[28] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from Twitter activity," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 3187–3196.

[29] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[30] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.

[31] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, Jul. 2017.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.

[34] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[37] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–11.

[38] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," *Inf. Sci.*, vol. 177, no. 18, pp. 3799–3821, Sep. 2007.

[39] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.

[40] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, 2015, pp. 31–39.

[41] I. Pirina and Ç. Çöltekin, "Identifying depression on reddit: The effect of training data," in *Proc. EMNLP Workshop SMM4H, 3rd Social Media Mining Health Appl. Workshop Shared Task*, 2018, pp. 9–12.

[42] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 1, pp. 214–226, Feb. 2021.

[43] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Springer, 2016, pp. 28–39. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-44564-9_3

[44] J. Zimmermann, T. Brockmeyer, M. Hunn, H. Schauenburg, and M. Wolf, "First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients," *Clin. Psychol. Psychotherapy*, vol. 24, no. 2, pp. 384–391, Mar. 2017.

[45] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," 2011, *arXiv:1103.2903*.

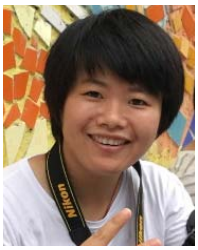
[46] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical Turk to create an emotion lexicon," in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, 2010, pp. 26–34.
[47] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Conf. Hum. Lang. Technol. Empirical Methods Natural Lang. Process.*, 2005, pp. 347–354.
[48] B. Juba and H. S. Le, "Precision-recall versus accuracy and the role of large data sets," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 4039–4048.
[49] T. Basu and C. A. Murthy, "A feature selection method for improved document classification," in *Proc. Int. Conf. Adv. Data Mining Appl.* Springer, 2012, pp. 296–305. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-35527-1_25

Luna Ansari is a Postdoctoral Researcher and a Research Assistant at Aalto University, Espoo, Finland. She received the Ph.D. degree in organizational science and did a research visit at Stanford University Scancor Research Institute, Stanfor, CA, USA. Parallel to that, she did master's studies in artificial intelligence and bachelor's studies in data science at Aalto University.

Her research interests include brain structure, cognitive-linguistics, NLP, computational linguistics, and social psychology. In particular, she is investigating the combination of sub-symbolic and symbolic AI for depression detection task.

Shaoxiong Ji is currently pursuing the Ph.D. degree with the Department of Computer Science, Aalto University, Espoo, Finland.

He was a Visiting Researcher with the Finnish Institute for Health and Welfare, Helsinki, Finland, and a part-time an Artificial Intelligence Scientist with Silo AI, Helsinki. Prior to Aalto, he did an M.Phil. research at The University of Queensland, Saint Lucia, QLD, Australia, worked as a Research Assistant and a Visiting Scholar with the University of Technology Sydney, Ultimo, NSW, Australia, and did a visiting research at Nanyang Technological University, Singapore. His research interests include machine learning (ML), natural language processing, and health informatics.



Qian Chen received the bachelor's degree in information and computational science from the Dalian University of Technology, Dalian, China. She is currently pursuing the Ph.D. degree with Nanyang Technological University, Singapore.

In particular, she is investigating depression patterns in multimodal learning for depression recognition tasks. Her main research interests are image sentiment analysis analyzing sentiment from image that includes general image sentiment analysis (images posted on social media platforms), facial expression recognition, and emotion recognition.

Erik Cambria (Fellow, IEEE) received the joint Ph.D. degree from the University of Stirling, Stirling, U.K., and the MIT Media Lab, Cambridge, MA, USA, through a joint program.

He is the Founder of SenticNet, Singapore, a Singapore-based company offering B2B sentiment analysis services, and an Associate Professor with Nanyang Technological University (NTU), Singapore, where he also holds the appointment of Provost Chair in computer science and engineering. Prior to joining NTU, he worked at Microsoft Research Asia, Beijing, China, and HP Labs, Bengaluru, India. His research interest includes neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems, and financial forecasting.

Mr. Cambria was a recipient of several awards, such as the IEEE Outstanding Career Award. He was listed among the AI's ten to watch. He was featured in Forbes as one of the five people building our AI future.