A HYBRID ENSEMBLE METHOD FOR DETECTING DEPRESSION

Vidya Y [1], Dr. Kalaiarasan[2]

Article History: Received: 06.04.2023 **Revised:** 22.05.202 **Accepted:** 01.08.2023 **Published**: 02.08.2023

Abstract

Over the past century, people have suffered from depression more frequently due to changes in lifestyles. Many cases of mental illness remain undiagnosed, even though the rates of diagnosis have improved in recent years It can be helpful to identify individuals at risk of depression or depressed using automated detection methods. In order to understand depression detection, it is necessary to represent and analyze features in language. Detecting depression using text classifiers is the topic of this article. Hybrid and ensemble methods are examined and compared with the aim of improving depression detection performance. Compared to hybrid models, ensemble models perform better in classification.Multiplying features and selecting the most appropriate features can result in enhanced performance.

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

[1]Computer Science & Engineering, Presidency University Bengaluru, India

Email: [1]vidya.yc@gmail.com, [2]kalaiarasan@presidencyuniversity.in

DOI: 10.31838/ecb/2023.12.s3.796

1. Introduction

A drastic change in lifestyles has led to a greater number of people suffering from depression in modern society.By 2030, depression is expected to be one of the three most common causes of disease [2]. Depression is known as "a disease of modernity" [1]. In addition to social stigma and a high rate of misdiagnosis, depression is difficult to diagnose and treat properly.Suicidal thoughts can develop in patients with mental disorders if they are not treated effectively [4]. Depression can thus be detected early for the benefit of both individuals and society.Depression symptoms can appear in various ways and to varying degrees [5].One source of depressive symptoms that people can recognize is language [6].Numerous cognitive and linguistic investigations [6] have shown that people with depression use language characteristics differently.They frequently utilize terms with negative meanings and the first-person singular (I, us, or we) pronoun [7].

As a platform for user communication, online social content contributes to automatic detection of mental disorders.Various researchers have introduced new forms of health care solutions in recent years using social networking platforms to study users' behavior [8], [9].Social media can also play a role in individuals' decision-making when it comes to seeking professional assistance due to the stigma associated with depression. A study of depression can benefit from the study of social media as an important source of information about individuals' opinions and feelings [10]. There have been various approaches and levels of granularity in research related to depression detection. Research on depression and other mental disorders, such as postpartum depression and post traumatic stress disorder (PTSD), has been conducted on a variety of social network sites (SNSs),

such as Reddit, Twitter, Facebook, and Weibo[12].

Deep learning (DL) and machine learning (ML) methods have been applied to depression detection primarily from the bottom up. Despite their usefulness in providing insights about word frequencies and statistical correlations, subsymbolic artificial intelligence (AI) methods are not sufficient to analyze narrative and understand dialog systems in sentiment analysis [13]. NLP methods have improved using deep learning methods, but their predictive power has been limited mainly due to DL's ability to learn from large quantities of data. Furthermore, communication involves a broader range of contributors, including cultural consciousness, social norms, and understanding the world. As a solution to these challenges, recent research in depression detection has applied symbolic AI techniques such as logical reasoning to apply top-down learning to depression detection. Natural language texts can be induced to exhibit more meaningful patterns even when using subsymbolic approaches and symbolic methods in combination [13].Therefore, automated depression detection must integrate symbolic learning strategies with subsymbolic approaches.

The ensemble method, in which several learning methods are combined, is another method that yields high accuracy [14].There has been extensive use of ensemble methods to solve a variety of predictive problems [14].

By extending existing knowledge on automated depression detection, the current study builds on these recent advancements. In this study, depression detection is improved as a text classification task as part of the contribution to the literature. The purpose of this study is to show how hybrid methods can enhance the effectiveness of symbolic and subsymbolic methods for depression identification.Three datasets are examined eight times each for this purpose.Different sentiment lexicons are utilized in the hybrid experiments, along with logistic regression (LR) for text classification.The experiments with the ensemble methods are conducted by combining DL approaches and lexicon-based models. Long short-term memory (LSTM) and AttentionLSTM are two DL methods included in this set of experiments.The use of hybrid approaches to automatically identify depression adds to the body of knowledge on the subject.

The following is a breakdown of this article.The second section focuses on automated depression detection and NLP techniques for text classification. Section III discusses the methods of learning.A description of the experiments conducted, an overview of the collected data, and the results of the exploratory data analysis are provided.In addition to comparing the two models, conclusions are made, limitations are discussed, and future research directions are suggested.

Related Work

Various NLP techniques have been developed for capturing the linguistic tendencies of users through the evaluation of textual data. Classifier models have been used to automate the examination of the relationship between language and mental states with the aid of feature extraction techniques.

N-grams [15], linguistic inquiry and word count (LIWC) [16], and bag of words (BOWs) [17, 18] are examples of statistical techniques that have been used to examine how single features can be extracted. Other studies have compared the effects of various machine learning strategies on single features.By merging distinct features, researchers have looked into strategies to increase classification accuracy. Several approaches are used in these studies, including term frequency-inverse document frequency (TF-IDF), linear discriminant analysis (LDA) [19], and 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. A Bell Lets Talk depression detection on the CLPsych2015 dataset was improved using word embeddings, for example, by Orabi et al. [22]. Models based on convolutional neural networks (CNN) performed better than those based on reinforcement learning (RNN).There was higher generalization power with CNN-based models combined with optimized embeddings [22]. In a study involving a small dataset, Benton et al. [15] evaluated the effectiveness of multitask learning (MTL) models.Based on the combination of feedforward multilayer perceptrons and MTLs, the authors predicted a set of mental state predictions. 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 Live Journal social networking service. A 30-topic extraction algorithm was also developed by Maupomé and Meurs [24] using unigram, bigram, and trigram frequency to combine an unsupervised topic extraction algorithm with a multilayer perceptron.Neural networks failed to perform well with a limited number of data points [24].Identifying the right set of features to analyze natural language is a significant challenge when designing natural language analysis.Text classification literature has been analyzed using a variety of methods in order to obtain a set of relevant features. There are a number of ways to carry out this analysis, including "BOWs," "bags of phrases," "bags of n-grams," "Word Net-based word generalizations," and "word embedding" techniques. According to Misra [25], in a recent review study, language feature specifications have been reused in similar semantically-related contexts using machine-learning and deep learning approaches. It is possible to extract linguistic, semantic, and statistical features from textual data by considering several levels of analysis, such as words, phrases, sentences, paragraphs, documents, and corpora.Text classification context is discussed in terms of four main approaches[25]. By applying lexico-syntactic patterns to documents, the first features can be extracted. It is possible to capture these patterns as part of speech (POS) tag patterns, i.e., when a sequence of words matches a specific expression.A second set of features is those that are based on semantic similarity and relatedness. Analyzing latent semantics is possible using vector space modeling, topic modeling, neural embedding, and neural network embedding. In addition to identifying semantic features, relationships between concepts can be analyzed across a broad range of text corpora in order to extract semantic features.The fourth method of feature extraction uses statistical analysis and feature engineering to discover statistical features contained within texts.

It is possible to detect mental states using linguistic cues, statistical features, and a user's posting pattern by using classification models built using linguistic cues and statistical features.The features included in this study are stress-related usage of language, timing and frequency of posting, sentiment and value contrast, i.e., the polarity of posting shifting between positive and negative. In Stankevich et al. [26], embedding and bag-of-words were applied to the recall and accuracy of TF-IDF models. However, TF-IDF models with morphological features were found to provide better performance and precision scores than embedding features (63% F1-score).Shen et al. [27] described online social behavior and clinical depression criteria based on online social behavior using six feature extraction approaches. Using a tagged Twitter dataset and a multimodal depressed dictionary learning model, we were able to get an F1-score of 85%.To assess the severity of depression among Twitter users, Tsugawa et al. [28] employed topic modeling to ascertain individuals' mental states based on their prior online behavior.

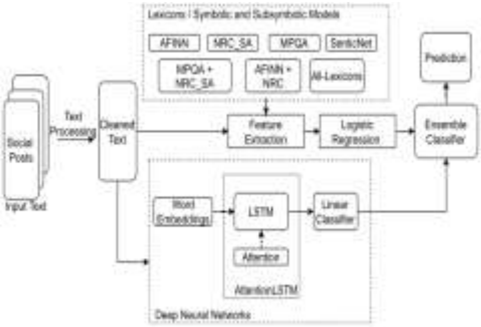


Fig 1.Flowchart for the suggested ensemble model.

The combination of symbolic and subsymbolic AI can be used in this context to enhance sentiment classification tasks, despite these studies providing valuable insights into the use of textual data in mental state detection.

2. Methods

A wide variety of problems can be solved with NLP technologies across many different domains, including text summarization, translation, and sentiment analysis. The hidden Markov model [29] was used in earlier attempts to solve these problems, which required extensive data engineering. DL has been used more frequently recently in NLP methods [30, 31].In recent years, powerful computing platforms have made end-to-end training possible as ML helps address a variety of problems, including speech recognition [32], image recognition [33], and natural language processing [34].

A bag-of-words representation of documents was used to represent text classification early on, and ML methods were applied without sequential processing of the words [31]. Due to their reliance on word embeddings, LSTM [35] neural networks have been used for text classification due to their ability to reduce the number of training samples.

The study draws on recent developments in speech and NLP, including RNNs, LSTMs, and attention-based models, to analyze and classify texts by identifying sentimental fragments. The flowchart of the proposed ensemble model is shown in Fig. 1.

A. LSTM Networks

The output of RNNs is dependent on the hidden state for previous tokens in a sequence, so their output is a kind of neural network architecture [34].In the field of natural language processing, LSTMs are widely used to solve text classification problems. RNNs with gated cells such as LSTM and gated cells LSTM are special kinds of RNNs.Although RNNs are theoretically capable of retaining information over time, they are difficult to use in practice given their inability to handle "long-term dependencies". There are four different subnetworks in LSTM cells: the cell state, the input gate, the forget gate, and the output gate. As part of the filtering process, the forget gate and input gate determine which information should be added to and removed from the cell state. Sigmoid functions are outputs from the output gate, which determines how the cell state and current token are to be combined.

B. Attention Mechanism

A learning algorithm uses the attention mechanism in order to focus more on certain inputs, while ignoring others [36]. A weighted arithmetic mean of all these subtexts and contexts is returned by this algorithm.Transformer models are new families of neural network architectures that are based on attention as the primary building block. Transformators are self-attention blocks coupled with feedforward networks in a transformer-based model. This system is composed of an encoder that encodes the input text and a decoder that produces the output [37]. It uses the context of a sentence as input to construct a transformer.There are three components to a context: the preceding sentence, the sentence itself, and the following sentence. The attention unit analyzes the entire sentence and drives the parts that are similar. As a result of detecting such similarity, a sequence can be examined in greater detail. In order to detect relationships between words very close to each other, the attention mechanism prioritizes the relationship between very distant items/words in a sequence.

C. Sentiment Lexicon: Hybrid Learning Approach

To conduct the experiments, the SenticNet sentiment lexicon was one of the most popular sentiment lexicons. Sentiment information is derived from this lexicon. All of the sentences in this training set are represented by different features, such as their polarity values, polarity labels, sensitivity, introspection, temper, and attitude. SenticNet assigns a specific value to each of these features for every word in the sentence.Hybrid machine learning (HML) is used to engineer and drive SenticNet's features. Models that include multiple techniques are called HML models by definition.Data preprocessing is performed by the first component and classification or prediction is performed by the second [38].

SenticNet's sentiment lexicon includes AI capabilities for symbolic and subsymbolic sentiment analysis [13]. To deconstruct multiword expressions into primitives, it employs logical reasoning as symbolic AI in conjunction with DL as subsymbolic AI [13]. The symbolic AI processes information from the top-down, whereas the subsymbolic AI takes a bottom-up approach to NLP. Utilizing both subsymbolic and symbolic DL methods simultaneously has more advantages than doing it separately [13].

D. Logistic Regression

Binary classifiers are trained using LR, a supervised machine learning method.As the training process determines the optimal weight vector w for the predictor h, it is based on a linear graph h(x) w^T x.

TABLE I
DATASETS SUMMARY

Dataset	No. of positive samples	No. of Negative samples
CLPYsych	327+150	246+150
Reddit	1,200	641
eRisk	770	3,728

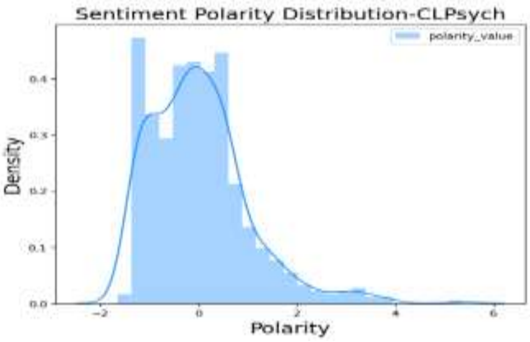
E. Ensemble Methods

In this study, ensemble methods were also used. In data science competitions, these approaches have often placed among the top winners for achieving record performance on challenging datasets.As a general rule, the ensemble approach is applied to reduce bias in the model (boosting), decrease variance in the model (bagging), and enhance predictions (stacking) [39].

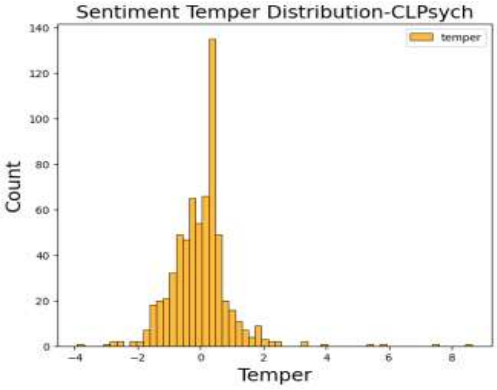
Combining the predictive results using the ensemble method is the current study's approach.In particular, the classification results produced by LSTMs and LRs are averaged. As a result of averaging multiple estimates, bagging-also called bootstrapping-lowers estimation variance of one model [39].As a result of parallel learning, each classifier empathizes different features due to its separate training. When LSTM and LR are combined, accuracy can improve, variance in training datasets can be reduced, and overall variance can be reduced [14].

3. Experiments

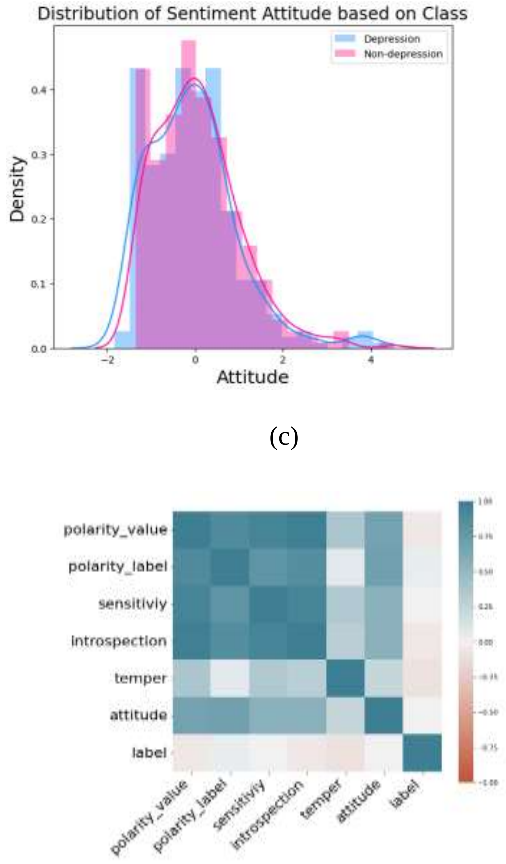
On three public datasets, depression detection experiments were conducted separately.PyTorch[1] and Scikit-learn[2] were used as frameworks and libraries.



(a)



(b)



(c)

(d)

Fig. 2. Visualization of sentiment features in the CLPsych dataset. (a) Distribution of sentiment polarity values. (b)Distribution of temper in the mind. (c) attitudes are distributed. (d) SenticNet feature correlation heatmap.

A. Datasets

This paper's experiments were based on the datasets described in Table I.

1) CLPsych 2015 Shared Task: To foster collaboration between psychologists and computer scientists, the Computational Linguistics and Clinical Psychology (CLPsych) initiative was launched in 2014 [40]. In particular, "shared tasks" have been defined as ways to study and compare similar prediction problems using different methods. A dataset containing posts from Twitter users with depression or PTSD (Depression and PTSD on Twitter). Three binary classification subtasks are included in the study: 1) depression versus control; 2) PTSD versus control; and 3) depression versus PTSD. An age- and gender-matched control user was assigned to each depression and PTSD patient, making a total of 1146 patients in the train partition. Approximately 600 participants took part in the study, 150 with depression, 150 with PTSD, and age- and gender-matched controls for each. A data missing issue in the training and testing sets causes the actual number of users to be 1711.

2) Reddit:The Reddit social media collection contains user posts from both depressed and non-depressed individuals.The dataset contained 1841 users, of which 1200 had good feedback and 641 had negative feedback [41]. Reddit, a social media platform with anonymity, is home to many stigmatic topics.Reddit users' posts on topics related to suicidal ideation and mental health difficulties were explicitly examined by researchers using Reddit data [42]. An 80:20 split rate was used to split the data between train and test sets after concatenation, random shuffle, and concatenation.A column of text comments and a column of labels corresponding to those comments made up the final data frame.Comments were categorized as depression or nondepression by assigning a 1 or 0.

3) eRisk Dataset: A dataset based on the eRisk (Early Risk Prediction) forum [43] was collected for this study.eRisk is a platform for creating reusable datasets and benchmarks and facilitating multidisciplinary research in the area of early risk detection technologies. Depression early warning signs were initially detected by the eRisk 2018 dataset. In the eRisk collection, which contains 4498 posts from both categories of users, there are 770 responses to depressive postings and 3728 responses to nondepressive postings.The data was split between train and test sets using an 80:20 split

rate after concatenation, random shuffling, and concatenation.

B. Data Preprocessing

These datasets are preprocessed using NLP techniques before moving on to the training stage. The posts are first tokenized to make them more readable. A second step involves removing punctuation and stop words and stemming to reduce word lengths and set them to their root forms. Learning algorithms can group words based on similarity through these steps.We filtered the datasets so that only actual comments were included.For the comments, lowercase letters have been utilized.The content was edited to remove the unnecessary whitespace token, user and subreddit mentions, and other formatting. Short remarks that are not clipped are required for the depression identification job. Numerous research have revealed a link between first-person pronoun usage and depression [44].

C. Sentiment Features

The sentiment lexicon employed in this work is comprised of SenticNet [13], AFINN [45], NRC [46], and MPQA [47].It is important to conduct exploratory data analysis in order to get insights and facilitate the interpretation of the results in the future.To further demonstrate these analyses, a variety of figures are presented below. "Polarity value" is depicted in Figure 2(a) as part of the SenticNet vocabulary. A right-skewed distribution is observed.

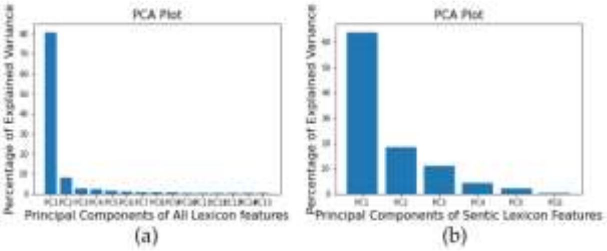


Fig. 3. The key elements of lexical features. (a) All lexicons. (b) SenticNet.

The SenticNet lexicon's "temper" feature (a characteristic state of being) is shown in figure 2(b).Based on SenticNet lexicon, figure 2(c) illustrates variance of class distribution for the "attitude feature" (a complex mental state involving feelings or thoughts). The two distributions are comparable to one other.Figure 2(d) displays the correlation between the SenticNet characteristics.Correlations close to 1 are indicated by green color, and correlations close to 1 by red color.

Principal component analysis (PCA) is used in exploratory data analysis to condense the original features into a smaller number of orthogonal variables. SenticNet and all lexicon features show collinearity, which is analyzed through PCA. PCA can be used to examine multicollinearity in data, especially when the variables are uncorrelated. The first principal component (PC) depicted in figure 3(a) may explain the majority of the variation for all lexical features. As a result, PC1 and PC2 can be used in 2-D graphs to represent the original data in an informative way. It is possible to determine which features have the greatest effect on the first PC based on Table II.Identifying each component's features can be accomplished by evaluating its loadings. According to Table II, strong subjectivity and MPQA-negative both have more explanatory power for all lexicons, as the loading scores for those two features are 0.2727 and 0.2695 respectively. The highest loading scores are obtained for SenticNet by polarity and sensitivity. Based on PCA, both lexicon data points form one cluster, which implies that data points are not very different from one another [Fig 4(a) and (b)], as the separation between data points is not very rigid.Given that SenticNet's loading values are so comparable (Table II), grouping the comments into one set involves more than simply one or two features.

According to Table III, the first two components account for 0.8067519 and 0.08012761 of the variation. Table III indicates that SenticNet components 1 and 2 account for 0.63942615 and 0.18486088 of the variation, respectively, and Fig. 3(b) demonstrates that the first PC explains a large portion of the data volatility.

D. Baselines and Settings

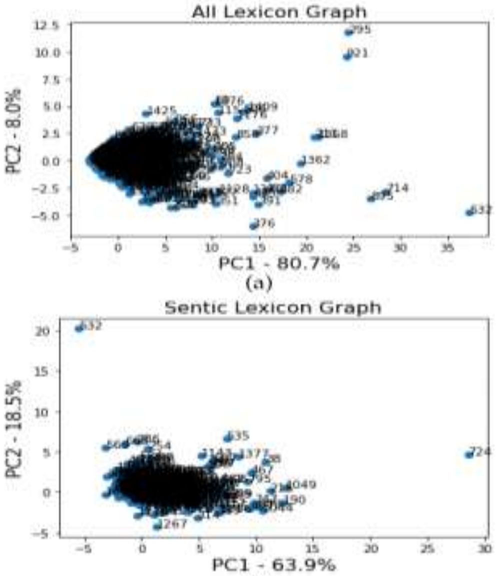
In order to establish a baseline for this investigation, four major models were developed, and two folds of comparisons were used. Each experiment is carried out using three separate datasets. On DL models, the initial fold of experiments is built.In particular, three architectures are utilized for this.These models make use of word embeddings that represent raw text.

TABLE II
FIRST PC FEATURE LOADINGS

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

TABLE III
A VARIATIONS EXPLAINED FOR PCA COMPONENTS

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



(a)

(b)

1) LSTM: The final hidden state of the LSTM network, which receives its input as the phrase with the shape of (batch size, length of sequences), is the output of the linear layer composed of logits for the positive and negative class.
The form of the final output layer is (batch size, output size). The following parameters were utilized for this model: a batch size of 32, a dropout probability of 0.2, a weight decay of 1e^{-2}, a learning rate of 2e^{-2}, and five epochs.

2) AttentionLSTM:Both the AttentionLSTM and the TorchText BucketIterator's output data batches have the same size.The output is (positive, negative), and the hidden layer's size is the same as the LSTM's hidden state's size. The number of distinct words makes up the vocabulary size, and pretrained GloVe word embeddings provide the embedding dimension.The following parameters were utilized in 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.

3) Hybrid Lexicon-Based LR: TBuilding on the lexicon-based comparisons, the second fold of experiments. In specifically, two different lexicons are compared with the acquired SenticNet lexicon when paired with LR. With the inverse of the regularization strength set to 1, a LogisticRegression model using Python's Scikit-Learn was created for these trials.

E. Evaluation Metrics
The classification methods shown above were evaluated using a number of assessment metrics. Precision, recall, F1 score, and accuracy are among these criteria. Accuracy is defined as the total number of correct predictions divided by the total number of predictions made on a dataset. Accuracy as a performance indicator is not indicative of the performance of the models for an unbalanced dataset since the number of data points from the

majority class (nondepression) will greatly outnumber the number of data points in the minority class [48].

As a result, depending on the class imbalance, excellent accuracy can be attained even for models with limited performance. Recall is the proportion of correct positive predictions made among all possible positive predictions, while precision measures the number of correct positive predictions made [48].Recall offers information about the missed positive predictions, whereas precision simply takes into account the right positive predictions [49]. Recall is particularly useful for identifying coverage of the minority class in the imbalance dataset.However, neither recall nor precision by themselves can offer a comprehensive overview of model performance [49]. The metric version most frequently employed for learning from an unbalanced dataset is F1, which combines the two.Each of the four measures is reported in this study. F1 score and accuracy have mostly been used to compare model performance.

4. Results

The goal of this study was to identify user depression across three datasets. The goal of combining these various NLP techniques was to determine which models and feature combinations would most effectively improve the performance accuracy of depression diagnosis.The tests listed above produced the results and performance measures that will be reviewed in this section.Three of the four classifiers used in the studies are built on artificial neural networks. These models use neural networks built on LR and RNNs, like LSTM and AttentionLSTM. The range of DL techniques made it possible to steadily increase the models' level of sophistication. The Python language's Pytorch and Scikit-learn modules are used to create these classifiers, and fivefold cross validation is performed to confirm the outcomes.

The evaluation metrics for the four classification models with eight lexicon-based features are displayed in Table IV.The combined results of the LSTM and all-lexicon LR classifiers utilizing the bagging-based ensemble technique are shown in Table V. For all three datasets, the total result is more accurate than any individual LSTM or all-lexicon LR. Overall, the ensemble models, particularly when applied to the Reddit dataset, yield the highest accuracy (75% accuracy and 0.77 F1 scores).

TABLE IV
CLPSYCH DATASET- COMPARISON OF DL AND LEXICON-BASED METHODS

Category	Model	Precision	Recall	F1 score	Accuracy
Deep Learning	Attention LSTM	**0.6481**	**0.6467**	**0.6469**	**0.6466**
	LSTM	0.5533	0.594	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_SA	0.6439	0.6433	0.6430	0.5566
	NRC	0.6439	0.6433	0.6430	0.6300
	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
ANALYSIS OF THE RESULTS OF THREE DATABASES USING ENSEMBLE METHODS

Dataset	**Model**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Reddit	LSTM	0.5333	0.5117	0.5512	0.5238
	All-Lexicons 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-Lexicons LR	0.6736	0.5027	0.5757	0.7513
	Ensemble	**0.8005**	**0.7455**	**0.7655**	**0.7555**
CLPsych	LSTM	0.5333	0.5940	0.5744	0.5222

	All-Lexicons LR	0.6436	0.6433	0.6431	0.6433
	Ensemble	**0.6550**	**0.6500**	**0.6509**	**0.6500**

The hybrid lexicon-based models typically perform second best behind ensemble models in this regard, as seen, for instance, in the case of the Reddit dataset with accuracy of 74% and 0.72 F1 scores for the hybrid all-lexicon model.

Furthermore, we see that among the hybrid lexicon-based models, the single-lexicon LR and bi-lexicon LR models frequently outperform the all-lexicon and Sentic LR models. This disparity is particularly evident in the CLPsych dataset, where accuracy for single- and bi-lexicon LR models is less than 60% whereas accuracy for all-lexicon and SenticNet LR models is 64%. NRC_SA, which achieves the maximum F1 with two of the three datasets, is the best feature among the single feature sets.

Word-based embedding representation of textual data is a key component of DL-based models. Social media users' texts can be distinctively their own and highly variable because they might repeat letters or words or use emoticons.Word embeddings might not accurately collect and depict the subtleties of data in a social media text as a result.Since hybrid techniques, in contrast to DL models, are based on sentiment lexicon properties to represent textual input, word embeddings may have a significant role in the performance disparity between the two models.

5. Conclusion And Future Directions

This study aims to identify depression using three social media datasets. As a result, a relationship between language use and depression has been examined and defined using a variety of text categorization techniques.DL pipelines were integrated with several sentiment lexicons.It was investigated how single-lexicon features and mixed lexicons affected each other.In DL-based and LR models, the combined set of features (using all-lexicon features) is shown.

Overall, the hybrid lexicon- and DL-based classification models underperform the ensemble models.The strength and effectiveness of ensemble models are demonstrated by the classifier's performance of 75% accuracy and 0.77 F1 scores, which achieved the maximum performance degree for identifying the existence of sadness in the Reddit social media dataset used in this work. The results also show that models based on multiple lexical feature sets perform better than models based on single-lexicon feature sets.

Despite the fact that key recent developments in subsymbolic AI are DL-based complex black-box models, the findings of this study demonstrate that classical models, like LR, can achieve outstanding performance by using sentiment lexical properties.Although this study demonstrates that the applied feature set enhances classification performance, the work can still be studied and enhanced, according on the assessment metrics' absolute values. In this work, DL architectures were utilized. Other text classification models, including CNNs and pre-trained language models based on transformers, might be included to the experiments.

Future research can increase the potential for improvement by leveraging POS tags and other techniques for handling unbalanced information.

IV. PUBLIC IMPACT

There are ethical issues that need to be resolved when looking at social media-based mental health evaluation.

Misclassification can affect mental health indicators and assessments from a mental health perspective. As a result, the features should be thoughtfully incorporated into mental health systems. Social workers may use the models to identify potential clients who require early intervention.The model's predictions, however, do not represent psychiatric diagnoses.

We advise anyone experiencing mental health problems to contact their local mental health helpline and, if available, seek professional assistance.Furthermore, it is crucial to retain transparency on who is driving which features and in what way. Frameworks for data ownership and protection are required to ensure that users are not harmed since the social stigma attached to mental illness deters those who are affected from seeking professional help.

Because it is advantageous for persons with mental health issues to seek therapy as soon as feasible, NLP can be utilized in a variety of ways to support this process.Future research can, for instance, look into how users' personality and symptoms of sadness are related.Since data privacy is a serious issue, we work to minimize the privacy impact while using social media for model training.

The datasets used in this work can be accessed by anybody. They include anonymous posts that are clearly accessible to the general public. We haven't made any effort to contact or identify any of the anonymous users. The gathered data are safely kept and password-protected. Additionally, throughout the data gathering and model training, there might be some concerns with bias, justice, uncertainty, and interpretability.It is crucial for future study to evaluate those difficulties.

6. 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. Lin- guistics 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: Attend- ing 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 Com- put. 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