Detecting the Severity of Depression in Online Forum Data by Leveraging Implicit Semantic Inferences

Shaveen Thialakratne, Viraj Welgama, Ruvan Weerasinghe

Abstract—**Depression, a prevalent mental health disorder with global implications, exerts a profound negative influence on individuals' lives. While the prediction of depression (as a binary classification task) is a well-established research area, depression severity detection is a new research direction with limited studies. In the context of detecting the severity of depression through online forum data, this research endeavors to offer two distinct solutions by employing Ada embeddings, GPT 3.5 Turbo, and LIWC as feature engineering techniques, while AutoSklearn serves as the ensemble learning algorithm. Notably, the outcomes of this study significantly outperform existing state-of-the-art models on both depression severity annotated datasets used in this research. The results also showcase the potential reuse nature of the proposed models in diverse data sources due to their high performance in both datasets. Furthermore, as a valuable practical outcome, a software prototype has been developed, capable of providing the depression severity level, along with associated symptoms and keywords, upon inputting an online forum post.**

Index Terms—**Depression, Language Models, Natural Language Processing, Machine Learning, Deep Learning**

I. Introduction

Depression, also known as major depressive disorder, is a debilitating mental health condition that has significant negative impacts on an individual's life. Globally, there are over 300 million individuals who suffer from major depressive disorder [1]. Research has demonstrated that depression is a preventable condition that can be treated with medical support and interventions [2]. Given the severity and the progressive nature of this mental condition, early detection of depression is essential for effective treatment. Nevertheless, approximately 70% of these individuals with major depressive disorder, do not receive treatment due to various reasons, such as social stigma, ignorance, negligence, and insecurity [3]. This has led many individuals to seek support and information online, particularly through online forums and social and media-sharing networks.

In recent times, social media platforms have become an integral part of people's daily lives, where they share their thoughts, feelings, and experiences with others [4]. Online forums and social media sharing networks are common examples of these platforms. Due to the global popularity and wide usage of these platforms , many individuals suffering from different mental health conditions have tended to seek support and information through online forums and social media sharing networks [5]. Online support forums are particularly useful in this regard due to their main focus on such mental health conditions . Examples of such forums include Beyond Blue Online Forum, Understand Forum, and 7 Cup Forum [6]. Other online forums include Reddit and Quora, while popular social media sharing platforms include Facebook, Instagram, Twitter, and YouTube.

The proliferation of such online interactions has provided valuable information for researchers in the mental health domain to develop computational techniques to automatically detect depression in individuals . Early identification of depression is crucial because it allows to identify and support individuals who require professional care [7]. Due to this reason, researchers worldwide have contributed to this emerging research direction by using various computational techniques to automatically detect depression by analyzing data from social media platforms [8]. Despite the significant progress in the field, most existing work has only focused on detecting depression as a binary classification task, i.e., identifying whether a given user post exhibits depression or not . The task of predicting the severity levels of depression in a given user post is rarely explored in this field of study [6]. However, the severity level of depression is critical to promptly identifying the appropriate therapeutic processes and medications for individuals . Identification of the severity level of depression can also prevent fatal accidents, such as self-harm and suicides.

With these considerations in mind, this study contributes to this nascent research direction with the primary focus of detecting the severity levels of depression from online forum posts by leveraging cutting-edge technologies in Natural Language Processing (NLP) and machine learning . More specifically , to the best of our knowledge , this is the first study in the field that performs fine-grained implicit semantic inferences by leveraging chat completion language models in order to elicit latent cues beyond the content in the user posts to accurately predict depression severity levels.

Correspondence: Shaveen Thialakratne (E-mail: shaveenthilakaratne@gmail.com) Received: 08-01-2024 Revised: 10-07-2024 Accepted: 17-07-2024

Shaveen Thialakratne, Viraj Welgama and Ruvan Weerasinghe are from University of Colombo School of Computing, Sri Lanka
(shaveenthilakaratne @gmail.com,wvw@ucsc.cmb.ac.lk, arw@ucsc.cmb.ac.lk)

DOI: https://doi.org/10.4038/icter.v17i2.7280

II. Related Work

A. Taxonomies for Detecting Depression and Severity

Depression is a complex mental disorder that poses a challenge to diagnosis due to the lack of direct laboratory tests. While some indirect laboratory tests on thyroid functioning are available [9], the typical approach to diagnosing depression involves analyzing the patient 's condition and measuring the severity through expert analysis based on individual incidents and experiences or using psychometric tests [6]. However, many patients do not seek medical assistance for depression due to reasons such as stigma and instead, obtain support from online platforms . Consequently , psychometric tests have become valuable diagnostic tools for depression as they are easy to use and self-administrative.

This section provides background on the diagnosis criteria and severity scales used in the literature and discusses some popular scales that are heavily utilized in depression research. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5), designed by the American Psychiatric Association , utilizes a binary -fashioned taxonomy with 13 binary statements divided into two main sections, namely, symptoms and additional required criteria [10]. Beck Depression Inventory (BDI), proposed by Beck et al., is a widely used psychometric test consisting of 21 sections categorized as multiple -choice questions with four statements of increasing severity . Each question is weighted from 0 to 3 according to the answer, with a maximum score of 62. BDI is rated for ages 13-80 and has a 4-point rating scale, namely , no depression , mild depression , moderate depression, and severe depression [11]. There is also a modified version of BDI, namely, the Children 's Depression Inventory (CDI), rated for ages 7-17 [12].

Another commonly used psychometric test is the Center for Epidemiologic Studies Depression Scale (CES-D) proposed by Radloff [13]. CES-D comprises 20 items and is rated for ages six and upwards, with a 4-point rating scale. In contrast to BDI, CES -D has an increasing frequency manner for the four answers for each question. Hamilton Rating Scale for Depression (HDRS, HRSD , or HAM -D), proposed by Hamilton [14], is used to detect depression severity in both pre- and post-treatment stages . The test comprises 21 items and is scored with a 5-point or 3-point scale, with the first 17 items being used for scoring.

Out of these psychometric tests, DSM -5 and BDI psychometric tests are important since these two are used to define the annotation criteria for the datasets.

B. Role of NLP, Social Media, and Online Forums in Mental Health Issues Detection

There is an important concept that glues social media, online forums, and mental health issues detection together, namely, self-disclosure . Self-disclosure is the telling of previously unknown information by a person so that it becomes shared knowledge with others, which includes depressed people disclosing their personal feelings and experiences to someone else [15]. This concept is a widely researched area in both psychological and computer-mediated communication literature [16]. Many studies have shown that computer-mediated communication with patients sets them the platform to reveal more symptoms, be honest, and provide candid answers than in traditional face-to-face interviews [17], [18]. In addition, it has also been identified that there is a significant increase in self-disclosure in textual communication rather than audio communication [19]. This particularly emphasizes the importance of NLP techniques in the context of social media sharing networks and online forums to detect mental health issues.

Most studies on depression detection have been conducted on social media platforms compared to online support forum data. Nevertheless, due to the focused nature of a particular topic in online support forums, they tend to be less noisy and credible than social media data. Addressing this gap, the current study aims to investigate online forums such as Reddit, Beyond Blue Online Forum, Understand Forum, 7 Cup Forum, and Depression Forums in this research.

C. Computational Techniques to Detect Depression

Depression detection using online data is an active research field where researchers have used different NLP techniques and text classification methods to automatically detect whether the owner of a given user post exhibits depression or not [1]. In other words, these studies have tackled the issue of depression detection as a binary classification task. Table I outlines the best models in this task based on the data sources used.

D. Computational Techniques to Predict Depression Severity

The field of depression severity levels detection is currently in its early stages of development and there is only a limited amount of literature available. This literature is mainly affiliated with the following datasets.

1) LT-EDI-ACL2022 shared task “Detecting Signs of Depression from Social Media Text” (abbreviated to DepSign-LT-EDI@ACL-2022)[1]
2) Depression Severity in Online Forum Data by Arachchige et al. (abbreviated to DSiOFData) [6].
3) CLEF eRisk workshop series.[2]

Kayalvizhi et al. [20] have produced the dataset for the shared task, “DepSign-LT-EDI@ACL-2022.” The primary source of the dataset is the online forum Reddit. This dataset includes three severity levels for posts: *not depressed*, *moderately depressed*, and *severely depressed*. The highest scores achieved in the shared task, based on metrics recognized by the shared task, are listed in Table II

[1] https://competitions.codalab.org/competitions/36410
[2] https://erisk.irlab.org/2021/index.html

TABLE I
DEPRESSION DETECTION TASKS' BEST RESULTS

Data Source	Features	Methods	Best Results
Reddit	BOW, UMLS	Ada Boost & SVM	F1: 0.75&0.98
	N-grams, topic modeling	MLP	F1: 0.64
	LIWC +N-grams	SVM	Acc.: 0.82
Tweets	LIWC, sentiment, time series	RF	AUC: 0.87
	LIWC, n-grams, topic modeling, sentiment	LR	AUC: 0.85
	N-grams, topic modeling	SVM	Acc. 0.69
	N-grams	Neural Network	AUC: 0.76
Facebook	LIWC, N-grams, topic modeling	LR	AUC: 0.72
LiveJournal	LIWC, topics modeling, mood tags	Regression Models	Acc.: 0.93
Blog posts	TF-IDF, topic modeling, BOW	CNN	Acc.: 0.78

TABLE II
BEST RESULTS OF THE SHARED TASK "DEPSIGN-LT-EDI@ACL-2022" AT LT-EDI-ACL2022

Accuracy	Macro Recall	Macro Precision	Macro F1-score
0.658	0.5912	0.586	0.583

Apart from the shared task "DepSign-LT-EDI@ACL-2022" research contributions, Arachchige et al. [6] have introduced a new corpus for detecting depression severity and proposed two distinct approaches: lexical similarity and n-gram models (AL1) that achieve higher recall and sentence similarity, and pre-trained transformer models (AL2) that achieve higher precision. Arachchige et al. [6] used the aforementioned feature extraction methods and machine learning models to evaluate severity classification tasks. The best results Arachchige et al. achieved are depicted in Table III. Both the best results in Table II and Table III are of great significance for this research as they serve as the baseline models to compare results.

TABLE III
BEST RESULTS FOR THE DSIOFDATA BY ARACHCHIGE ET AL.

Accuracy	Recall	Precision	F1-score
82.15	79.349	80.537	79.938

With respect to the computational techniques to predict depression severity levels, CLEF eRisk workshop series has a notable amount of literature. The research in this area is evaluated using four different metrics: Average Hit Rate (AHR), Average Closeness Rate (ACR), Difference between Overall Depressions Levels (DODL), Average Difference between Overall Depressions Levels (ADODL), and Depression Category Hit Rate (DCHR) [21]. This section presents the most successful research using the aforementioned metrics.

Spartalis et al. [22] achieved an AHR score of 35.46% in 2021 using the MeanPostSVM classifier, which combines feature-based transfer learning with machine learning classification. However, the best overall AHR result was obtained by Burdisso et al. [23], which was 41.43%. They utilized the SS3 classifier, a supervised learning model for text classification, to identify severity levels. The UPV-Symanto team achieved the highest ACR score of 73.17% using a temporal user representation based on the evolution of certain linguistic features over time. They also employed an improved version of RoBERTa, named RoBRoBERTaERTA [21], as a secondary approach.

Wu et al. [24] achieved the highest ADODL score of 83.59% using a deep-learning approach based on the pre-trained model RoBERTa. They also achieved the best DHCR score in 2021, which was 41.25%. On the other hand, Abed-Esfahani et al. [25] achieved the best overall DHCR score of 45% by using the distance between the answers and all of the sentences in the user's writing history.

Notwithstanding the significant research contributions so far in this domain, the existing studies have ignored the importance of augmenting and enriching implicit semantic information using novel advancements in NLP such as chat completion language models. Augmenting and enriching the given textual data with such latent semantic cues is vital to gain a deeper understanding of the context of the user post. Motivated by this, the present study leverages novel implicit semantic technologies in NLP and tailors them to the problem at hand by performing a meticulous feature engineering phase in the machine learning pipeline.

III. DATA SOURCES

A. Primary Dataset 1-"DepSign-LT-EDI@ACL-2022"

Kayalvizhi et al. have developed a gold standard dataset to detect severity levels of depression and released it on the shared task "DepSign-LT-EDI@ACL-2022", [20]. The posts for the dataset was collected from Reddit forums since this online forum platform offers a greater volume of textual data compared to other social media platforms such as Facebook and Twitter. The authors utilized the "pushshift" API to scrape the data from Reddit [20].

The authors selected several subreddits, which are groups of forums focused on specific topics within Reddit, for the data retrieval process. These subreddits include: *r/Mental Health*, *r/depression*, *r/loneliness*, *r/stress* and *r/anxiety*. The target variable of the dataset represents the severity level of depression corresponding to each post. The depression severity levels in this dataset range from 0-2 as outlined below:

- 0 - Not depressed
- 1 - Moderately depressed
- 2 - Severely depressed

TABLE IV
LABEL DISTRIBUTION OF THE PRIMARY DATASET 1

Severity Level	Label Distribution
No depression	2733
Moderate depression	3705
Severe depression	763

A total of 20,088 posts were initially annotated. Among them, the two annotators assigned the same label to 16,613 posts, which were selected to form this dataset. It consists of 156,676 sentences and 26,59,938 words. On average, each post contains 9.42 sentences and 159.2 words. The number of sentences in a post range from 1 to 260, while the number of words ranges from 1 to 5065. Table IV outlines the data points available for each label in the dataset.

B. Primary Dataset 2 - "DSiOFData"

Arachichige et al. developed a gold standard dataset for depression severity detection in 2021 with the primary objective of fulfilling the gap of lack of limited datasets in the field of study. What sets this dataset apart from primary dataset 1 (discussed in Section III-A) is its emphasis on online mental health support forums, which have been specifically designed to address mental health issues such as depression. This enables the present study to evaluate the suitability of the proposed models in such online mental health support forums. Despite the presence of these forums, almost all previous studies in this field of study neglected to explore such in their machine learning models.

Arachchige et al. have utilized the following four online mental health support forums to develop their dataset: Beyond Blue Online Forum[3], Understand Forum[4], Cup Forum[5] and Depression Forums[6]. The data was extracted from these forums using a custom-built web scraper [6].

The label of this dataset represents the severity of depression associated with a given online forum post. This value is derived from the diagnostic criteria established by the DSM-V [26] and BDI [11] psychometric tests. The severity levels of depression, ranging from 0 to 3, are classified into four categories as shown below.

- 0 - No depression
- 1 - Mild depression
- 2 - Moderate depression
- 3 - Severe depression

There are 2293 total posts in this dataset, which are made from 58619 sentences and 1199103 words. This gives an average of 25.56 sentences per post and an average of 522.94 words per post. The labels are distributed as mentioned in Table V.

TABLE V
LABEL DISTRIBUTION OF THE PRIMARY DATASET 2

Severity Level	Label Distribution
No depression	1962
Mild depression	182
Moderate depression	85
Severe depression	64

In order to facilitate a comprehensive analysis of detecting severity levels of depression from online support forum posts, this study utilizes both of these primary datasets with annotated depression severity levels (in contrast to almost all existing literature which only uses a single dataset to validate their computational models). Prior to the construction of the machine learning modes, they went through meticulous cleaning, preprocessing, exploratory data analyzing (EDA), and data imbalance handling processes as described in the subsequent sections.

IV. METHODOLOGY

Detecting the severity of depression from online forum data is a relatively new research topic in the mental health domain. This research is comprised of three tasks based on this problem direction.

1) **Task 1**: Depression detection from online forum data
2) **Task 2**: Prediction of depression severity level from online forum data
3) **Task 3**: Development of a generic prototype interface to extract depression-related symptoms and keywords from a given online forum post and predict the depression severity level of it using models developed for tasks 1 and 2.

Figure 1 depicts the overview of the project methodology and design. Our solution to these tasks is comprised of three feature engineering methods namely, Ada embeddings, GPT 3.5 Turbo, and LIWC. For the primary dataset 1, the feature engineering methods Ada embeddings and GPT 3.5 Turbo were used for the best model. For primary dataset 2, all three feature engineering methods, Ada embeddings, GPT 3.5 Turbo, and LIWC were used for the best model. Additionally, AutoSklearn was able to achieve the best results as the best-performing learning algorithm for both datasets. These best-performing feature engineering methods and the learning algorithm are discussed below.

A. Ada Embeddings

The text-embedding-ada-002 model is a second-generation embedding model developed by OpenAI that has been specifically developed to replace the four previously available first-generation embedding models developed. The model offers significant cost savings and is optimized for text completion, thereby making it suitable for a range of applications such as search, clustering, recommendations, anomaly detection, and text classification tasks. It supports up to 8191 max input tokens and has 1536 output dimensions.

OpenAI recommends utilizing the text-embedding-ada-002 model for embedding tasks, as it is the only second-generation model and offers superior performance compared to first-generation models.

[3] https://www.beyondblue.org.au/
[4] https://www.depression-understood.org/
[5] https://www.7cups.com/
[6] https://www.depressionforums.org/

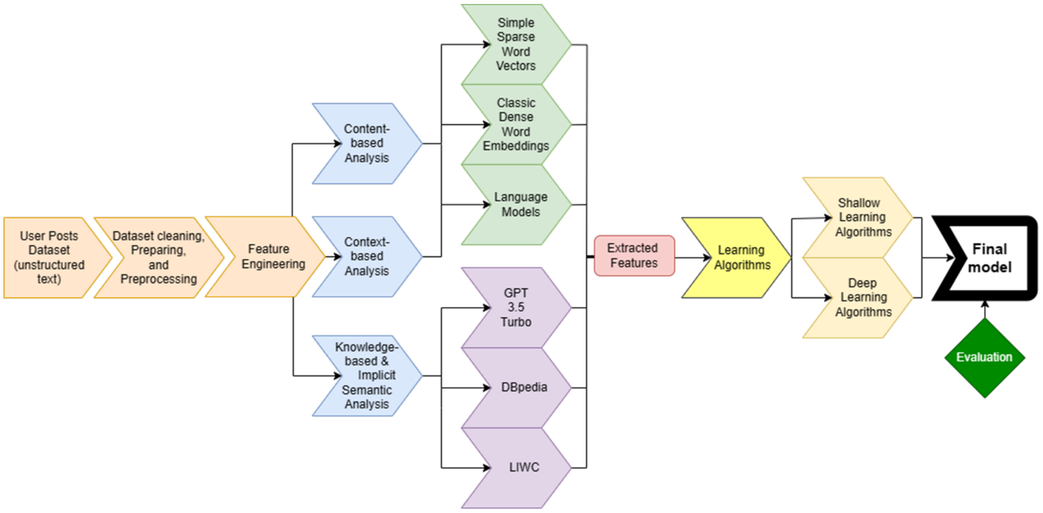

User Posts Dataset (unstructured text)
Dataset cleaning, Preparing, and Preprocessing
Feature Engineering
Content-based Analysis
Context-based Analysis
Knowledge-based & Implicit Semantic Analysis
Simple Sparse Word Vectors
Classic Dense Word Embeddings
Language Models
GPT 3.5 Turbo
DBpedia
LIWC
Extracted Features
Learning Algorithms
Shallow Learning Algorithms
Deep Learning Algorithms
Final model
Evaluation

Fig. 1. Overview of the project methodology and design

B. GPT 3.5 Turbo

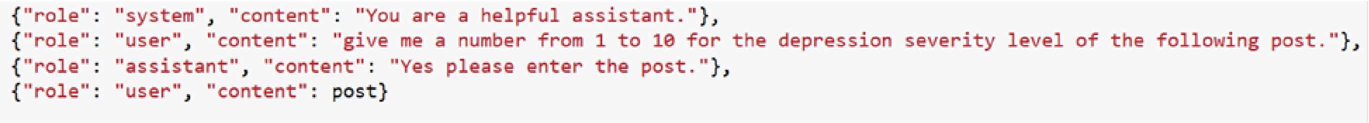
The GPT 3.5 Turbo is an advanced language model developed by OpenAI, which represents an improved version of the GPT-3 model. This model serves as the underlying language model for the widely known chatGPT application and has been specifically tailored for chat-based applications, including conversational agents and chatbots. The model utilizes NLP to generate responses that mimic human-like interactions, resulting in realistic conversations with users.

The GPT 3.5 Turbo has been extensively trained on vast amounts of text data, enabling it to understand the context of a conversation and provide accurate and relevant responses. Due to its sophisticated language abilities, this model, according to OpenAI, excels in many other non-chat use cases in addition to being appropriate for chat-based applications. It's also important to remember that the GPT 3.5 Turbo is a 10x more affordable option than the current GPT 3.5 versions.

In the context of this research, several features were achieved by utilizing a series of steps with the GPT 3.5 Turbo model.

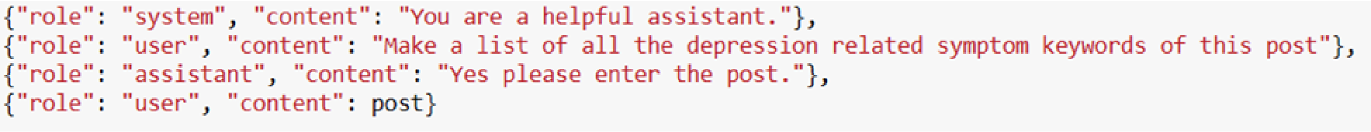
1) **Preparing data:**
In primary dataset 1 which is the 'DepSign-LT-EDI@ACL-2022' dataset, a considerable proportion of the posts were marked with the term "[removed]" at the end to indicate that the respective post had been removed. However, the GPT 3.5 Turbo model does process this token as it is and often refrains from generating any output on behalf of the post. Consequently, the term "[removed]" was eliminated from the dataset, and a novel feature column called "removed" was introduced to signify whether or not the corresponding data point had been removed. This modification enabled more efficient processing of the dataset by the GPT 3.5 Turbo model.
2) **Providing questions:**
A manual testing procedure was applied here by going over 1000 posts with different questions and answers for each post. The comprehensive process led to the identification of two optimal questions that yielded the most valuable information from the posts. The two questions were selected based on their ability to effectively retrieve pertinent information from the corpus.
These two questions are depicted in Figure 2 and 3 respectively.
An example answer for each question is listed below.
Question 1:
- *Based on the content of the post, I would rate the depression severity level as a 9 out of 10. The language used suggests a strong sense of hopelessness, worthlessness, and a lack of social support, which are signs of severe depression. The mention of suicidal thoughts and past suicide attempts also indicates a high level of distress and a need for immediate support and help.*

Question 2:
- *Depression related symptom keywords in this post include: sorrowful, worried, losing my sanity, inescapable reality, sense of meaning, pain, existential pressures. It is important to note that while some of these keywords may be associated with depression, they can also be related to other mental health conditions or simply emotional distress.*
3) **Extracting features from the answers:**
The severity level generated by GPT 3.5 Turbo was found to have a mild correlation with the actual severity label. However, its strength was not sufficient to enable direct usage. Consequently, various techniques were employed to extract new features from the responses. The initially extracted information is presented and elucidated below.
- **Severity Answer:** This refers to the answer provided by GPT 3.5 Turbo to the question 1.
- **Symptom List Answer:** This refers to the answer provided by GPT 3.5 Turbo to the question 2.



```
{"role": "system", "content": "You are a helpful assistant."},
{"role": "user", "content": "give me a number from 1 to 10 for the depression severity level of the following post."},
{"role": "assistant", "content": "Yes please enter the post."},
{"role": "user", "content": post}
```

Fig. 2. Question 1 for GPT 3.5 Turbo model



```
{"role": "system", "content": "You are a helpful assistant."},
{"role": "user", "content": "Make a list of all the depression related symptom keywords of this post"},
{"role": "assistant", "content": "Yes please enter the post."},
{"role": "user", "content": post}
```

Fig. 3. Question 2 for GPT 3.5 Turbo model

- **Extracted Severity Number:** This refers to the severity level number out of 10 suggested in the Severity Answer. For range values, its mean was taken.
- **Extracted Symptom List:** This refers to the extracted list of symptoms from the Symptom List Answer.

Subsequent to the initial extraction of information, a more advanced set of features was derived through the process of inference, filtering, and selection. A comprehensive analysis was conducted on a plethora of features, following which these final features were identified and are presented below along with their explication.

- **Extracted Severity Number:** The severity level number provided by GPT 3.5 Turbo was selected as a final feature.
- **Number of Symptoms:** The number of symptoms presented by the Symptom List Answer.
- **Symptom encode:** This contains 7 features of frequent symptom categories namely: attempted suicide, suicide ideation, medication, disease, self-harm, depression, and anxiety. These categories contain over 40 root symptoms.
- **Severe Keys:** Number of symptoms that adhere to severe level of depression.
- **Moderate Keys:** Number of symptoms that adhere to mild/moderate level of depression.

C. *LIWC*

LIWC (Linguistic Inquiry and Word Count) is a software application developed by researchers at the University of Texas at Austin [27]. It is a computational tool that utilizes a lexicon of words and linguistic features to analyze and classify textual data. LIWC's primary objective is to identify the affective, cognitive, and structural components of language use in written or spoken communication. LIWC assesses words based on more than 70 dimensions, including affect, cognition, social processes, and topical categories, providing researchers and practitioners with invaluable insights into an individual's psychological state or communication style. LIWC has gained widespread acceptance and usage in various domains, including psychology, marketing, and political science, to examine language usage and predict behavioral or attitudinal outcomes.

D. *AutoSklearn*

AutoSklearn was able to achieve the best results as the best-performing learning algorithm for both datasets, For Automated Machine Learning (AutoML), AutoSklearn is an open-source Python package that uses meta-learning to choose the best machine learning algorithms, hyperparameters, and pre-processing methods for a given dataset. The scikit-learn-based library uses Bayesian optimization to find the most effective machine-learning pipeline. Numerous machine learning algorithms, such as support vector machines, decision trees, neural networks, and linear regression, are available in the library. Moreover, it employs meta-learning to learn from the performance of prior models, which guides the search for the best model. The ensemble methods used by AutoSklearn combine the results of multiple models, further enhancing performance.

AutoSklearn is superior to conventional machine learning techniques in a number of ways. It can first automate the selection and fine-tuning of machine learning models, saving time and lowering the possibility of human error. By using meta-learning to learn from the performance of previous models and incorporating this knowledge into the search for the best model, it has the potential to improve machine learning performance.

V. Experiments and Results

A. *Evaluation Plan of Primary Dataset 1*

This section discusses the evaluation plan used to compare and validate the proposed machine learning models in terms of performance metrics and baseline models. The performance metrics used to evaluate the results for the primary dataset 1 in the shared task "DepSign-LT-EDI@ACL-2022" are *Accuracy*, *Macro Precision*, *Macro Recall*, and *Macro F1-Score*. Following the same notion, this study also utilizes these metrics to evaluate the performances of the proposed machine learning models. To facilitate the comparison of the results, this study selects the top-performing model in the shared task "DepSign-LT-EDI@ACL-2022" as the primary baseline model. More specifically, the best-performing model in the shared task was developed by Poswiata et al. [28] and utilizes an ensemble of RoBERTa$_{large}$ and DepRoBERTa models.

To assess and compare the performance of the proposed models in this study with the shared task model results, it is necessary to compute performance metrics using the test dataset provided in the "DepSign-LT-EDI@ACL-2022" shared task. However, the challenge involved with this is that the true labels for the test dataset have not been made publicly available by the shared task. In order to address this challenge and enable comparison with the current best model, the following two main steps are conducted.

1) Implementing the Current Best Model: In order to implement the current best-performing model in the "DepSign-LT-EDI@ACL-2022" shared task, firstly, this study thoroughly reviewed and comprehended the corresponding paper by Poswiata et al. [28], which has details on the model design, methodology, and implementation. Subsequently, this study accessed the publicly available source code for this model and implemented it into our system. During this process, several bugs related to portability were encountered, which were systematically resolved to ensure the proper functioning of this model in our system.

2) Cross Validation: After the development of the best-performing model from the "DepSign-LT-EDI@ACL-2022" shared task, a stratified sampling technique was employed to partition the provided dataset with true labels into separate train, dev, and test splits with 5201, 1000, and 1000 instances respectively.

This process was repeated five times, in order to enable cross-validation of performance while mitigating the effects of bias from a single train-dev-test split.
For each Train-Dev-Test Split out of the 5:

- The 5201 instances of the Train split were used to fine-tune the RoBERTa$_{large}$ and DepRoBERTa models.
- The 1000 instances of the Dev set were used to validate the fine-tuning.
- The 1000 instances of the Test set without labels were used to make the ensemble of the RoBERTa$_{large}$ and DepRoBERTa models predict the severity levels.
- The predicted severity levels were compared against the true severity level label to get the performance metrics (Accuracy, Macro Precision, Macro Recall, Macro F1-score) for each Train-Dev-Test split.

The results of each Train-Dev-Test split were aggregated together using simple mean. The aggregated results are depicted in the comparison Table VI presented under results discussion and these results are compared in the same table with our best-performing model that underwent the same cross-validation process for the five Train-Dev-Test splits.

B. Evaluation Plan for Primary Dataset 2

As mentioned in Section V-A the evaluation plan of a primary dataset, has two aspects to consider; namely, metrics and the baseline model. The performance metrics used to evaluate the results for the primary dataset 2 as published by the original paper are *Accuracy*, *Precision*, *Recall*, and *F1-Score* [6]. Following the same notion, the present study uses these metrics to validate and compare results for this dataset.

The baseline model selected for this dataset is the best-performing model utilized in the original study titled "A Dataset for Research on Modelling Depression Severity in Online Forum Data" by Arachchige et al. [6]. The original paper employed various algorithms for multiclass prediction using the annotated dataset. In the original study, the best results for the multiclass classification task utilizes unigram-count with neural networks (Table III).

The present study incorporates 5-fold cross-validation to our selected best model to minimize the bias and aims to compare the final results of Table III with the cross-validated outcomes obtained from our best-performing model.

C. Experiments

The present study employs novel AI-based models based on a vast amount of feature engineering methods and learning algorithms. As feature engineering methods the present study utilized simple sparse word vectors, classic dense word embeddings, language models, and other tools such as LIWC and DBpedia. As learning algorithms, the present study utilized both shallow and deep learning algorithms. Otherwise stated, this research has developed novel, comprehensive, and diverse sets of experiments with the use of different experimental settings with the intention of facilitating a thorough as well as nuanced analysis in the discipline.

D. Results Discussion of Primary Dataset 1

The best-performing proposed AI-based model for primary dataset 1 is obtained by using Ada embedding with GPT 3.5 Turbo feature engineering methods and AutoSklearn ensemble learning algorithm. To compare the performances of the proposed AI-based models with the state-of-the-art results, the best-performing model from the "LT-EDI@ACL-2022" shared task [28] was implemented and validated using the same cross-validation splits, as described in Section V-A. Table VI depicts results obtained for our best-performing model and the "LT-EDI@ACL-2022" shared task best-performing model. In addition, the other top results of the shared task from places 2-5 have also been listed in Table VI for comparison purposes.

As shown in Table VI, the proposed model from this research that uses Ada embedding with GPT 3.5 Turbo (ie., the experiment with the experiment ID V7) has outperformed the best-performing model from the 'LT-EDI@ACL-2022' shared task by 5.5%. The improvements of our model over the best-performing shared task model, in terms of macro precision, macro recall, and accuracy are 8.4%, 3.9%, and 1.9% respectively. The improved performance of our model clearly showcases that detailed implicit semantic inferences are crucial to enhance the prediction performances of depression severity levels (in contrast to the 'LT-EDI@ACL-2022' shared task models). To the best of our knowledge, this is the very first study in the discipline that uses chat completion language models to deduce depression severity levels. In other words, to the best of our knowledge, none of the previous studies have utilized chat completion language models to infer a comprehensive overview of implicit semantics in the textual

TABLE VI
RESULTS COMPARISON TABLE FOR PRIMARY DATASET 1

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score
Our Best Model	0.663000	0.656011	0.623844	0.632101
Shared Task Best Model	0.644000	0.571602	0.584719	0.576699
Shared task 2^{nd} Place Model	0.6330	0.5732	0.5394	0.5523
Shared task 3^{rd} Place Model	0.6253	0.5720	0.5303	0.5467
Shared task 4^{th} Place Model	0.6250	0.5806	0.5218	0.5426
Shared task 5^{th} Place Model	0.6179	0.5704	0.5263	0.5422

data. Such deep semantic inferences are vital to gain a better understanding of the background and context of the text, which is crucial for complex NLP tasks such as depression severity level detection. The 'LT-EDI@ACL-2022' shared task results, which are limited in incorporating such additional implicit semantic inferences may have impacted their low prediction performances over the models proposed by this research. More specifically, the strengths of this proposed model are as follows.

- **Capturing comprehensive implicit semantic details:** The proposed AI-based model attempts to capture a holistic view of the given textual data by augmenting and enriching additional implicit semantic details related to depression. This may enable the opportunity for the AI-based model to analyze critical depression-related implicit semantic cues that may provide valuable insights to predict the depression severity level label.
- **Co-modeling complementary semantic models:** This study observed a significantly high-performance gain with the complementary integration of Ada embedding with GPT 3.5 Turbo. In other words, even though Ada embedding provided a better performance improvement than the state-of-the-art results, by co-modeling it with GPT 3.5 Turbo, this study demonstrated the best prediction performance. This indicates the cruciality of having multiple semantic models that elicit implicit details of the user posts in the problem of depression severity level detection.
- **Additional semantic inferences in postprocessing:** In the proposed best-performing AI-based model, this study also performs an additional semantic inference phase in postprocessing to improve the representation of the input to the machine learning model (a detailed explanation is provided in Section IV-B). This indicates that the semantic inferences that are tailored to depression, as well as a better representation of the extracted features, are critical in the machine learning workflow to obtain high performance.

Due to the significant improvement of the results (as shown in Table VI), we have enrolled in the 2023 version of the same shared task, namely, "Detecting Signs of Depression from Social Media Text - LT-EDI@RANLP 2023"[7].

E. Results Discussion of Primary Dataset 2

The best-performing model for this dataset was obtained by utilizing Ada embedding, GPT 3.5 Turbo, and LIWC feature engineering methods (discussed in Sections IV-A, IV-B, and IV-C respectively). The above-discussed feature extraction methods were combined with AutoSklearn (discussed in Section IV-D) to build the proposed ensemble machine-learning model. The strengths of this proposed AI-based model are comparable to the best-performing model for the primary dataset 1, as discussed in Section V-D. To summarize, these strengths are utilizing a comprehensive feature extraction process by augmenting and enriching implicit semantic information about depression to obtain a deeper understanding of the context of the user posts alongside complementary integration of multiple semantic models (i.e., Ada embedding and GPT Turbo 3.5), and additional postprocessing of the semantic details to enhance the machine learning input representation. In addition to the primary dataset 1 strengths, another strength observed from this model is its use of extra details on a given user post such as affective, cognitive, and structural components of language use by using the LIWC tool.

F. Best Performing Models in Primary Datasets

In this section, the best-performing proposed AI-based models for both primary datasets are analyzed to identify which methodologies work best in multiple data sources. In other words, the primary datasets used for this research include data from different sources, as described in Section III. To summarize, the data source used for the primary dataset 1 is Reddit online forum [20], whereas the data sources used for the primary dataset 2 are online support forums [6]. Due to the differences in the data sources used in these two datasets, if a proposed AI-based model in this research performs well in both datasets, it illustrates the potential reuse of that model/methodology to diverse mediums of user posts, where the writing styles may vary. To facilitate this analysis, the top two best-performing models of both datasets were selected. To summarize, the best-performing models of the two datasets are as follows.

- **Primary Dataset 1**
  - **Best Performing Model:** Ada embedding + GPT 3.5 Turbo
  - **Second Best Performing Model:** Ada embedding
- **Primary Dataset 2**
  - **Best Performing Model:** Ada embedding + GPT 3.5 Turbo + LIWC
  - **Second Best Performing Model:** Ada embedding + GPT 3.5 Turbo

[7] https://codalab.lisn.upsaclay.fr/competitions/11075

To achieve the above-discussed improved prediction performances, our best-performing models leverage the following three powerful feature engineering methods.

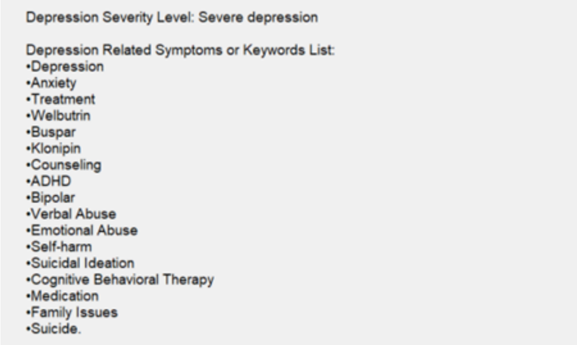
1) **Ada embeddings**
To summarize, this text embedding method captures certain implicit semantic relationships of textual data in addition to explicit content-related features and constructs a dense vector representation in a multi-dimensional vector space. In this vector space, posts that share higher semantic similarity are likely to reside in close proximity adhering to the distributional hypotheses. The dimensionality of the vector embeddings used for this research to represent a given user post is 1536, as discussed in detail in Section IV-A. The experimental results indicate that this is a very powerful text embedding model in the domain of depression severity level detection, as the use of this text embedding method alone was able to achieve significantly improved results over both baseline models in each dataset.
2) **GPT 3.5 Turbo**
The novel methodology that this research proposed by leveraging GPT 3.5 Turbo plays a significant role in augmenting additional implicit semantic details to a given user post (a detailed explanation is provided in Section IV-B). To the best of our knowledge, no prior research has used such kinds of chat completion feature extraction methods to obtain a detailed picture of additional implicit semantic inferences related to the user posts. The complementary integration of these augmented additional implicit semantic features with Ada embeddings has enabled us to obtain a thorough understanding of the implicit semantics in the textual data, which may be the reason for its significantly high-performance gain over the existing state-of-the-art results.
3) **LIWC**
In addition to the Ada embedding and GPT 3.5 Turbo, the best-performing model for primary dataset 2 also utilizes LIWC features. LIWC is a proprietary tool that specializes in analyzing textual data, as discussed in Section IV-C. In short, LIWC's primary objective is to identify the affective, cognitive, and structural components of language use in written or spoken communication. This proposed AI-based model assesses words in a given user post by utilizing LIWC-based features on 117 dimensions, including details such as affect, cognition, health, social processes, and psychological aspects.

One prominent observation is that the model involving Ada embedding + GPT 3.5 Turbo has performed well in both datasets. This indicates the potential reuse nature of this proposed model in diverse data sources such as Reddit, Beyond Blue, 7 Cup, Understand, and Depression online forums.

Another observation of this analysis is the majority of these top AI-based models share methodologies involving Ada embedding and GPT 3.5 Turbo, which indicates the efficacy of integrating multiple semantic models to perform nuanced implicit semantic analysis.

G. Software Prototype

As a practical byproduct, this study has employed an auxiliary task of developing a software prototype to assess depression severity levels. The working of prototype is that it takes a user post of an online forum or even a new post describing a patient's situation as input and will output identified depression severity level alongside depression-related symptoms or keywords identified through that post. An example is shown in Figure 4.



Depression Severity Level: Severe depression

Depression Related Symptoms or Keywords List:
•Depression
•Anxiety
•Treatment
•Welbutrin
•Buspar
•Klonipin
•Counseling
•ADHD
•Bipolar
•Verbal Abuse
•Emotional Abuse
•Self-harm
•Suicidal Ideation
•Cognitive Behavioral Therapy
•Medication
•Family Issues
•Suicide.

Fig. 4. Software prototype example

H. Major Contributions

Through this research, to the best of our knowledge, the study was able to shed light on several novel areas in the depression severity levels detection problem domain. These major contributions are summarized below and discussed in detail in Section IV.

- Being the very first study in the depression severity levels detection discipline to perform a circumstantial implicit semantic inferences component by leveraging chat completion language models, in order to analyze fine-grained depression-related implicit semantic cues beyond the content given in the user posts.
- Demonstrating the importance of co-modeling multiple semantic models through the identification of complementary, cutting-edge NLP techniques that enhances the prediction performances of the AI-based models.
- Showcasing the importance of enriching and tailoring the semantic information to the problem at hand by performing a postprocessing phase to enhance the semantic details representation of the machine learning input.
- The experimental results demonstrate a significant improvement of the results of the proposed AI-based models than the existing state-of-the-art results in predicting the severity levels of depression.
- The best performing proposed AI-based models have the potentiality of reusing in various data sources, enabling the opportunity to provide wider community benefits.
- Developing a software prototype as part of this research, enabling the opportunity to perform user-based diagnosis

using the proposed best-performing AI-based model by integrating it into online forums.

VI. Conclusion & Future Work

To perform a comprehensive analysis, this study utilizes two primary datasets with annotated depression severity levels, in contrast to previous literature that merely uses a single dataset to validate their models. Overall, the experimental results demonstrate significant improvements in the prediction performances of the proposed models than the state-of-the-art results. To the best of our knowledge, this is the first study in the discipline that performs a circumstantial implicit semantic inferences phase by leveraging chat completion language models. The purpose of using chat completion language models is to augment and enrich depression-related background details to capture implicit semantic cues beyond the content written by the user. The results are evidence that such detailed semantic augmentation and enrichment are crucial to gain a deep understanding of the textual data.

This paper also demonstrates the importance of co-modeling multiple semantic models in order to elicit semantic details in diverse layers. Additionally, this study showcases the importance of postprocessing to enrich and tailor the derived semantic information to better represent semantic details in the machine learning input. Another important observation is that our best-performing model showcases its potential reuse nature in various data sources, enabling it the opportunity to provide wider community benefits. The main reason for the reuse nature may be due to its reliance on the additional implicit semantic cues that have been augmented and enriched using chat completion language models than merely relying on the user content. To the best of our knowledge, such reusable analyses have never been performed in the discipline. As a practical byproduct, the present study has implemented a software prototype application to predict depression severity levels of user-inputted textual data.

In the future, this study intends to be extended by tracking the depression severity progression of a user by chronologically analyzing user posts by leveraging diachronic word embeddings along with the proposed AI-based models of this research. This will enable the moderators in online support forums to monitor how the depression severity levels of individuals have evolved over time.

References

[1] 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.

[2] R. F. Muñoz, W. R. Beardslee, and Y. Leykin, "Major depression can be prevented," *American Psychologist*, vol. 67, pp. 285–295, 2012. Place: US Publisher: American Psychological Association.

[3] P. Adekkanattu, E. T. Sholle, J. DeFerio, S. B. Johnson, and T. R. Campion, "Ascertaining Depression Severity by Extracting Patient Health Questionnaire-9 (PHQ-9) Scores from Clinical Notes," p. 10.

[4] S. Paul, J. Kalyani, and T. Basu, "Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks," Aug. 2018.

[5] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, no. 1, pp. 128–137, 2013. Number: 1.

[6] I. A. Nanomi Arachchige, R. Weerasinghe, University of Colombo School of Computing, Sri Lanka, V. H. Jayasuriya, and University of Kelaniya, Sri Lanka, "A Dataset for Research on Modelling Depression Severity in Online Forum Data," pp. 144–153, 2021.

[7] R. Martınez-Castano, A. Htait, L. Azzopardi, and Y. Moshfeghi, "Early risk detection of self-harm and depression severity using BERT-based transformers," p. 16.

[8] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and Classification of mental illnesses on social media using RoBERTa," Nov. 2020. arXiv:2011.11226 [cs].

[9] M. P. Hage and S. T. Azar, "The Link between Thyroid Function and Depression," *Journal of Thyroid Research*, vol. 2012, p. e590648, Dec. 2011. Publisher: Hindawi.

[10] A. P. Association, *Depressive Disorders: DSM-5® Selections*. American Psychiatric Pub, July 2015. Google-Books-ID: g4cJDAAAQBAJ.

[11] A. T. BECK, C. H. WARD, M. MENDELSON, J. MOCK, and J. ERBAUGH, "An Inventory for Measuring Depression," *Archives of General Psychiatry*, vol. 4, pp. 561–571, June 1961.

[12] M. Kovács, "Rating scales to assess depression in school-aged children.," *Acta paedopsychiatrica*, 1981.

[13] L. S. Radloff, "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population," *Applied Psychological Measurement*, vol. 1, pp. 385–401, June 1977. Publisher: SAGE Publications Inc.

[14] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, pp. 56–62, Feb. 1960.

[15] A. N. Joinson and C. B. Paine, *Self-disclosure, Privacy and the Internet*, vol. 1. Oxford University Press, Sept. 2012.

[16] S. Balani and M. De Choudhury, "Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, (Seoul Republic of Korea), pp. 1373–1378, ACM, Apr. 2015.

[17] J. H. Greist, M. H. Klein, and L. J. Van Cura, "A computer interview for psychiatric patient target symptoms," *Archives of General Psychiatry*, vol. 29, pp. 247–253, Aug. 1973.

[18] M. Ferriter, "Computer Aided Interviewing in Psychiatric Social Work," *Computers in Human Services*, vol. 9, pp. 59–66, Apr. 1993. Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1300/J407v09n01_09.

[19] A. N. Joinson, "Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity," *European Journal of Social Psychology*, vol. 31, pp. 177–192, Mar. 2001.

[20] K. S and T. D, "Data set creation and empirical analysis for detecting signs of depression from social media postings," Feb. 2022. arXiv:2202.03047 [cs].

[21] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview)," p. 24.

[22] C. Spartalis, G. Drosatos, and A. Arampatzis, "Transfer Learning for Automated Responses to the BDI Questionnaire," p. 13.

[23] S. G. Burdisso, M. Errecalde, and M. Montes-y Gomez, "UNSL at eRisk 2019: a Unified Approach for Anorexia, Self-harm and Depression Detection in Social Media," p. 18.

[24] S.-H. Wu and Z.-J. Qiu, "A RoBERTa-based model on measuring the severity of the signs of depression," p. 10.

[25] P. Abed-Esfahani, D. Howard, M. Maslej, S. Patel, V. Mann, S. Goegan, and L. French, "Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings," p. 9.

[26] A. P. Association and A. P. Association, eds., *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, D.C: American Psychiatric Association, 5th ed ed., 2013.

[27] R. Boyd, A. Ashokkumar, S. Seraj, and J. Pennebaker, *The Development and Psychometric Properties of LIWC-22*. Feb. 2022.

[28] R. Poświata and M. Pere\lkiewicz, "OPI@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pretrained Language Models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, (Dublin, Ireland), pp. 276–282, Association for Computational Linguistics, May 2022.