DEEP KNOWLEDGE-INFUSION FOR EXPLAINABLE DEPRESSION DETECTION

A PREPRINT

Sumit Dalal¹, Sarika Jain¹, and Mayank Dave¹
¹National Institute of Technology Kurukshetra, India

September 5, 2024

ABSTRACT

Discovering individuals' depression on social media has become increasingly important. Researchers employed ML/DL or lexicon-based methods for automated depression detection. Lexiconbased methods, explainable and easy to implement, match words from user posts in a depression dictionary without considering contexts, and little attention has been paid to how the word can be associated with the depression-related context. While the DL models can leverage contextual information, their black-box nature limits their adoption in the domain. Though surrogate models like LIME and SHAP can produce explanations for DL models, the explanations are suitable for the developer and of limited use to the end user. We propose a Knolwedge-infused Neural Network (KiNN) incorporating domain-specific knowledge from DepressionFeature ontology (DFO) in a neural network to endow the model with user-level explainability regarding concepts and processes the clinician understands. Further, commonsense knowledge from the Commonsense Transformer (COMET) trained on ATOMIC is also infused to consider the generic emotional aspects of user posts in depression detection. The model is evaluated on three expertly curated datasets related to depression. We observed the model to have a statistically significant (p<0.1) boost in performance over the best domain-specific model, MentalBERT, across CLEF e-Risk (25% MCC[↑], 12%F1[↑]). A similar trend is observed across the PRIMATE dataset where the proposed model performed better than MentalBERT (2.5% MCC\(\gamma\), 19%F1\(\gamma\)). The observations confirm the generated explanations to be informative for MHPs compared to post-hoc model explanations. Results demonstrated that the user-level explainability of KiNN also surpasses the performance of baseline models and can provide explanations where other baselines fall short. Infusing the domain and commonsense knowledge in KiNN enhances the ability of models like GPT-3.5 to generate application-relevant explanations.

Keywords A

ttention, Depression, Explainability, Neural network, Ontology, and Social media

1 Introduction

Large population suffering from various mental disorders join social communities, a popular means of online communication for sharing and helping others. People tend to write daily posts covering feelings, physical moments, food habits, exercise, and music choices. Information from these posts is considered for the user's mental health assessment. Researchers observed signs of depression in user's social data well before their first diagnosis. Moreover, the social data can be collected non-intrusively. These advantages help health professionals detect the user's mental health without much interference in the user's life.

ML/DL or lexicon-based approaches are manipulated to analyze the big data generated on social media platforms for depression detection. In lexicon-based approaches, users' social posts are searched for terms from specific dictionaries (like antidepressant or depression-related). If the frequency of the terms from these dictionaries in the user's posts crosses a threshold, the user is assigned a depressed label. While applying the lexicon-based method has been known to be explainable and easy to implement [1, 2]. Their possible disadvantage, however, is that they only consider if

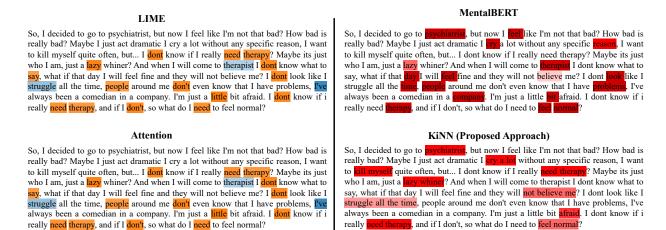


Figure 1: Comparision of explanations produced by LIME, MentalBERT, Self-Attention, and the proposed KiNN model. The objective is to enhance the ability to provide user-level explanations for depression detection. KiNN highlights variable-length phrases centered around depression, unlike others, which focus on unigrams lacking context for user understand-ability.

terms in user posts match the depression lexicon and don't consider how a word can be associated with the depression-related context. For example, Figure 1 illustrates two sentences: "I cut my wrist" and "I have my hair cut." Assuming that the word 'cut' belongs to the depression dictionary, only the former sentence should be evaluated as having depression. However, the latter sentence could also have depression if the methods of prior work [3, 4] are applied. In other words, if the context is incorrectly captured, a model using a depression lexicon created by experts may not be able to accurately assess the risk of depression [5].

While the ML/DL models can leverage the contextual information, their black-box nature limits adoption in the domain [6]. They fail to provide explanations grounded in the knowledge that aligns with MHPs [7] and depend on surrogate models like LIME and SHAP to explain the decisions [8, 9]. However, these system-level explanations are helpful to developers and of limited use to the end user.

Expert-level explainability can be defined as the connections between an AI model's collective experiences from training and the real-world entities and definitions that make sense to domain experts [10]. Consider the following expression, "For the past several weeks, I have no to little interest to write my life any better than it is at the moment." is an example of "little interest or pleasure in doing things" which is the first question in Patient Health Questionnaire (PHQ-9), which is a clinical instrument for measuring the severity of depression. Phrases like for the past several weeks, no to little interest are MHP-explainable concepts. If a user's post containing the above-mentioned sentence is classified as **depressed** by an AI model, then such phrases should be the focus. In this direction, Explainable AI has gained attention in natural language processing (NLP) for mental health, as it provides "explanations" for the BlackBox model's decisions [11]. The explanations in LIME and SHAP are obtained through training an interpretable classifier that matches the BlackBox's outputs. Examples of such BlackBox AI models are BERT, RoBERTa, Longformer, and other self-attention-based language models [12], [13], [14]. The attention visualizations are limited to MHPs and require post-hoc explainability techniques, which have other issues [15]. Supporting BlackBox models with clinical knowledge, like PHQ-9, Diagnostic and Statistical Manual for Mental Health Disorders (DSM-5), which MHPs often use, would result in models capable of delivering *user-level explanations*. Alternatively, recent studies have demonstrated the importance of clinical knowledge and expertise in creating labeled datasets that improve the quality of explanations from BlackBox models [11], [16]. However, constructing such datasets with infused knowledge has issues with quality (e.g., agreement scores between MHPs), transferability, cost, and effort. Figure 1 displays significant features by various approaches considered for generating explanations. Existing methods (LIME, Mental-BERT, and Attention) highlight unigrams most of which have low domain-specific meaning. However, the proposed approach can identify the domain-specific and bigger grams (bi or tri) compared to the previous ones. Bigger grams have more significance than smaller ones as they contain more contextual information. For domain-specific results, background knowledge is introduced in terms of the domain ontology.

Knowledge infusion is divided into three levels: a.) Shallow Infusion b.) Semi-Deep Infusion c.) Deep Infusion. In shallow infusion, the knowledge and the infusion method are shallow as syntactic and lexical knowledge is considered through word embedding models. While in semi-deep infusion, external knowledge is involved through attention

mechanisms or learnable knowledge constraints acting as a sentinel to guide model learning. Further, in deep infusion, a stratified representation of knowledge representing different levels of abstraction is employed in different layers of a deep learning model to transfer knowledge that aligns with the corresponding layer in the layered learning process.

Our main contributions are as follows:

- Machine learning and semantic technologies driven hybrid model is proposed for depression detection on the social media data.
- 2. Mental health-related contextual information is captured with domain-centric vector space representation of the n-grams and phrases in user posts.
- 3. Semantic information is retrieved from the advanced FeatureOnto ontology that adds depression-related conceptual phrases to the existing ontology version.
- 4. Classification results of the proposed approach are compared with state-of-the-art machine learning and deep learning methods.

2 Literature

2.1 Knowledge-infusion in Depression Detection

Various aspects of a social media user, along with the user's social postings, have been utilized in detecting depression, for example, the social media user's profiles and his/her behavior on the social platform [17, 18]. The downside is that these models are trained on several irrelevant contents, which might not be crucial for detecting a depressed user. Besides, this content hurts the overall efficiency and effectiveness of the model. To improve the model performance, authors infused knowledge at various levels using dictionaries/lexicons, which were reported to be effective in capturing specific linguistic or domain-specific characteristics to detect depression [19, 20]. The authors of [21, 20, 22] mapped the words from users' textual posts into Lexicons like Linguistic Inquiry and Word Count (LIWC) [23], Affective norms for English words (ANEW) [24], and NRC [25] for infusing linguistic knowledge in the classification process. While the authors of [26, 27, 28, 29, 19, 20] focussed on domain-specific knowledge like antidepressants. Few other works which considered domain-specific knowledge like exercise-related information are [30], and disease or symptoms-related information are [31, 32, 28, 29].

2.2 Explainable Depression Detection

Sometimes, the explanation is confused with the interpretation work; however, both are closely related but not the same. Like in [33], authors explored and interpreted multi-variate time series features for depressed versus non-depressed users in depression detection. For this purpose, they exploited the descriptive text of the depression symptoms to check the similarity index between each user post and the symptoms to calculate different time series features for a user. Here, the time-series features are interpreted, but the reason for the classification output is not produced.

Authors of [34] provided a glimpse of the role of attention weights in explaining depression classification results by hierarchical attention neural model. In [27], multiple features besides the similarity index of user posts with depression symptoms and antidepressants have been considered for depression detection from Reddit posts. The relation between the features and the representations learned by the post-level attention network is studied to interpret the depression classification results. Similarly, the authors of [35] emphasized post-level attention weights from hierarchical attention networks for producing depression classification explanations. The authors in [36] attempted to generate explanations using representation learned by attention mechanism only on the words in a user post, with accompanying limitations. Also, in [37], word-level attention representations from hierarchical attention networks are inspected to analyze unigrams and bigrams relevant for classification. However, the authors of [38] visualized the significant words and posts based on the attention weights from a hierarchical attention network of word and post levels to explain the classification results. They employed multiple features, including symptoms and antidepressant count, in depression detection from social media posts. Though a BiLSTM-CNN with Attention Network is provided for depression classification in [39], they considered the word frequency of depressive and non-depressive samples from the training set to find feature-level explanations. These works focus on attention weights, but authors in [40] designed a new classifier. They updated the SS-3 classifier to accommodate grams of various lengths instead of uni-grams only from training samples for depression detection. The classifier works on a decision tree where n-gram nodes are given some weight depending on their participation in the classification process. All works mentioned here are considered user posts for explanation except the work in [41]. Numerical features like the number of followers, positive/negative emotion words, and posts have been deployed for explanation generation by TreeSHAP, a game approach for explaining decision tree outputs.

Table 1: Symbols Table

Symbol Abbreviation				
u_i	ith social media user's posts			
z_{u_i}	contextualized representation of <i>ith</i> user's posts			
	returned by KiNN			
$\frac{u_i^{pt}}{\hat{u}_i^{pt}}$	phrase tagged posts from ith user			
\hat{u}_i^{pt}	encoded phrase tagged posts			
\overline{W}	Writer or Speaker of the post/comment			
L	Listeners (others)			
E_{IW}	Intent of W			
E_{EW}	Effect on W			
E_{RW}	Reaction of W			
E_{EL}	Effect on L			
E_{RL}	Reaction on L			
W	Noun token			
l	Noumber of noun tokens in user post u_i			

From the literature, we observed that only a few authors focussed on explanation generation regarding depression classification output. Some worked on redesigning the depression classifier to produce explainable output, while others concentrated on attention networks for explanation generation. However, none of these produced the user-level explanations our model tends to do. Moreover, the lexicons used for knowledge infusion lack domain-specific information incorporated at shallow or semi-deep levels. Further, prior work focused on matching a word in a post with a depression dictionary without considering contexts, and little attention has been paid to how the word can be associated with the depression-related context.

3 Knowledge-infused Neural Network (KiNN)

3.1 Architecture

KiNN is a deep neural model with domain and commonsense knowledge infusion through multiple dense and attention layers. KiNN has a unique aspect of three-stage knowledge infusion: shallow infusion via domain-specific embeddings, semi-deep infusion via knowledge incorporation during neural network weight learning, and deep infusion via multiple layers of knowledge introduction through attention mechanism at various levels. The attention layers enable the model to compute the significance score of words/phrases from user posts compared to the CPGs (here DepressionFeature and UMLS) and sentimental aspects (like COMET). It has three self-attention layers as shown in Figure 2 which focuses on depression and sentiments. The first two layers independently find attention scores of phrase-tagged user posts related to depression concepts from DepressionFeature ontology (DFO) and emotional aspects from COMET. The third attention layer takes the concatenated output from the two and decides upon the importance of words/phrases compared to the depression and emotional aspects cumulatively.

KiNN is supported by two external modules: *Knowledge Infusion from Domain-specific Knowledge Graphs*¹ and *Explanation Visualizer*. Figure 2 shows the architecture of KiNN with DepressionFeature Ontology and COMET transformer. User posts collected from social platforms, Reddit in this case, have been passed to the *Phrase Tagging* module. The module employs the DepressionFeature Ontology to tag the domain-specific phrases from user posts.

MentalBERT [42] is a BERT-based large language model pre-trained on mental health-specific data from social media platforms. BERT tokenizer, which considers words or sub-words as tokens, not phrases. Moreover, depending on the context, a single word could have more than one representation in the dimensional space. Hence, to calculate a fixed vector, phrases are considered to be present standalone, and representation from the last four layers of the encoder was added and averaged. Phrases are considered over words as tokens because phrase tagging preserves the meaning of sentences following the transformer tokenization process [43].

The **DepressionFeature Ontology** (**DFO**) is a structured person and disease-specific data model to detect and monitor mental disorders in social media users'. It contains distinct features from multi-modal data regarding various mental disorders [44, 45]—for example, depression-related text phrases or image characteristics from social media users' postings. DepressionFeature ontology is populated with significant topics and static and dynamic-length phrases

¹The current architecture of KiNN uses two domain-specific knowledge graphs. Other forms of knowledge can also be explored.

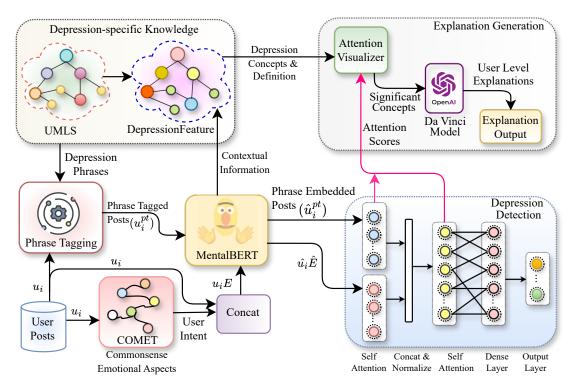


Figure 2: Knowledge-infused Neural Network (KiNN) model considering different aspects of user posts-domain and commonsense for depression detection. MentalBERT provides context-specific embeddings for phrase-tagged user posts. COMET provides nine aspects related to user posts, five of which are mental-health related and hence considered to check the intent of user posts. KiNN allows visualization of various attention layers for user-level explanations.

(uni, bi, tri, or quad grams) across all depressive posts in the PRIMATE, CLEF e-Risk, and CAMS datasets. Topics have been extracted through unsupervised LDA, while for phrases, besides the NLTK library's "ngram" function, they deployed three neural keyphrase extraction methods. These methods are KeyBERT [46], KeyBART [47], and KeyBERT incorporating part-of-speech tagging and TF-IDF [48]. DepressionFeature ontology also reuses instances from the Patient Health Questionnaire-9 Depression Ontology (PHQ-9 DO) [49]. PHQ-9 DO has been developed to produce concept-level and user-understandable explanations where each category (or class in ontological terminology) represents a question from the PHQ-9.

This work complemented the DepressionFeature ontology with another biomedical knowledge source, the **unified medical language system** (**UMLS**), for expert-curated synonyms and depression-related definitions of the phrases [50]. Further, we have considered the **MentalBERT** [42] to find similar (synonym or related) phrases used by social media users. Phrases in UMLS and MentalBERT are checked for cosine similarity with phrases from the ontology, and a threshold of cosine score of =>0.80 is considered to find the related or synonym phrases. All these resources (DepressionFeature ontology, UMLS, and MentalBERT) help identify and tag depressive phrases in user posts.

COMET [51], a generative transformer model trained on ATOMIC [52] is employed to extract everyday inferential if-then emotional aspects from user posts. This knowledge is present in the form of textual descriptions mainly focused on the speaker and listeners of the input. There are total nine different if-then aspects: i) the intent of W, ii) the need of W, iii) attribute of W, iv) effect on W, v) wanted by W, vi) the reaction of W, vii) effect on others, viii) wanted by others, ix) the reaction of others. Here W is the writer (or speaker of the post/comment) while others are listeners. As an example, with the input "Person W gives him a compliment," the intent of W would be W wanted to be nice. Following the conviction of [52, 53, 54] who utilized these aspects in emotion and mental health-related tasks, we selected five aspects: intent of W (E_{IW}), effect on W (E_{EW}), reaction of W (E_{RW}), effect on others (E_{EL}), and reaction of others (E_{RL}). Considering the close relation between emotions and mental states, this selection is expected to be appropriate.

Explanation Generation visualizes the attention patterns in KiNN against each attention block. The attention weights reflect the model's internal representation of the relationships. KiNN also represents dependencies between different words/phrases in the input and depression or sentimental aspects in the attention blocks. The visualizer identifies key parts of the input text regarding how the model processed the information and performed *depression classification*.

Algorithm 1 KiNN - Training Loop

```
1: I, E, R, W, and L stands for Intent, Effect, Reaction, Writer, and Listener
 2: while loop over batch data samples (u_i, y_i) do
                                                                                         \triangleright (u_i, y_i) is input pair indexed by i
 3:
        Set input u_i
                                                                                                  > tokenized user document
        obtain u_i^{pt}
                                ⊳ input after mapping to the Phrase Lexicon (i.e. DepressionFeature) (Run algorithm 2)
 4:
        obtain \hat{u}_i^{'pt} = EL(u_i^{pt})
 5:
                                                                      ▷ EL represents embedding layer (i.e. MentalBERT)
        for k, where k \leftarrow 1 to K do
                                                                           \triangleright K is no. of commonsense aspects in COMET
 6:
            E_{I/E/R,W/L} = CL(u_i)
                                                                       ▷ CL represents commonsense layer from COMET
 7:
        obtain u_i E = Con(u_i E_{IW} E_{EW} E_{RW} E_{EL} E_{RL})
                                                                                       > Con represents concatenation layer
 8:
        obtain \hat{u}_i \hat{E} = EL(u_i E)
 9:
                                                                      ▷ EL represents embedding layer (i.e. MentalBERT)
        obtain S_{\hat{u_i}^{pt}} = SA(\hat{u}_i^{pt})
10:
                                                                                                       obtain S_{\hat{u}_i\hat{E}} = SA(\hat{u}_i\hat{E})

⊳ Self-Attention layer

11:
        obtain h_i = LN(Con(S_{\hat{u_i}^{pt}}S_{\hat{u_i}\hat{E}}))
12:
                                                                                  obtain S_{h_i} = SA(h_i)
13:

⊳ Self-Attention layer

        obtain z_{u_i} = DL(h_i)
14:
                                                                                                        \hat{y}_i(u_i) = \sigma(\Theta^T(\oplus z_{u_i}))
15:
                                                                                                   ▶ Row concat and Predict
        If L(\hat{y}_i(u_i), y_i(u_i)) \leq \varepsilon, break
16:
```

The end-to-end training in KiNN generates a contextualized representation that any classifier can use. The current version of KiNN employs *feed-forward neural network* (*FFNN*) for classification.

3.2 Algorithm

Depending on the attention layers KiNN has two variants named KiNN 1 and KiNN 2. KiNN 1 has cross-attention for $\hat{u}_i\hat{E}$ while KiNN 2 has self-attention network. Algorithm 1 presents the working process of KiNN.

User posts in the form of paragraphs or sentences represented as u_i are passed to KiNN, which outputs a contextualized representation z_{u_i} . Depending on the downstream task of detecting depression or causes of depression (see the Datasets section), the representation is given to a *feed-forward neural network* (*FFNN*) classifier or the output layer. The process involves transforming u_i into u_i^{pt} , where "pt" stands for phrase tagging. The phrases were identified and tagged using multiple domain-specific knowledge sources, as illustrated in Figure 2. Phrase tagging helps infuse depression knowledge in user posts. Domain-specific LM, MentalBERT [42] embeds the user posts (u_i^{pt}) into a vector (\hat{u}_i^{pt}) to pass it through a self-attention layer. This attention layer focuses on significant parts of the post while considering the infused knowledge via tagged phrases.

In parallel, the emotional aspects along with commonsense are extracted for a user document (u_i) through pre-trained COMET on the ATOMIC knowledge graph. Each user document and corresponding text descriptions of five selected aspects $(E_{IW}, E_{EW}, E_{RW}, E_{EL})$, and E_{RL} returned by COMET are concatenated (u_iE) and embedded using MentalBERT encoding $(\hat{u}_i\hat{E})$. This time, the user documents are not phrase tagged as the COMET is trained using BERT tokenizer, which considers words or sub-words as tokens, not phrases. The domain knowledge is infused with phrases; in mental health, phrases have more significance than words and sub-words. This significance is also visible in our previous studies [49].

The embedded representation of u_i and emotional aspects $\hat{u_i}\hat{E}$ is passed through the self-attention network. The network provides a significance score to each aspect according to their relevance to the user post. Further, the output from self-attention layers is merged and normalized before being transferred to a single attention layer to decide the relevance of domain-specific and emotional (commonsense) aspects. This way KiNN finds if a user document is mental health-related or general emotion (commonsense) related. Depending on the task, hidden representation, z_{u_i} , of user posts generated by KiNN is fed into a classification head to obtain the final classification label or vector, $\hat{y}_i(u_i)$. We can interpret which aspects contributed the most predictive value toward the final classification by examining the attention matrices. Algorithm 1 provides a formal pseudocode for the complete process within KiNN. Notably, line 4 and 6 the associated domain knowledge source, which at present is DepressionFeature, can be changed.

Algorithm 2 discusses the process of mapping the user posts into DepressionFeature Ontology (DFO). Phrases from posts unavailable in DFO are searched in UMLS and the top three concepts are extracted and tagged in the corresponding user post. If UMLS also does not contain any phrase then MentalBERT is employed to look for similar phrases. These will further be looked into UMLS for top concepts and tagged in user posts for processing by Algorithm 1.

Algorithm 2 Lexicon Mapping

```
1: Initialize DepressioFeature ontology (DFO) and UMLS
2: while loop over data sample (u_i) do
                                                                                  \triangleright (u_i) is input user post indexed by k

    b tokenized user document

3:
        Set input u_i
        Tag tokens for Part of Speech to keep NOUNS only
 4:
 5:
        for W, where W \leftarrow 1 \ to \ l do
                                                                               \triangleright l is no. of noun tokens in user post u_i
           if DFO has W then
 6:
 7:
               Tag u_i with the matching phrase
           else if DFO does not contain W then
 8:
               Find Synonyms and Similar tokens having similar contextual information (Employed MentalBert)
9:
               if DFO contains Synonym or Similar token then
10:
                   Tag u_i with the matching phrase
11:
12:
               else if DFO does not have a Synonym and UMLS has W or its Synonym then
13:
                   Tag u_i with the matching phrase
14:
                   Import top three Concept Unique Identifiers (CUIs) and their definition from SNOMED CT
```

4 Experimental Setup

4.1 Datasets

CLEF e-Risk (Type: *Binary*; Context: *Depression*): CLEF e-Risk sourced from r/depression subreddit consists of user posts and comments. We have considered the CLEF e-Risk dataset released in 2021. This version of CLEF is annotated with the Beck Depression Inventory (BDI), a CPG used by MHPs for detecting depression [55, 56]. However, there is no existing lexicon for BDI, so we leverage an existing PHQ lexicon created by Yazdavar as the source of clinical groundedness [31]. The dataset comprises at most 2000 Reddit posts per user for 828 users, out of which 79 users have self-reported clinical depression and 749 are control users. The control set consists of random Redditors interested in discussing depression.

PRIMATE (**Type:** *Multi-label*; **Context:** *Depression*): The PRIMATE dataset was constructed to train conversational agents to judge which PHQ-9 questions are answerable from the user's online post. The dataset comprises 2,003 posts sourced from Reddit's r/depression_help subreddit. Each post has been labeled with nine binary annotations that correspond to whether the post addresses one of the nine PHQ-9 items. The dataset has been established as a gold standard for assessing depression severity, demonstrating an inter-annotator agreement (using Fleiss Kappa [57]) of 0.85 among six MHPs affiliated with the National Institute of Mental Health and Neurosciences (NIMHANS) in Bangalore, India. The anonymized dataset is made publicly available by [58].

CAMS (Type: Multi-Class; Context: Mixed Depression and Suicide): The CAMS dataset comprises 5051 instances and is designed to identify the primary causes of MH problems by categorizing social media posts into six causal classes [59]. These classes, which are based on the underlying reasons for mental illness and derived from relevant literature, include 'No reason' (C0), 'Bias or abuse' (C1), 'Jobs and careers' (C2), 'Medication' (C3), 'Relationship' (C4), and 'Alienation' (C5). The dataset is presented in a ¡text, cause, inference; format, where 'text' is the user post with 'cause' referring to the labeled reason behind the mental disorder mentioned in the post, and 'inference' indicates the key phrases (or relevant concepts) in the post that expert annotators considered when assigning labels. The annotations were reviewed by a clinical psychologist and a rehabilitation counselor and validated using Fleiss' Kappa interobserver agreement, achieving a substantial agreement of 0.61.

4.2 Implementation and Training Details:

While training KiNN on *CLEF e-Risk Dataset* user post length is taken as 2000 i.e. a maximum of 2000 tokens (grams) are embedded for a user. Other parameters are two classes (Depressed vs. Non-Depressed), sixteen training and validation batches, fifteen training epochs, and a learning rate of one e-03. For *PRIMATE dataset* the parameters were 9 classes, 150 user post length, 16 training and validation batches, 25 training epochs, and 1e-03 learning rate. Further, to train KiNN on *CAMS Dataset* of 6 classes (C0-C5), maximum text length is taken as 50 with a training and validation batch size of 128. The training epochs were 25 and the learning rate is placed at 1e-03. The training is completed on an Intel Xeon server (8GB) which took about 8 hours.

4.3 Evaluation Metrics:

Four evaluation metrics have been chosen to assess the effectiveness of KiNN across the three datasets. Quantitative evaluation is done using standard performance indicators such as Precision (P), Recall (R), F1, and Matthew Correlation Coefficient (MCC) scores. The creators of the original dataset employed these metrics to guarantee a fair benchmark and enable insightful comparisons. The micro average for P and R is found to be the same in most of the cases. So we have calculated the macro average for P, R, and F1 which provides equal weight to each class in multiple classes.

5 Results & Discussion

Results from KiNN on multiple depression-related datasets are compared with popular generic and domain-specific transformers. This section presents qualitative results along with a user-level explanation generation process and results from the proposed model and one best baseline model.

5.1 Quantitative Results

KiNN *on CLEF e-Risk Dataset:* Table 2 reveals consistent improvement from **KiNN** over baseline methods in the CLEF e-Risk dataset. In terms of precision, recall, and F1 score, KiNN 2 outperforms all other models, followed by KiNN 1. However, the MCC scores reveal a slightly different pattern. KiNN 1 achieves the highest MCC score, indicating better overall classification performance in terms of true and false positives and negatives. In comparison with the best generic transformer model BERT, KiNN 2 showed especially remarkable benefit, particularly in MCC (19%) compared to precision (30.6%), recall (41.6%), and F1 (34.4%) scores. However, while comparing KiNN 2 with the second-best domain-specific transformer model, MentalBERT, significant enhancements were observed across various standard performance metrics. Specifically, in the case of MCC (25.2%) compared precision (4.5%), recall (20.5%), and F1(13%) scores. Similar results were obtained in the case of Naive Bayes (NB) and textCNN models. MCC is high for both the **KiNN** variants which signifies better overall performance of the classifier.

KiNN *on PRIMATE:* For precision, recall, and F1 score, KiNN 2 again emerges as the top-performing model, with KiNN 1 following closely. Interestingly, in terms of MCC, KiNN 1 achieves the highest score, indicating better overall classification performance. From generic transformer models, BERT gives higher scores than other models. KiNN 2 has provided an approximate gain of 10%, 14%, 10%, and 1% over BERT in terms of precision, recall, F1, and MCC scores respectively. From domain-specific models MentalBERT second-best model after KiNN.

KiNN *on CAMS*: provided "inference," reported by annotators as an explanation behind their labeling. Such ground-truth explanations enable a more thorough assessment of the fidelity of the KiNN's attention allocation. Table 2 shows the performance gains of KiNN over CAMS baselines.

KiNN: Though MentalBERT performed best in terms of F1 and MCC scores while BioBERT is best in precision and recall, KiNN variants demonstrated competitive performance across precision, recall, F1 score, and MCC. However, KiNN 1 achieves the higher precision and F1 score on this dataset, while KiNN 2 achieves the higher MCC.

5.2 User-level Explanations

For producing user-level explanations of depression classification results GPT 3.5 is employed. The "text-DaVinci-003" model is prompted in Python through the LangChain library [60]. The prompt is provided with user posts and significant words/phrases from MentalBERT or concepts from KiNN. Depression-related concepts are assigned higher attention scores by KiNN as compared to the MentalBert. GPT considers these concepts or words during explanation generation (see italicized words in Table 3). However, phrases from the MentalBERT model did not produce a comparable pattern in GPT 3.5's behavior. The reason could be observed from Figure 1 which displays highlighted text from MentalBERT and the proposed KiNN model. However, most of the time MentalBERT highlights all text with equal likelihood, and GPT 3.5 focuses on certain terms during the explanation-generating process.

The method establishes a connection between attention words and concepts in DepressionFeature Ontology by computing their cosine similarity. The visualizer then extracts top-ranked concept(s) with a similarity greater than 0.80, considering them potential concepts for generating explanations using GPT 3.5. GPT 3.5 utilized the mapped concepts, as illustrated in Figure 1, to generate an explanation, as shown in Table 3.

Models	CLEF e-Risk (in %)				
Widaels	P	R	F1	MCC	
NB	58.5	53.0	55.0	20.3	
textCNN	57.3	51.2	54.2	22.6	
BERT	36.9	34.4	35.6	23.6	
RoBERTa	31.4	25.1	27.8	17.3	
MentalBERT	63.5	<u>56.5</u>	<u>58.0</u>	18.8	
BioBERT	63.0	55.5	57.0	17.4	
KiNN 1†	64.0	69.0	66.0	32.7	
KiNN 2†	67.5	76.0	70.0	42.6	
		PRIMATE (in %)			
BERT	48.7	42.6	45.4	13.8	
RoBERTa	44.2	41.5	42.8	$\overline{11.4}$	
MentalBERT	54.0	35.0	37.0	12.4	
BioBERT †	44.0	36.0	38.0	6.5	
KiNN 1†	53.0	52.0	52.0	19.8	
KiNN 2†	58.0	56.0	56.0	14.7	
		CAMS (in %)			
MentalBERT	47.0	43.0	43.0	33.8	
BioBERT	49.0	36.0	34.0	29.2	
KiNN 1†	44.0	45	44.0	32.9	
KiNN 2†	41.0	41.0	41.0	29.1	

Table 2: Classification results on CLEF e-Risk, PRIMATE, and CAMS. The best performance is indicated in bold, while the second best is underlined. \dagger (p < 0.05) indicates statistically significant results when comparing the best to the second-best metric.

6 Conclusion

The BlackBox nature of AI models limits their practical implementation in the healthcare domain specifically in mental health assessment due to its subjective behavior. Banking upon a system that does not provide reasoning for its decisions could be too costly in the mental health domain costing someone's life. Further explanations produced by ad-hoc models are rarely understandable to end users. Present models miss domain knowledge and other aspects of a person's personality like commonsense or emotional. To handle the challenges, we presented **KiNN**, a deep knowledge-infused learning network that considers domain knowledge (e.g., DepressionFeature, UMLS) and multiple emotional aspects of users' text (COMET) infused through at multiple levels through attention networks to provide user-level explainability in critical domains like MH.

The performance of our inherently explainable model is comparable to the latest Language Model Models (LLMs) trained on both generic and domain-specific datasets, including BERT, BioBERT, and MentalBERT. The model's performance on all three datasets can be found in Table 2 and Table 3. Moreover, the LLMs are trained on very large datasets and costly resources along with large time and space requirements compared to our model.

The inclusion of clinical domain knowledge, such as Clinical Practice Guidelines (CPG), is crucial for the identification of various symptoms and experiences within unstructured text. LM algorithms can enhance their learning capabilities by incorporating a wider range of mental health (MH) targets. This expansion enables them to provide improved support for diagnosis, treatment, and intervention strategies.

The KiNN model demonstrated superior performance in identifying relevant concepts in unstructured text compared to the baseline models. The infusion of knowledge in natural language processing for mental health involves the incorporation of logical reasoning, such as clinical practice guidelines and emotional aspects. This enables the AI model to engage in continuous learning, ensuring its robustness and resistance to errors. In addition, the establishment of a shared semantic understanding between human experts and language models facilitates collaborative decision-making.

Input Text

So, I decided to go to psychiatrist, but now I feel like I'm not that bad? How bad is really bad? Maybe I just act dramatic I cry a lot without any specific reason, I want to kill myself quite often, but... I dont know if I really need therapy? Maybe its just who I am, just a lazy whiner? And when I will come to therapist I dont know what to say, what if that day I will feel fine and they will not believe me? I dont look like I struggle all the time, people around me don't even know that I have problems, I've always been a comedian in a company. I'm just a little bit afraid. I dont know if i really need therapy, and if I don't, so what do I need to feel normal?

MentalBERT on Words/Phrases

The text describes the narrator's inner struggle with determining if they truly need therapy or if they are just being dramatic. It also highlights their fear and uncertainty about seeking help and their desire to feel "normal". These concepts are connected to their decision to see a psychiatrist and their concerns about being perceived a certain way by others.

KiNN on Concepts

The text discusses the narrator's decision to see a psychiatrist and their doubts about whether they truly need therapy. The concepts identified relate to the narrator's mental state, including feelings of depression (crying and suicidal thoughts), self-doubt (worrying about being seen as a "lazy whiner"), and the desire to feel "normal." These concepts reflect a potential struggle with mental health issues and the need for professional help.

Table 3: Illustrative instances showcasing the quality of explanations from KiNN compared to ClinicalT5, employing GPT3.5. The explanations generated by KiNN exhibit an application-relevant explainability as they accurately capture the context within the user's expression by using concepts in DepressionFeature Ontology. Conversely, other GPT3.5(MentalBERT) explanations appear to deviate from the intended context.

Acknowledgment

This work is supported by the University Grants Commission (UGC) of India. Any opinions, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the UGC.

References

- [1] Anastasia Kotelnikova, Danil Paschenko, Klavdiya Bochenina, and Evgeny Kotelnikov. Lexicon-based methods vs. bert for text sentiment analysis. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 71–83. Springer, 2021.
- [2] Elena Razova, Sergey Vychegzhanin, and Evgeny Kotelnikov. Does bert look at sentiment lexicon? In *International Conference on Analysis of Images, Social Networks and Texts*, pages 55–67. Springer, 2021.
- [3] Shreya Ghosh and Tarique Anwar. Depression intensity estimation via social media: a deep learning approach. *IEEE Transactions on Computational Social Systems*, 8(6):1465–1474, 2021.
- [4] Ronghao Pan, JAG Díaz, and Rafael Valencia-García. Umuteam at erisk@ clef 2023 shared task: transformer models for early detection of pathological gambling, depression, and eating disorder. *Working Notes of CLEF*, pages 18–21, 2023.
- [5] Nut Limsopatham and Nigel Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (volume 1: long papers), pages 1014–1023, 2016.
- [6] Reem Nadeem. 60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care pewresearch.org. Link, 2023. [Accessed 13-08-2023].
- [7] Ben Cushing. AI-Driven Clinical Care Guidelines Can Lead to Better Patient Outcomes healthtechmagazine.net. Link, 2023. [Accessed 13-08-2023].

- [8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [10] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.
- [11] Ayah Zirikly and Mark Dredze. Explaining models of mental health via clinically grounded auxiliary tasks. *CLPsych* 2022, page 30, 2022.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* preprint arXiv:1810.04805, 2018.
- [13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [14] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. arXiv:2004.05150, 2020.
- [15] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. *arXiv preprint arXiv*:2201.10295, 2022.
- [16] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525, 2019.
- [17] Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. Early risk detection of pathological gambling, self-harm and depression using bert. *arXiv preprint arXiv:2106.16175*, 2021.
- [18] Anca Dinu and Andreea-Codrina Moldovan. Automatic detection and classification of mental illnesses from general social media texts. In *Proc. of the Int. Conf. on Recent Advances in Natural lang. Processing (RANLP 2021)*, pages 358–366, 2021.
- [19] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. arXiv preprint arXiv:2007.02847, 2020.
- [20] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*, 2021.
- [21] Annika Marie Schoene, Alexander Turner, Geeth Ranmal De Mel, and Nina Dethlefs. Hierarchical multiscale recurrent neural networks for detecting suicide notes. *IEEE Trans. on Aff. Comp.*, 2021.
- [22] Ning Wang, Fan Luo, Yuvraj Shivtare, Varsha D Badal, KP Subbalakshmi, Rajarathnam Chandramouli, and Ellen Lee. Learning models for suicide prediction from social media posts. *arXiv preprint arXiv:2105.03315*, 2021.
- [23] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.
- [24] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ..., 1999.
- [25] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.
- [26] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh Int. AAAI Conf. on weblogs and social media*, 2013.
- [27] Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C Park. Feature attention network: Interpretable depression detection from social media. In *Proc. of the 32nd Pacific Asia Conf. on Language, Info. and Computation*, 2018.
- [28] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF* (*Working Notes*), 2018.
- [29] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans. on Know. and Data Eng.*, 32(3):588–601, 2018.

- [30] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in twitter. In *Eighth Int. AAAI Conf. on weblogs and social media*, 2014.
- [31] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198, 2017.
- [32] Antoine Briand, Hayda Almeida, and Marie-Jean Meurs. Analysis of social media posts for early detection of mental health conditions. In *Canadian Conf. on Artificial Intelligence*, pages 133–143. Springer, 2018.
- [33] Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, page 119538, 2023.
- [34] Julia Ive, George Gkotsis, Rina Dutta, Robert Stewart, and Sumithra Velupillai. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proc. of the Fifth Workshop on CLPsych.: From Keyboard to Clinic*, pages 69–77, 2018.
- [35] Akkapon Wongkoblap, Miguel A Vadillo, Vasa Curcin, et al. Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study. *JMIR Mental Health*, 8(8):e19824, 2021.
- [36] Rohan Kshirsagar, Robert Morris, and Sam Bowman. Detecting and explaining crisis. *arXiv preprint* arXiv:1705.09585, 2017.
- [37] Ivan Sekulić and Michael Strube. Adapting deep learning methods for mental health prediction on social media. *arXiv preprint arXiv:2003.07634*, 2020.
- [38] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 25(1):281–304, 2022.
- [39] Tapotosh Ghosh, Md Hasan Al Banna, Md Jaber Al Nahian, Mohammed Nasir Uddin, M Shamim Kaiser, and Mufti Mahmud. An attention-based hybrid architecture with explainability for depressive social media text detection in bangla. *Expert Systems with Applications*, 213:119007, 2023.
- [40] Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. τ -ss3: a text classifier with dynamic n-grams for early risk detection over text streams. *Pattern Recognition Letters*, 138:130–137, 2020.
- [41] Lei Tong, Zhihua Liu, Zheheng Jiang, Feixiang Zhou, Long Chen, Jialin Lyu, Xiangrong Zhang, Qianni Zhang, Abdul Sadka, Yinhai Wang, et al. Cost-sensitive boosting pruning trees for depression detection on twitter. *IEEE Transactions on Affective Computing*, 2022.
- [42] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv* preprint arXiv:2110.15621, 2021.
- [43] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. Ucphrase: Unsupervised context-aware quality phrase tagging. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 478–486, 2021.
- [44] Sumit Dalal, Sarika Jain, and Mayank Dave. Featureonto: A schema on textual features for social data analysis. *OpenReview.Net*, 2022.
- [45] Sarika Jain, Sumit Dalal, and Mayank Dave. An ontology for social media data analysis. In *Semantic Intelligence: Select Proceedings of ISIC 2022*, pages 77–87. Springer, 2023.
- [46] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert. Zenodo, 2020.
- [47] Mayank Kulkarni, Debanjan Mahata, Ravneet Arora, and Rajarshi Bhowmik. Learning rich representation of keyphrases from text. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 891– 906, 2022.
- [48] Soheil Danesh, Tamara Sumner, and James H Martin. Sgrank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction. In *Proceedings of the fourth joint conference on lexical and computational semantics*, pages 117–126, 2015.
- [49] Sumit Dalal, Deepa Tilwani, Manas Gaur, Sarika Jain, Valerie Shalin, and Amit Seth. A cross attention approach to diagnostic explainability using clinical practice guidelines for depression. arXiv preprint arXiv:2311.13852, 2023.
- [50] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

- [51] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317, 2019.
- [52] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035, 2019.
- [53] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*, 2020.
- [54] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv* preprint arXiv:2106.01071, 2021.
- [55] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer, 2016.
- [56] Aaron T Beck, Robert A Steer, Gregory K Brown, et al. Beck depression inventory. Harcourt Brace Jovanovich New York:, 1987.
- [57] Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.
- [58] Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. *arXiv preprint arXiv:2205.13884*, 2022.
- [59] Muskan Garg, Chandni Saxena, Veena Krishnan, Ruchi Joshi, Sriparna Saha, Vijay Mago, and Bonnie J Dorr. Cams: an annotated corpus for causal analysis of mental health issues in social media posts. *arXiv preprint arXiv*:2207.04674, 2022.
- [60] Harrison Chase. Langchain python.langchain.com. https://python.langchain.com, 2023. [Accessed 16-08-2023].