Explainable Zero-Shot Modelling of Clinical Depression Symptoms from Text

Nawshad Farruque*†, Randy Goebel*†, Osmar R. Zaïane *†, and Sudhakar Sivapalan‡
* *Alberta Machine Intelligence Institute (AMII), Edmonton, AB, Canada*
† *Department of Computing Science, Faculty of Science, University of Alberta, Edmonton, AB, Canada*
‡ *Department of Psychiatry, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada*
Email: {nawshad,rgoebel,zaiane,sivapala}@ualberta.ca

***Abstract*—We focus on exploring various approaches of *Zero-Shot Learning* (ZSL) and their explainability for a challenging yet important supervised learning task, notorious for training data scarcity, i.e. *Depression Symptoms Detection* (DSD) from text. We start with a comprehensive synthesis of different components of our ZSL modelling and analysis of our ground truth samples and Depression symptom clues curation process with the help of a practicing Clinician. We next analyze the accuracy of various state-of-the-art ZSL models and their potential enhancements for our task. Further, we sketch a framework for the use of ZSL for hierarchical text-based explanation mechanism, which we call, *Syntax Tree-Guided Semantic Explanation* (STEP). Finally, we summarize experiments from which we conclude that we can use ZSL models and achieve reasonable accuracy and explainability, measured by a proposed Explainability Index (EI). This work is, to our knowledge, the first work to exhaustively explore the efficacy of ZSL models for DSD task, both in terms of accuracy and explainability.**

1. Introduction

Earlier studies have shown that, young people who are suffering from Depression, often show help-seeking behavior by speaking out through social media posts. There have been ample research done to date, e.g. [1], [2], that successfully lays the foundation of analyzing posts to extract signs of Depression with reasonable accuracy. However, to detect and confirm signs of Depression in the most clinically accurate way, a Clinician needs to uncover clinical symptoms exhibited by an user on a day-to-day basis.

Since collecting large volumes of expert human (e.g. Clinicians) annotated data is a daunting task, we exploit powerful language models available these days to formulate a Zero-Shot Learning (ZSL) [3] approach for detecting clinical clues of depressive symptoms from Tweets. Moreover, to further understand the efficacy of these models we not only report the accuracy of these models but we also build a framework for explainability analysis [4] of these models. In summary our contributions are as follows:

1) We use state-of-the-art language models, their learned representations and a few subject matter techniques to augment those representations to build our ZSL framework.
2) Since a ZSL task requires minimal clues that can help it to label an "unseen" sample, we carefully curate the clues of depressive symptoms with the help of a practicing Clinician, the nine symptoms of Major Depressive Disorder in the Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition (DSM-5) [5], and validated Depression rating scales.
3) We propose a text explainability algorithm called *Syntax Tree-Guided Semantic Explanation* (STEP), that encourages multiple short and hierarchical phrases inside a Tweet to explain its label.
4) In companion with the previous point, we propose an *Explainability Index* (EI) that is used to grade the explainability mechanism.

2. ZSL Model Preliminaries

ZSL models predict possible membership of a sample to an unseen label, i.e the label it did not see in the training time. For example, in our problem settings, Given, a Tweet, T, it has a label, L_i where, $L_i \in \{L_1, ...L_m\}$ if it has a strong *membership-score* with any of its descriptors, l_j where, $l_j \in L_i$ and $L_i = \{l_1, ...l_n\}$. Here the descriptor l_j is a representation of the label L_i. For example, consider that one of our Depression symptoms L_i is "Low Mood" and the descriptors representing L_i is a set, $l = \{Despondency, Gloom, Despair\}$. If T has a strong membership-score with any members of l, we can say T has the label L_i = "Low Mood.".

We use mainly two broad families of ZSL models in this paper, such as, embedding (both sentence and word) models and Natural Language Inference (NLI) pre-trained models. For embedding models, we represent T and each of the l_j's using various classic and state-of-the-art word and sentence embedding models and measure their membership-scores based on how close they are in the vector space or cosine distance. For NLI models, we extract the probability type of entailment-scores which shows the membership-score for a T with respect to each of the l_j's, see Figure 1.

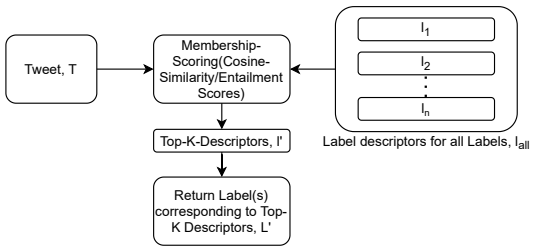

Figure 1. An overview of ZSL Framework

2.1. Ground Truth Dataset (GTD) Curation for DSD Task

Our *Ground Truth Dataset* (GTD) for Depression symptoms was collected and rigorously annotated by experts at the University of Montpellier [6] for binary Depressive Post Detection (DPD) task. We use GTD and label its entries for nine clinical Depression symptoms as per the DSM-5. Before annotating the dataset, we first identify confusing Tweets which could be member of several Depression symptom categories. Later, we arrange a one-on-one discussion session with a Clinician, to confirm possible membership of those Tweets for one or more Depression symptoms categories according to several Depression rating scales, such as, PHQ-9, MADRS, BDI and CES-D [7]. Finally, based on the clinical insights combined with rating scale concepts, we then annotate 255 Tweets for Depression symptoms. We annotate the same dataset twice by the same annotator and achieve a 0.85 test-retest [8] reliability coefficient score.

2.2. Label Descriptors (l) Curation for ZSL

First, we curate our label descriptors (l) as follows: we separate the minimal description of the Depression symptoms or **Header** for DSM-5, called **DSM-Header (DH)** and MADRS [9], called **MADRS-Header (MH)** and slightly elaborated description of the symptoms, or **Lead** for MADRS, called **MADRS-Lead (ML)**. We curate a list of elaborated descriptions of Depression symptoms concept with the help of all the rating scales (as listed in the previous section), and discussion with the Clinician, which we call **All**. In addition, we combine only the headers of DSM-5 and MADRS for corresponding Depression symptoms, which we call **MADRS+DSM-Header (MH+DH)**. Finally, we use a hand curated and expert annotated Depression symptoms lexicon, named **SSToT** [10]. It is to be noted that the choice of good headers and leads are based on availability of those in the Depression rating scales or manuals. For example, we choose both header and lead for MADRS but only header for DSM, because MADRS lead provides significantly more useful description for each Depression symptoms compared to its headers but for DSM only headers are sufficient.

TABLE 1. A GLIMPSE OF FEW DEPRESSION SYMPTOMS LABELS (L) AND SOME HEADERS AND LEADS THAT CONSTITUTE OUR l

Sample of Depression Symptoms, L	DSM-Headers (DH), l	MADRS-Headers (MH), l	MADRS-Leads (ML), l
Disturbed sleep	Insomnia, Hypersomnia	Reduced sleep	Reduced duration of sleep, Reduced depth of sleep
Anhedonia	Loss of interest, Loss of pleasure	Inability to feel	Reduced interest in surroundings, Reduced ability to react with adequate emotion

2.3. Representation of Tweets and Label Descriptors for ZSL

Here we separately discuss about various embedding based representation techniques for our Tweet, T and Depression symptoms label descriptor, l_j.

2.3.1. Word-Embedding-Family (WEF). We use several classic word embedding models, including Google News (Google) [1], Twitter Glove (Glove) [2], Twitter Skip-gram Embedding (TE) [11], Depression Specific Embedding (DSE) trained on Depression specific corpora [12], Depression Embedding Augmented Twitter Embedding (ATE) [12], NLI pre-trained Roberta Embedding (Roberta-NLI) [13] and Universal Sentence Encoder Embedding (USE) [14].

2.3.2. Average Word Vector Models (WV-AVG). If we assume a Tweet, T or a label descriptor, l_j (see Section 2) as our sentence, and each sentence, S consists of n words, i.e., $S = \{W_1, ...W_n\}$, *"wv"* is a function that returns the vector representation of a word, then a sentence as an averaged word vector can be expressed as follows:

$$\frac{\sum_{i=0}^{n} wv(W_i)}{n} \tag{1}$$

2.3.3. Word Vector Mapper Models (WV-MAPPER). As originally proposed in [12], we learn a least square projection matrix, M_w, between the word vectors of the common vocabulary V of both source and target embeddings, see Equation 2. This learned matrix is then used to adjust word vectors of source embedding, then later used to build WV-AVG sentence representation.

$$M_w^* = \arg\min ||wv(V_S)^\top M_w - wv(V_T)||^2 \tag{2}$$

2.3.4. Sentence Embedding Family (SEF). We use state-of-the-art Roberta-NLI and USE sentence embeddings which are transformer based models and multi-task pre-trained on NLI and semantic textual similarity tasks (STS) (i.e. Roberta-NLI) and sentiment analysis tasks as well (i.e USE).

2.3.5. Vanilla Sentence Vector Models (SV). Provided a sentence, S, its sentence vector is represented as $sv(S)$.

1. https://code.google.com/archive/p/word2vec/
2. https://nlp.stanford.edu/projects/glove/

2.3.6. Sentence-to-Word Vector Mapper Models (SV-WV-MAPPER). We use the same formulation as stated in Equation 2, however, while learning the projection matrix M_s^*, here we use the sentence vector of a source word and learn its projection to word vector of the target word for the common vocabulary between the source and target embeddings. All the other notations are the same as noted earlier:

$$M_s^* = \arg\min ||sv(V_S)^\top M_s - wv(V_T)||^2 \qquad (3)$$

2.4. Natural-Language-Inference (NLI) Model

We use the Facebook-BART [15] model, which uses BERT and GPT hybrid pre-training on NLI task and performs very well in ZSL settings. It has been found to be a very effective ZSL model, which can pretty accurately predict whether a given label (in our case label descriptor, l_j) entails a particular sample, in our case a Tweet, (T). This mechanism provides a probability score $\in [0, 1]$ of entailment for each label descriptor.

2.5. ZSL Top-k-Label-Membership Formulation

At the heart of our Top-k-Label-Membership formulation is an algorithm that determines the membership of a Tweet, T with all the descriptors, l_{all} for all labels, L. We later sort the descriptors based on their membership-scores with T in descending order (assuming higher score means better membership), and get $l_{all-sorted}$ (see Algorithm 1). Finally, we return the labels, $L' \subset L$ represented by the top-k descriptors, $l' \subset l_{all-sorted}$ as our candidate labels for the Tweet, T (see Algorithm 2).

Algorithm 1: Sorted-Descriptors

Input : T, l_{all}, $mode$
Output : $l_{all-sorted}$

$l_{all-sorted} \leftarrow \emptyset$;
membership-score-dictionary $\leftarrow \emptyset$;
if *mode is "Embeddings"* **then**
  foreach $l \in l_{all}$ **do**
    membership-score-dictionary[l] $\leftarrow$ 1 - cosine-distance(T, l) ;
  end
end
else if *mode is "NLI"* **then**
  foreach $l \in l_{all}$ **do**
    membership-score-dictionary[l] $\leftarrow$ entailment-prob-score(T, l) ;
  end
end
$l_{all-sorted} \leftarrow$ descriptors(sort-desc(membership-score-dictionary)) ;
return $l_{all-sorted}$;

2.5.1. Embedding Family Models. For this family of models, we use cosine similarity or (1 - cosine-distance) to determine the membership of a vector representation of Tweet, T to the same of any of the descriptors in l_{all}. **Centroid:** In this scheme, we represent each label, L_i with the average representation vectors of all of its descriptors, which we call "centroid". For example, in the centroid-based method, T has label L_i if T has a strong membership-score with $centroid(L_i)$, where $L_i = l = \{l_1, l_2, ..., l_n\}$ and l is the set of descriptors. Then we return $L' \subset L$, i.e. the top-k labels, based on the descending order of the cosine similarity with T (as described earlier) as candidate labels for T. **Top-k Centroid:** Similar to centroid membership, instead of considering all the descriptors of L_i, we use the Top-k descriptors based on the cosine similarity. **NLI Family Models:** As mentioned in Section 2.4, NLI models provide probability scores for entailment for a Tweet, T to its descriptor, l_j. We follow a similar procedure as for the embedding family models except we use the entailment probability scores to find the final candidate labels, L' for T.

Algorithm 2: Label-Predictor

Input : L, $l_{all-sorted}$, k
Output : L'

$L' \leftarrow \emptyset$;
$n \leftarrow 0$;
while $n < k$ **do**
  foreach $l' \in l_{all-sorted}$ **do**
    foreach $L_i \in L$ **do**
      if $l' \in L_i$ **then**
        $L' \leftarrow L' \cup L_i$
      end
    end
    $n \leftarrow n + 1$
  end
end
return L' ;

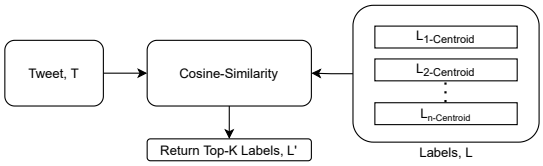

Figure 2. An overview of ZSL-Centroid Methods

3. ZSL Model Explainabilty

Our first algorithm "Syntax Tree-Guided Semantic Explanation (STEP)" respects the syntax tree-based compositionality to explore n-grams inside the Tweet, this compositionality may also contribute to the Tweet having a particular label (see Section 3.1). Our second algorithm, we call n-gram based explanation (ngramex), naïvely divides a Tweet in its constituent n-grams (where "n" is pre-determined) to help explain a Tweet for its label (see Section 3.2). Finally, in Section 4.3 we propose an Explanation Index (EI) function that provides higher scores for multiple minimal

explanations for a Tweet-Label as opposed to single or lengthy explanations.

3.1. Syntax Tree Guided-Semantic Explanation (STEP)

First, we start by approximating semantic understanding of a Tweet, T as a whole (or the label expressed by it), then we gradually explore the nodes of the syntax tree for T in breadth-first manner and find out which n-grams (children of those nodes) also express the same label, until all the nodes have been traversed. Finally, STEP returns the set of n-grams (where "n" is dynamic and $n \in \mathbb{Z}^+$) or "explanations," E, in descending order of membership-score with the top label-descriptor corresponding to the Tweet label, see Algorithm 3. It is to be noted that the label for T returned by $Label(T)$ at line 2 in Algorithm 3 is the label corresponding to top-1 label-descriptor returned by our Algorithm 2.

Algorithm 3: STEP

```
   Input  : T
   Output : E
1  Tree ← Syntax-Tree(T) ;
2  Tweet-Label ← Label(T) ;
3  Explanation-Dictionary ← ∅ ;
4  while not Tree.traversedAllNode() do
       ; // Traversing the Tree in
         Breadth-First order and from
         left-to-right nodes
5      foreach node ∈ Tree do
6          node-Label ← Label(n-gram(node));
7          if node-Label == Tweet-Label then
8              node-Score ← Score(n-gram(node),
                 Tweet-Label) ;
9              Explanation-Dictionary[n-gram(node)] ←
                 node-Score
10         end
11     end
12 end
13 E ←
     explanations(sort-descending(Explanation-Dictionary) ;
14 return E
```

3.2. N-gram based Explanation(ngramex)

In this algorithm, we simply partition T into some pre-defined length of n-grams. Later we identify n-grams which have the same label as T, and return the list of those n-grams according to the descending order membership-score with a label the same way as described in previous section.

4. Experimental Design

We design our experiments to enable analysis with respect to model accuracy and explainability. We report two experiments to confirm the accuracy of our models, such as in (1) **Depression Symptoms Detection from Tweets (DSD)** task, which is our core task and (2) **Depressive Post Detection (DPD)** task, which confirms the predictive capability of our models in general to identify depressive vs. non-depressive Tweets. In terms of explainability, we formulate an explanation index (EI) score and analyze how different models perform in terms of it.

4.1. Train and Test Data-sets

For our DSD task, We have 255 Tweets annotated with expert insight for Depression symptoms except the symptom "Retardation" because our SSToT lexicon does not contain any samples for that category, so for fair comparison we did not consider it. We randomly split our data-set into 80% train-set (204 Tweets) and 20% test-set (51 Tweets). We have three such sets and we report our accuracy scores averaged over those.

For DPD task, we use rigorously human annotated 500 Depressive and non-Depressive Twitter posts. We partition it to 30 stratified train-test splits. Later, we report our accuracy score averaged over those.

4.2. Accuracy Scores

Since our DSD task is a multi-label classification task and our data is rather imbalanced, we use a standard Micro-F1 score to evaluate our ZSL models. For the binary depressive posts detection task, we use a standard F1-score.

4.2.1. Depressive Symptoms Detection (DSD) Task. We perform experiments on all the combinations of our ZSL family models and Depression label descriptors curation strategies earlier described in Sections 2.3 and 2.1 respectively. In addition, we run these experiments for various configurations of top-k $= \{1, 3, 6, 9\}$ label descriptors. However, to analyze and discuss our results, we report the best models under each of the ZSL families. In Table 2, we report these results, where each model is named as: [ZSL-Model-Name(Label-Descriptor-Name)]-[Top-k]. For the baseline, we use BERT ("bert-base-uncased" under "Hugging-Face" transformer library [3]) fine-tuned for our Depression symptoms dataset and few naïve and random classifiers.

4.2.2. Depressive Post Detection (DPD) Task. For this task, we use our membership-scores for various symptoms as the feature representation for the Tweets, then send that to an SVM classifier and compare its performance with LIWC [16] feature representation based SVM classifier along with random uniform, all-majority-class and all-false prediction baselines. We use an SVM classifier because it is found to be best performer, given our small dataset.

4.3. Explanation Index Score (EI-Score)

We propose an Explanation Index (EI) score to evaluate our ZSL Models in terms of their explainability. We report

3. https://huggingface.co/transformers/

EI scores for both STEP and ngramex, and analyze their agreement over different samples to compare and contrast. Let us assume a set of explanations, $E = \{e_1, e_2, ...e_n\}$ for a particular Tweet for its label. Each e_i corresponds to an n-gram explanation of a Tweet for its label. A function "length" returns the number of words in e_i, and the function "rank" returns the rank of a particular e_i in E. Since e_i's are in sorted order under E, the lower the rank the better the explanation. We can express our EI-Score for E as follows,

$$\frac{\sum_{i=0}^{n} EI_i}{n} \tag{4}$$

where,

$$EI_i = LengthScore(e_i) * RankScore(e_i) * Relevance(e_i) \tag{5}$$

$$LengthScore(e_i) = 1 - (length(e_i)/length(Tweet)) \tag{6}$$

$$RankScore(e_i) = 1 - (rank(e_i)/n) \tag{7}$$

$$Relevance = \begin{cases} 1 & \text{if } Label(Tweet) == Label(e_i) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

We can see that EI scores are higher for multiple explanations over a single explanation, and short explanation over lengthy explanations. It is possible that ngramex with a certain "n" can have a better score according to this scoring system, however, ngramex has a high possibility of returning non-salient explanations which are not useful to humans (see Table 5 in Section 5.0.3).

5. Results Discussion

Here we discuss the performance of our models based on two broad categories of performance measures, such as, (1) accuracy and (2) explainability as follows,

5.0.1. Depression Symptoms Detection (DSD) Task Accuracy. We observe that NLI models are the best and Sentence Embedding Famlty (SEF) models are on par indicates, NLI and sentence embedding models with their semantic similarity pre-training, are inherently better in ZSL tasks. All the ZSL models have significantly better accuracy than supervised BERT-fine-tuned model, naïve and random baselines, such as **All-True:** means all the labels are predicted as positive, **All-False:** means all the labels are predicted as negative and **Random-Uniform:** means the labels are predicted either positive or negative based on random discrete-uniform distribution [4]. These experiment results gives us a hope that we can start with ZSL to gather more data before training supervised models.

5.0.2. Depressive Post Detection (DPD) Task Accuracy. We see significant discriminatory capability of ZSL models than the baselines when their prediction scores (i.e., cosine similarity or entailment-probability) for various Depression symptoms are used to represent a Tweet and were fed to SVM for the task. In Table 3, we report the best model's i.e Facebook-BART's Depression prediction capability.

4. https://numpy.org/doc/1.16/reference/generated/numpy.random.randint

TABLE 2. DSD TASK MICRO-F1 SCORES FOR THE BEST MODELS UNDER THE ZSL FAMILIES

ZSL-Family	Model-Name	Micro-F1
WEF	**DSE(MH+DH)-Top-1**	**0.4557**(± 0.0383)
	Glove-ATE(DH)-Top-3	0.3785(± 0.0430)
	ATE(DH)-Centroid-Top-9	0.3589(± 0.0231)
	DSE-Top-k-Centroid(MH+DH)-Top-1	0.3761(± 0.0607)
SEF	**USE(SSToT)-Top-1**	**0.5142**(± 0.0444)
	USE-Mapped(MH+DH)-Top-1	0.4730(± 0.0121)
	USE-Mapped-Centroid(MH+DH)-Top-3	0.3711(± 0.0222)
	USE-Mapped-Top-k-Centroid(MH+DH)-Top-3	0.3711(± 0.0222)
NLI	**Facebook-BART(MH)-Top-1**	***0.5205***(± 0.0196)
Baselines	BERT-Fine-tuned	0.3299(± 0.0246)
	All-True	0.2323 (± 0.0119)
	Random-Uniform	0.2094 (± 0.01323)
	All-False	0.0 (± 0.0)

TABLE 3. F1 SCORES IN DPD TASK

Features	F1-Score
Facebook-BART(MH)-Top-1	0.7830(± 0.0278)
LIWC-Score	0.7404(± 0.0311)
Random-Uniform	0.5102(± 0.0455)
All-Majority-Class	0.6966(± 0.0)
All-False	0.0(± 0.0)

5.0.3. EI-Score. We observe that EI-score wise, sentence embedding based model (USE(SSToT)-Top-1) achieves significantly better score than all the other methods followed by word embedding based, DSE(MH+DH)-Top-1 and NLI based Facebook-BART(MH)-Top-1, See Table 4. Interestingly, the Facebook-BART achieves significantly high accuracy for DSD task, although in-terms of explainability it's worst among the other models, which confirms the inefficacy of entailment scores compared to cosine-similarity to find out salient n-gram explanations. We also observe that ngramex and STEP EI-score usually agrees with each other, although for USE(SSToT)-Top-1, this difference is significant, this could be due to the fact that STEP is capable of extracting explanations which are semantically consistent compared to inconsistent ngrams often extracted by ngramex.

In table 5, we see two examples, where in first example

TABLE 4. EI-SCORES FOR TOP-3 ZSL MODELS REPORTED AT TABLE 2

Models	STEP EI-Score (avg.)	ngramex EI-Score (avg.)
DSE(MH+DH)-Top-1	**0.1605**(± 0.0971)	**0.1769**(± 0.1125)
USE(SSToT)-Top-1	***0.2439***(± 0.0988)	***0.1978***(± 0.1164)
Facebook-BART(MH)-Top-1	0.1261(± 0.1155)	0.1398(± 0.1275)

TABLE 5. TOP 2 EI EXPLANATIONS FOR THE LABEL "FEELING WORTHLESS" FOR TWO TWEET EXAMPLES, WHERE STEP & NGRAMEX DISAGREE FOR TOP EI-SCORING ZSL MODEL: (USE(SSTOT)-TOP-1)

Tweet	Condition	Exps (STEP)	Exps (ngramex)
"No one understands me"	$EI(STEP) > EI(ngramex)$	"No one", "No one understands me"	"No one understands", "one understands me"
"I feel like utter shit"	$EI(STEP) < EI(ngramex)$	"feel like utter shit", "shit"	"I feel like", "feel like utter"

STEP explanations provide high score (0.15) than the same for ngramex (0.1), the reason for EI-Score penalization for ngramex is that, the first explanation is almost the same size as the original Tweet. In the second example, ngramex EI-Score is higher (0.21) than STEP (0.18), here the EI-score penalization for STEP is because of the same reason as first example, however, if we see the semantic quality of the explanations, STEP explanations are better than ngramex.

6. Earlier Work

Most of the earlier work in text based Depression classification can be divided into two broad categories such as, (1) Post level signs of Depression detection [6], [17] and (2) User-level signs of Depression detection [1], [10]. It is to be noted that task (1) is often an important prerequisite of task (2). Even importantly, for clinically meaningful user-level signs of Depression detection, we need to have models that can identify post level signs of clinical Depression symptoms. There have been some efforts put to date for Depression symptoms detection task, such as, [10], [17]. However, most of these works depend on either small and labor intensive gathering of human annotated Tweets or large amount of Tweets for the same through simple rule based distant supervision mechanism which tend to gather noisy Tweets. In this work, we outline a purely ZSL approach to find the semantic similarity relationship between our samples and the label descriptors (which correspond to a certain label). Further Most of the earlier work did not consider the explainability and a need for their explainability evaluation for Depression symptoms task, which is also our primary contribution in this work.

7. Ethical Concerns

Our project is approved by research ethics office for all aspects of data privacy, ethics and confidentiality. To preserve anonymity, all the Tweets sample used in this paper are paraphrased and no user identifier is provided.

8. Acknowledgements

We are grateful to the AMII and CIFAR for their generous support for this project.

9. Conclusion

In this paper we address a challenging task, i.e. Depression Symptoms Detection (DSD) from Text. The main challenge in this task is the scarcity of labelled data. Hence we show that using various learned representation techniques and their enhancements, we can formulate an explainable ZSL approach for this task, which performs better than a fine-tuned BERT-based supervised baseline for the same, provided that our training data is very small. Further we also evaluate the efficacy of our ZSL model explainability with our proposed EI-Score and discuss how different models perform in-terms of providing precise and meaningful explanations.

References

[1] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media." in *ICWSM*, 2013b, p. 2.

[2] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "Clpsych 2015 shared task: Depression and ptsd on twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 31–39.

[3] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks." in *AAAI*, vol. 1, no. 2, 2008, p. 3.

[4] S. Atakishiyev, H. Babiker, N. Farruque, R. Goebel, M. Kima, M. H. Motallebi, J. Rabelo, T. Syed, and O. R. Zaïane, "A multi-component framework for the analysis and design of explainable artificial intelligence," *arXiv preprint arXiv:2005.01908*, 2020.

[5] F. Edition *et al.*, "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, 2013.

[6] M. J. Vioulès, B. Moulahi, J. Azé, and S. Bringay, "Detection of suicide-related posts in twitter data streams," *IBM Journal of Research and Development*, vol. 62, no. 1, pp. 7–1, 2018.

[7] N. C. C. for Mental Health (UK *et al.*, "The classification of depression and depression rating scales/questionnaires," in *Depression in Adults with a Chronic Physical Health Problem: Treatment and Management*. British Psychological Society, 2010.

[8] L. Guttman, "A basis for analyzing test-retest reliability," *Psychometrika*, vol. 10, no. 4, pp. 255–282, 1945.

[9] F. Holländare, G. Andersson, and I. Engström, "A comparison of psychometric properties between internet and paper versions of two depression instruments (bdi-ii and madrs-s) administered to clinic patients," *Journal of medical Internet research*, vol. 12, no. 5, p. e49, 2010.

[10] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," in *ASONAM, 2017*, 2017, pp. 1191–1198.

[11] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle, "Multimedia lab@ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations," in *Proceedings of the workshop on noisy user-generated text*, 2015, pp. 146–153.

[12] N. Farruque, O. Zaiane, and R. Goebel, "Augmenting semantic representation of depressive language: From forums to microblogs," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 359–375.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[16] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[17] D. Mowery, H. Smith, T. Cheney, G. Stoddard, G. Coppersmith, C. Bryan, and M. Conway, "Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study," *Journal of medical Internet research*, vol. 19, no. 2, 2017.