IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Improving Health Mention Classification of Social Media Content using Contrastive Adversarial Training

PERVAIZ IQBAL KHAN[1,2], SHOAIB AHMED SIDDIQUI [1,2], IMRAN RAZZAK [3], ANDREAS DENGEL [1,2], AND SHERAZ AHMED [1]

[1]German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
[2]TU Kaiserslautern, 67663 Kaiserslautern, Germany
[3]UNSW, Sydney, Austrailia

Corresponding author: Pervaiz Iqbal Khan (pervaiz.khan@dfki.de).

ABSTRACT Health mention classification (HMC) involves the classification of an input text as health mention or not. Figurative and non-health mention of disease words makes the classification task challenging. Learning the context of the input text is the key to this problem. The idea is to learn word representation by its surrounding words and utilize emojis in the text to help improve the classification results. In this paper, we improve the word representation of the input text using adversarial training that acts as a regularizer during fine-tuning of the model. We generate adversarial examples by perturbing the word embeddings of the model and then train the model on a pair of clean and adversarial examples. Additionally, we utilize contrastive loss that tries to learn similar representations for the clean example and its perturbed version. We train and evaluate the method on three public datasets. Experiments show that contrastive adversarial training improves the performance significantly in terms of F1-score over the baseline methods of both $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ on all three datasets. Furthermore, we provide a brief analysis of the results by utilizing the power of explainable AI.

INDEX TERMS Health Mention Classification, Contrastive Adversarial Training, Tweet Classification.

I. INTRODUCTION

HEALTH mention classification (HMC) deals with the classification of a given piece of text as health mention or not. This helps in the early detection and tracking of a pandemic which enables health departments and authorities in managing the resources and controlling the situation. The input text is gathered from the social media platforms such as Twitter, Facebook, Reddit, etc. The collection process involves crawling the aforementioned platforms based on keywords containing disease names. The keyword-based data collection does not consider the context of the text and hence contains irrelevant data. For example, a tweet "I made such a great bowl of soup I think I cured my own depression"[1] contains a disease of "depression", but this is used figuratively. Another tweet "Hearing people cough makes me angry. I cannot explain it"[1] contains "cough" in it, but this does not show that a person is having a cough. Non-health and a figurative mention of disease words in these cases pose challenges to the HMC task. So, the question arises of how to address these challenges? One way is to consider surrounding words of the disease words that will give the context of the text. Another way is to leverage the emojis in the text as figurative mentioning text may contain smileys, whereas the actual disease mentioning text may contain emojis of sad faces, etc.

Transformer methods [1] are good at capturing the contextual meanings of the words and have shown success in many natural language processing (NLP) tasks. BERT [2] is a transformer model pre-trained on a large unlabelled text corpus for language understanding, and can be fine-tuned on downstream tasks such as text classification [3]. It considers the words on the left and right sides of a given word while learning a representation for it. In this way, it achieves the contextual representation of a given word. BERT randomly masks 15% of the tokens in the corpus and then tries to predict masked tokens during the training process. RoBERTa [4] is an improvement over the BERT

[1]This tweet is taken from Twitter

using dynamic masking of words instead of static 15% masking of the words. Further, it is trained on 1000% more data than BERT. Existing health mention classification tasks use both non-contextual, and contextual representations for the given text [5]–[9]. However, contextual representations have improved the performance of the classifier over non-contextual representations. Some methods use emojis present in the tweet text for the classification task. [5] extracts the sentiment information from the given tweet and passes it as an additional feature with textual features. [9] converts emojis into text using Python library[2] and then utilizes this emoji text as a part of tweet text.

Adversarial training (AT) [10] works as a regularizer and improves the robustness of the model against adversarial examples. The key idea is to add a gradient-based perturbation to the input examples, and then train the model on both clean and perturbed examples. In contrast to images, this technique is not directly applicable to text data. [11] applies perturbations to word embeddings for the task of text classification. [12] utilizes a contrastive loss for learning features in computer vision (CV). The idea is, that the input image is perturbed by adding some augmentation, and during training contrastive loss pushes both clean and augmented examples together while it pushes other examples away from these examples. Contrastive loss helps the model learn noise-invariant image feature representation. [13] proposes contrastive adversarial for text classification that improves the performance over the baseline methods. In this work, we propose contrastive adversarial training on the task of HMC, additionally using contrastive loss during the fine-tuning of the two transformer models. Specifically, we add perturbation to the embedding matrix of BERT and RoBERTa using Fast Gradient Sign Method (FGSM) [10]. Then we train both the clean and perturbed training examples simultaneously. Our method outperforms both $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ baseline methods on three public datasets. Generally, deep learning models are regarded as black boxes, i.e., it is not clear what information in the input influences the models to make their decisions. European Union adopted new regulations to implement a "right to explanation" which means a user can ask for the explanation of a decision made by the algorithm [14]. Explainable AI focuses on explaining the decisions made by algorithms. In this paper, we leverage explainable AI capabilities to visualize the words that contribute to the model decision. The main contributions of this paper are:

- We propose contrastive adversarial training as a regularizer for HMC and evaluate the performance of the proposed method on three public datasets.
- We show that our method improves HMC performance over the existing methods on three public datasets.
- We provide the analysis of our best-performing model, i.e., RoBERTa decisions by leveraging the power of explainable AI.

[2]https://pypi.org/project/emoji/

The rest of the paper is organized as follows: In section II, we discuss the related work, whereas in section III we present our method for HMC. In section IV, we give experimentation detail. In section V, we present the results and analysis of the experiments. In section VI, we provide the conclusion of the paper.

II. RELATED WORK

In this section, we discuss existing work in the literature related to adversarial training, contrastive learning, and health mention classification of tweets.

A. ADVERSARIAL TRAINING

Adversarial training (AT) has been studied in many supervised classification tasks such as object detection [15]–[17], object segmentation [17], [18] and image classification [10], [19], [20]. AT is the process of training the model to defend against malicious "attacks" and increase network robustness. AT involves the training of the model simultaneously with adversarial and clean examples. These malicious attacks are generated by perturbing the original input examples, so that the model predicts the wrong class label [21], [22] for them. FGSM proposed in [10] is the method for generating adversarial examples for images. [11] extends FGSM to NLP tasks such that it perturbs word embeddings instead of original text inputs and applies the method to both supervised and semi-supervised settings with Virtual Adversarial Training (VAT) [23] for the latter. Recent works propose to add perturbations to the attention mechanism of transformer-based methods [24]–[26]. Compared to single-step FGSM, [21] applies the multi-step approach to generate adversarial examples that proves more effective as compared to single-step FGSM, however it increases the computational cost due to the inner loop that iteratively calculates the perturbations. [27] proposes free adversarial training, where the inner loop calculates the perturbation as well as gradients with respect to the model parameters and updates the model parameters. [26] also uses the free AT algorithm and adds gradient accumulation to achieve a larger effective batch. It also applies perturbations to word embeddings of LSTM and BERT-based models similar to [11]. In our work, we generate adversarial examples using one-step FGSM and perform contrastive learning with clean examples to learn the representations for the input examples.

B. CONTRASTIVE LEARNING

Self-supervised contrastive learning methods, such as MoCo [28], SimCLR [12], and Barlow Twins [29] have narrowed down the performance gap between self-supervised learning and fully-supervised methods on the ImageNet [30] dataset. It has also been applied successfully in the NLP domain. The main idea of contrastive learning is to create positive pairs to train the models. Various methods have been used to create these pairs. [31] uses back-translation to generate another view of the input data. [32] uses the word and span deletion, reordering, and substitution of words, whereas [33]

crops and masks sequences from an auxiliary Transformer to create positive pairs. [34] performs supervised contrastive learning [35] by treating training examples of the same class as positive pairs. To generate positive examples, [36] uses different dropout masks on the same data and treats premises and their corresponding hypotheses as positive pairs and contradictions as hard negatives in the NLI datasets [37], [38]. In our work, we train an original input and its adversarial example in parallel. We further use Barlow Twins [29] as an additional contrastive loss during fine-tuning of models to learn similar representations for the original and its adversarial example.

C. HEALTH MENTION CLASSIFICATION

[7] presents a new method namely Word Embedding Space Partitioning and Distortion (WESPAD) for health mention classification on Twitter data. WESPAD first learns to partition and then distort word representations, which acts as a regularizer and adds generalization capabilities to the model. This method also solves the problem of little training examples for the positive health mentions in the dataset. Although, this method improves the classification accuracy, distorting the original word embedding causes information loss. [6] uses non-contextual word embeddings for tweet health classification. It applies the preprocessing on the given tweet and extracts non-contextual word representations from it, and then passes these representations to Long Short-Term Memory Networks (LSTMs) [39]. LSTMs-based classifier outperforms Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Decision Trees. [8] uses a two-stepped approach for tweet classification. First, it detects whether the disease word is mentioned figuratively or not, and then, it uses this information as a new binary feature combined with other features and applies a convolutional neural network (CNN) for the classification. The usage of this additional feature improves the classification results. This method does not work well on figurative mention tweets, especially the disease word "heart attack", one of the most widely used words in the figurative sense. [5] adds 14k new tweets to the existing health-mention dataset "PHM2017" [7]. It also uses emojis by converting them into string representations using the Python library [3]. As a preprocessing, it normalizes the URL and user mentions in the tweet. This work experiments with both non-contextual representations such as word2vec [40] as well as with contextual representations like ELMO [41] and BERT [2] and incorporates sentiment information using WordNet [42], VAD [43], and ULMFit [44]. It combines the output of the Bi-LSTM [45] with sentiment information to produce a final binary output that represents classification results. Experiments show that combining BERT and VAD outperforms other methods. [9] uses permutation-based word representation method [46] for health mention classification and leverages the emojis as a part of the tweet text by converting them into a text representation. [47] presents a new dataset of Reddit posts called the Reddit health mention dataset (RHMD) and classifies a given post as health mention or not by combining the symptom or disease terms with user behavior. [48] presents a COVID-19 personal health mention (PHM) dataset containing labeled tweets and proposes a dual CNN for the detection of health mention tweets. The dual CNN consists of a primary network called P-Net, and an auxiliary network called A-Net where A-Net helps P-Net to alleviate the class-imbalance issue.

In this paper, we exploit the adversarial training combined with contrastive learning on the task of HMC. For this purpose, we generate adversarial examples using FGSM and employ Barlow Twins [29] as a contrastive loss. We evaluate our method on 3 public datasets.

III. METHOD

In this section, we describe the basics of the transformer-based encoder for text classification. Then we discuss adversarial training and contrastive loss. Finally, we discuss how to combine these ideas to improve the HMC score. Figure 1 shows the overall architecture of the model.

A. TRANSFORMERS BASICS

Let $\{x_i, y_i\}_{i=1,...,N}$ be training examples in the dataset and 'M' be a pre-trained model such as BERT or RoBERTa. Each training example is represented as tokens of sequences, i.e., $x_i = [CLS, t_1, t_2, ..., t_T, SEP]$ as input to M that outputs contextual token representations $[h^L_{CLS}, h^L_1, h^L_2, ..., h^L_T, h^L_{SEP}]$, where 'L' denotes number of layers in 'M'.

To fine-tune pre-trained model 'M', a softmax classifer is added as a final layer that takes the hidden representation h^L_{CLS} of the $[CLS]$ token. A model 'M' is trained by minimizing cross entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} log(p(y_i, c|h^i_{[CLS]})) \quad (1)$$

where 'C' denotes the number of classes in the dataset, and 'N' is the number of training examples in a batch.

B. ADVERSARIAL TRAINING

AT involves perturbing the inputs to the model that cause misclassifications. FGSM is proposed by [10] to generate perturbed examples. The model is trained on both clean and adversarial examples in parallel which improves the model's robustness against adversarial attacks. Let, 'r' be the small perturbation to the input example x_i, and y_i be the ground truth. Then we maximize the loss function:

$$max\mathcal{L}(f_\theta(x_i + r), y_i), s.t. \|r\|_\infty < \epsilon, \text{where } \epsilon > 0 \quad (2)$$

where $\mathcal{L}(f_\theta(x_i + r), y_i)$ is the loss function and f_θ is the neural network parameterized by θ.

[3] https://pypi.org/project/emoji/

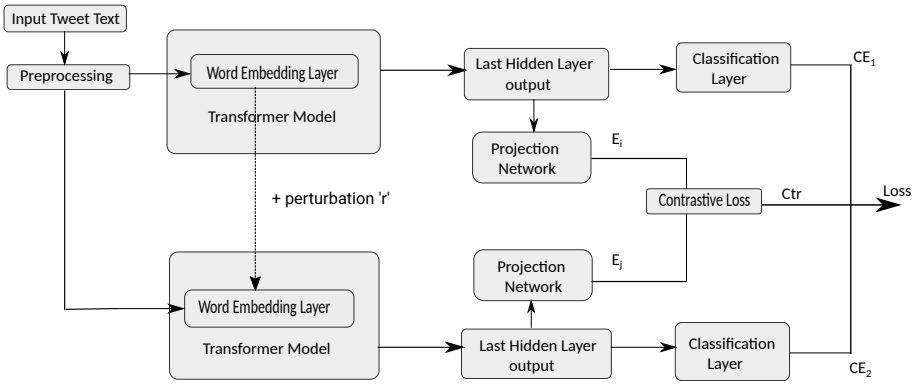


FIGURE 1: Method for contrastive adversarial training. Every input example goes through preprocessing step, then it completes its forward pass through transformer model. We perturb the embedding matrix using Fast Gradient Sign Method (FGSM) to generate adversarial examples. Then we train perturbed example, and it also completes its forward pass. We also utilize contrastive loss represented as 'Ctr'. The final loss is the weighted sum of two cross-entropy losses and a contrastive loss.

To produce the perturbation 'r', Equation (2) can be simplified as follows:

$$r = -\epsilon sign(\nabla_{x_i} \mathcal{L}(f_\theta(x_i), y_i)) \tag{3}$$

To generate adversarial examples, similar to [11] we perturb the embedding matrix $E \in \mathbb{R}^{d_v \times d_h}$ where d_h is hidden unit size and d_v is vocabulary size in the transformer model 'M'. At the end of each forward pass, we calculate the gradient of the loss function given in equation (1), with respect to embedding matrix 'E', instead of input examples as given in equation (3) to calculate the amount of perturbation. We add this perturbation to the embedding matrix and the network goes through another forward pass using the adversarial example. Finally, we calculate another classification loss against the adversarial example.

C. CONTRASTIVE LEARNING

Given a pair of clean and perturbed examples, we want to learn their representation similar to each other while learning different representations for the examples that are not from the same pair. To learn this representation, we leverage contrastive learning as a part of fine-tuning process. To this end, we employ the Barlow Twins loss proposed by Zbontar et al. [29] that is based on the redundancy reduction principle. The equation for the Barlow Twins is given as follows:

$$\mathcal{L}_{ctr} = \sum_{i=1}(1 - M_{ii})^2 + \beta \sum_{i=1} \sum_{j \neq i} M_{ij}^2 \tag{4}$$

where $\mathcal{L}_{ctr}$ is a Barlow Twins, $\sum_{i=1}(1 - M_{ii})^2$, and $\sum_{i=1} \sum_{j \neq i} M_{ij}^2$ represent invariance, and redundancy reduction terms respectively, and β is the trade-off parameter between two terms. M is a square matrix and computes the cross-correlation between clean example embeddings (E^{clean}), and the adversarial example embeddings ($E^{perturbed}$). Values of M vary between -1 (representing a perfect anti-correlation), and +1 (representing a perfect correlation). M_{ij} is computed as follows:

$$M_{ij} = \frac{\sum_{b=1}^{N} E_{b,i}^{clean} E_{b,i}^{perturbed}}{\sqrt{\sum_{b=1}^{N} (E_{b,i}^{clean})^2} \sqrt{\sum_{b=1}^{N} (E_{b,i}^{perturbed})^2}} \tag{5}$$

where i, j, represents the index of the matrix M, and b represents batch samples.

We extract the final hidden states $h_{[CLS]}^{L}$ from both clean and adversarial examples and pass them to three layers multi-layer perceptron (MLP) that projects the hidden units of embeddings from 1024 to 300. Then, we pass these projected units to Barlow Twins loss which aims at learning similar representations for the clean and adversarial examples. Although the original Barlow Twins method projects the image features to higher dimensions, in our experiments, lower-dimensional projection works well. The projection network's first two linear layers consist of input and output dimensions of 1024 and the final layer consists of input dimension of 1024 and output dimension of 300. Every linear is followed by a 1-D batch normalization layer and ReLU as an activation function except the final linear layer.

Similar to [13], we take the weighted average of two classification losses (for clean and its adversarial example) and the contrastive loss (represented by $\mathcal{L}_{ctr}$) as given below:

$$\mathcal{L} = \frac{(1 - \lambda)}{2}(\mathcal{L}_{CE_1} + \mathcal{L}_{CE_2}) + \lambda \mathcal{L}_{ctr} \tag{6}$$

where λ controls the weightage of losses, and $\mathcal{L}$ represents the total loss.

IV. EXPERIMENTS

In this section, first, we discuss the used datasets for training and evaluating our method. Then we give the pre-processing and training details for the method.

A. DATASETS

We use three datasets to train and evaluate contrastive adversarial training. These datasets can be accessed at https:

//github.com/pervaizniazi/HMCDatasets. The detail of each dataset is given as follows:

1) PHM2017

This dataset is an extended version of the PHM2017 dataset provided by [5]. We split the dataset into 65%, 15%, and 20% for the train, validation, and test sets, respectively. This dataset contains data related to 10 diseases, namely, Alzheimer's, cancer, cough, depression, fever, headache, heart attack, migraine, Parkinson's, and stroke. There were 15,742 tweets at the download time, out of which 4,228 tweets were health mentions (HM), whereas 7,322 and 4,192 tweets were non-health mentions (NHM) and figurative mentions (FM), respectively.

2) COVID-19 PHM

This dataset contains tweets related to COVID-19 for HMC task where every tweet example is labeled as one of the four categories., i.e., self-mention, other-mention, awareness, and non-health. There were 9,219 tweet examples available at the time of download. We use the proportion of 8:1:1 for train, validation, and test set split following [48]. Similar to [48] we combine self-mention, other-mention, and non-health categories to tackle the class imbalance issue.

3) RHMD

RHMD dataset contains 10,015 posts from Reddit platform [47]. Every post contains one of the 15 disease or symptom terms such as migraine, asthma, siabetes, PTSD, depression, cough, addiction, Alzheimer, OCD, headache, fever, allergy, cancer, stroke, and heart attack. Every tweet example has a label of one of the four categories, i.e., personal health-mention (PHM), non-personal health mention (NPHM), figurative mention (FM), and hyperbolic mention. The public version of dataset combines figurative and hyperbolic health mention classes.

B. PREPROCESSING

Each tweet goes through the preprocessing pipeline before going through the model. We first convert emojis in the tweet to text using Python library[2]. Then we remove all the user mentions, URLs, hashtags, and special characters. This preprocessing makes the emojis a part of the tweet text.

C. TRAINING DETAILS

We conduct experiments by using $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ as baseline models. Then we apply contrastive adversarial training using these models. For all the experiments, we set a fixed learning rate of $1e^{-5}$ and fine-tune models for 10 epochs. For $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ as baselines, we search over a batch size of $\{16, 32\}$. For contrastive adversarial training, we perform grid search over $\lambda \in \{0.1, 0.2, 0.3\}$, and $\epsilon \in \{0.02, 0.005, 0.001, 0.0001\}$. To compare results with existing methods, we apply 10-fold cross-validation on PHM2017 and RHMD datasets. For 10-fold cross-validation, we choose the best validation hyper-parameters of batch size, λ, and ϵ, and then report average results across 10-folds. We set a maximum sequence length of 64, 68, and 215 for PHM2017, COVID-19 PHM, and RHMD datasets, respectively. For Barlow Twins loss, we choose the default hyperparameters values.

V. RESULTS AND ANALYSIS

We fine-tune two transformer models namely $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ and use these models as the baseline for the task of HMC. For contrastive adversarial training, we use these models with three losses, i.e., two classification losses (for cleaned and adversarial examples) and a contrastive loss, and take the weighted average of these losses.

Table 1 shows the test set results on three datasets for baseline and contrastive adversarial training (denoted as AT + Ctr). On the PHM2017 dataset, BERT + AT + Ctr improves the performance over the baseline by 1.23% and 1.5% in terms of macro F1-score and micro F1-score, respectively. RoBERTa + AT + Ctr improves macro and micro F1-scores of 0.30% and 1.0% respectively, on the PHM2017 dataset. On the RHMD dataset, both macro and micro F1-scores improve by 1.0% and 1.33% respectively over the baseline training method for RoBERTa + AT + Ctr. However, BERT + AT + Ctr degrades the performance over the baseline in terms of both macro and micro F1 scores on the RHMD dataset. On the COVID-19 PHM dataset, BERT + AT + Ctr and RoBERTa + AT + Ctr improve macro F1-score by 0.62% and 4.14% respectively, over their baseline methods. Micro F1-scores improve by 0.5% and 4.5% by BERT + AT + Ctr and RoBERTa + AT + Ctr, respectively over their baseline methods on the COVID-19 PHM dataset.

In Figure 2, we plot the embedding on the validation set of all three datasets for the baseline and contrastive adversarial training of our best performing model, i.e., RoBERTa. We reduce the learned embeddings to lower dimensions using principal component analysis (PCA). The embedding plots show that different embeddings are learned for the baseline and contrastive adversarial training. In Figure 3, we plot the receiver operating characteristic (ROC) curve for the test sets of all three datasets for the baseline and adversarial training. Figure 3a and 3b visualize ROC curves on PHM2017 dataset for BERT and RoBERTa respectively. As shown in Figure 3a, the area of the ROC curve (AUC) for BERT + AT + Ctr is higher than the BERT baseline. However, the AUC for RoBERTa + AT + Ctr is slightly lower than the baseline method as shown in Figure 3b. For the PHM-COVID-19 dataset, the AUC for contrastive adversarial training for both BERT and RoBERTa models is higher than the baseline methods as shown in Figure 3c, and Figure 3d, respectively. As our task on the RHMD dataset is multi-class classification, we plot one vs all ROC curves for it. As shown in Figure 3e, the AUC of the BERT baseline is higher than its contrastive adversarial training. The AUC of RoBERTa + AT + Ctr is higher for FM vs rest as compared to the baseline model as shown in Figure 3f. However, for other classes,

TABLE 1: Results measured on the test set of all the three datasets for BERT and RoBERTa baselines and contrastive adversarial training. RoBERTa with contrastive adversarial training improves F1-score over other experiment settings. * shows statistically significant improvement ($p < 0.05$) of contrastive training over baseline training using McNemar's test.

Model	PHM2017		RHMD		COVID-19 PHM	
	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
BERT	91.81	91.5	81.10	81.33	79.09	79.0
BERT + AT + Ctr	93.04*	93.0*	80.08	80.0	79.71	79.5
RoBERTa	93.45	93.0	81.02	81.0	75.64	75.5
RoBERTa + AT + Ctr	**93.75**	**94.0**	**82.02**	**82.33**	**7**9.78*	**8**0.0*

TABLE 2: Comparison of our method with L. Lu et al. [48] on COVID-19 PHM dataset for binary classification task. Results are micro-averaged Precision, Recall, and F1-score the test set. However, these results are not directly comparable due to data samples mismatch.

Model	Precision	Recall	F1-score
L. Lu et al. [48]	**81.11**	78.29	79.07
BERT + AT + Ctr (Ours)	79.75	**79.67**	79.71
RoBERTa + AT + Ctr (Ours)	80.39	79.23	**79.78**

TABLE 3: Comparison of our method with Naseem et al. [47] on RHMD dataset for 3-class classification setting only. Results are micro-averaged Precision, Recall, and F1-score on the 10-fold validation.

Model	Precision	Recall	F1-score
Naseem et al [47].	81.0	81.0	81.0
BERT + AT + Ctr (Ours)	82.1	81.97	81.97
RoBERTa + AT + Ctr (Ours)	**83.43**	**83.27**	**83.23**

TABLE 4: Comparison of our method with some of the existing methods on an extended version of PHM2017 dataset. Results are micro-averaged Precision, Recall, and F1-score on the 10-fold validation. Our method is directly comparable to only Khan et al. method [9] because the distribution of the dataset does not match with other methods dataset.

Model	Precision	Recall	F1-score
Jiang et al. [6]	72.1	95	81.8
Karisani et al. [7]	75.2	89.6	81.8
Biddle et al. [5]	75.6	92	82.9
Khan et al. [9]	89.1	88.2	88.4
BERT + AT + Ctr (Ours)	93.4	93.8	93.55
RoBERTa + AT + Ctr (Ours)	**94.25**	**94.35**	**94.3**

AUC for baseline methods is higher than the contrastive adversarial training.

A. COMPARISON OF OUR METHOD WITH OTHERS WORK

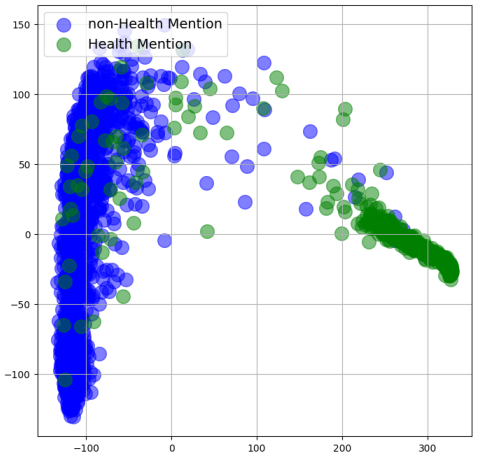
In Table 2, we compare the performance of our method with L. Lu et al. [48] on the COVID-19 PHM dataset for binary classification setting. Our method performs better than L. Lu et al. [48] method in terms of F1-score. However, this is not a fair comparison due to sample mismatch in both experiments. In Table 3, we compare our results with Naseem et al. [47] on the RHMD dataset. Our contrastive adversarial training method for both BERT and RoBERTa improves precision, recall, and F1-score over Naseem et al. [47] method. RoBERTa with contrastive adversarial training improves precision, recall, and F1-scores by 2.43%, 2.27%, and 2.23% respectively over the Naseem et al. [47] method. We present the comparison of our method with some of the work in the literature on the PHM2017 dataset in Table 4. Our method improves precision, recall, and F-score as compared to the work in literature. RoBERTa + AT + Ctr achieves the precision, recall, and F1-score of 94.25%, 94.35%, and 94.3% respectively. In Table 5, we present the results that to see whether the adversarial training or contrastive loss improves the model's performance. Results show that adversarial training improves the F1-score over the baseline method in two of the three datasets. Adding the contrastive training further improves the performance in terms of F1-score in comparison to the adversarial training only on all three datasets.

B. VISUALIZING THE INFLUENTIAL WORDS

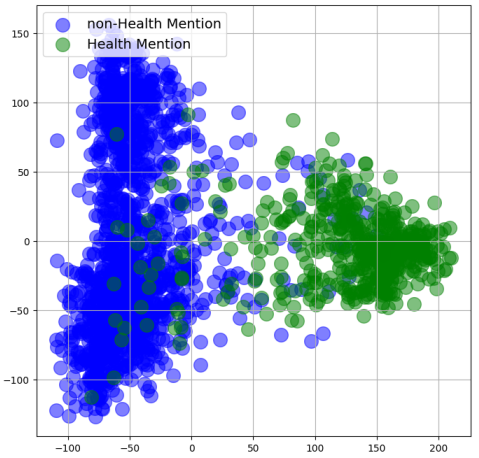
Deep learning models are black boxes in nature, i.e., it is unclear which features of the input influence the deep learning model to reach a decision. Hence, the use of deep learning in critical applications such as healthcare is questionable. European Union announced new regulations to implement a "right to explanation" which means a user can ask for the factors contributing to the decision of the deep learning model. Explainable AI [49] focuses on providing the internals of the model in a human-understandable way to explain the factors influencing the model decision. Especially, various methods explain the model decision by feature, neuron, and layer importance, also known as layer attribution algorithms [50]. In this paper, we visualize the important words that influence the model in reaching the classification decision using transformers-interpret library [51] based on *Integrated Gradients* algorithm [52]. In the Integrated Gradients algorithm, initially, there is no input word to the model. Then, words are gradually added and their impact on the predictions is observed. In this way, the influence of words from the input

FIGURE 2: Embeddings of baseline and contrastive adversarial training methods for the validation sets of the three datasets.

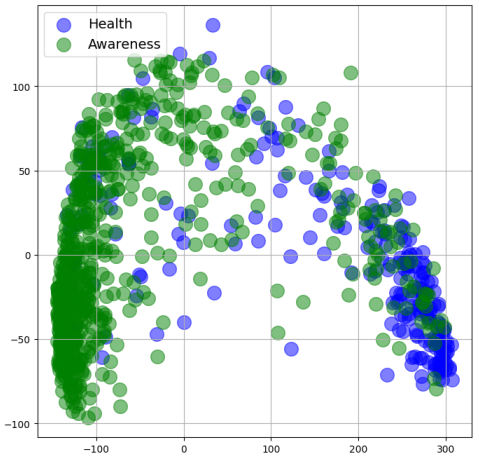
(a) Embeddings of RoBERTa baseline on PHM2017 dataset.



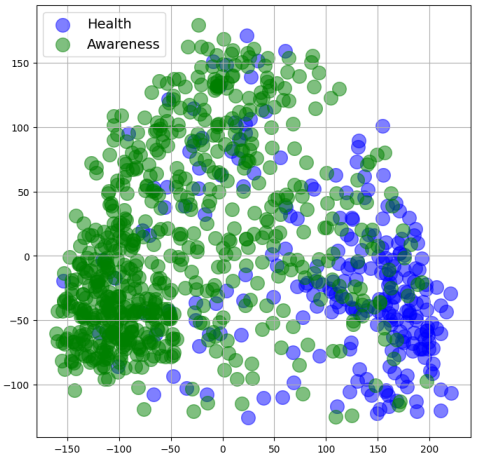
(b) Embeddings of RoBERTa + AT + Ctr on PHM2017 dataset.



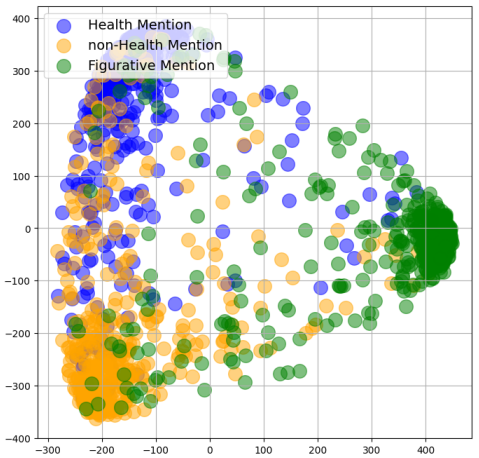
(c) Embeddings of RoBERTa baseline on PHM-COVID-19 dataset.



(d) Embeddings of RoBERTa + AT + Ctr on PHM-COVID-19 dataset.



(e) Embeddings of RoBERTa baseline on RHMD dataset.



(f) Embeddings of RoBERTa + AT + Ctr on RHMD dataset.

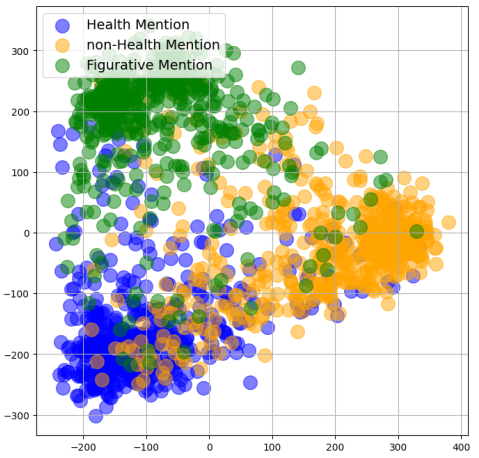


TABLE 5: Results showing the impact of adversarial training (AT) and contrastive adversarial training (AT + Ctr) in terms of macro F1-scores on all the three datasets. Additional use of contrastive further increases the F1-scores on all the datasets.

Model	PHM2017	RHMD	COVID-19 PHM
RoBERTa	93.45	81.02	75.64
RoBERTa + AT	93.49	80.32	79.29
RoBERTa + AT + Ctr	**93.75**	**82.02**	**79.78**

TABLE 6: Visualizations for RoBERTa baseline represented by Baseline and RoBERTa contrastive adversarial represented by Ctr model showing important words that influence by the model for its classification decision. Green highlighted words are those which contributed to the model classification decision. Red highlighted words are those which opposed the model decision. Here, GT stands for ground truth. The prediction column indicates whether the model's prediction is correct or not.

Dataset	GT	Prediction	Model	Word Importance
PHM2017	HM	✗	Baseline	#s just finished rolling my post depression joint so that I can smoke after my therapist session tomorrow #/s
		✓	Ctr	#s just finished rolling my post depression joint so that I can smoke after my therapist session tomorrow #/s
	NHM	✗	Baseline	#s I just straight ened my hair out of depression wow look at me #/s
		✓	Ctr	#s I just straight ened my hair out of depression wow look at me #/s
RHMD	FM	✗	Baseline	#s A British sk yd iver plunged a terrifying 2 000 feet when his parachute malfunction ed yet survived with a stroke of nearly incredible luck . #/s
		✓	Ctr	#s A British sk yd iver plunged a terrifying 2 000 feet when his parachute malfunction ed yet survived with a stroke of nearly incredible luck . #/s
	HM	✗	Baseline	#s Me playing Qu ipl ash with my friend who has cancer What s the difference between me and cancer My friend U hh . . . what Me You won t beat me #/s
		✓	Ctr	#s Me playing Qu ipl ash with my friend who has cancer What s the difference between me and cancer My friend U hh . . . what Me You won t beat me #/s
COVID-19 PHM	NPHM	✗	Baseline	#s so my dad calls to ask me to write my well in case i died from cor ona #/s
		✓	Ctr	#s so my dad calls to ask me to write my well in case i died from cor ona #/s

on prediction is calculated. In Table 6, we plot some randomly selected examples from the test sets of three datasets and analyze the importance of words in the classification decision of the best performing model, i.e., RoBERTa + AT + Ctr. The first tweet example from the PHM2017 dataset, "just finished rolling my post depression joint so that I can smoke after my therapist session tomorrow" is HM and classified by RoBERTa + AT + Ctr as HM. The words like "rolling, post, depression, after, and session" influence the model for classifying this tweet as HM. The words "join and so" contribute towards NHM classification. The model RoBERTa baseline wrongly classifies this tweet as NHM. "just, so, and smoke" are resulting in the model's prediction of NHM, whereas words "finished, depression, join, and therapist" are opposing the model prediction as NHM. The tweet "I just straightened my hair out of depression wow look at me" is classified correctly as NHM by RoBERTa + AT + Ctr. The words "I, just, hair, depression, and wow" influence the model to predict the tweet as NHM, whereas words such as "straightened, of, look, at, me" influence it to predict as HM. On the other hand, RoBERTa baseline wrongly predicts it as HM and the words such as "wow, look, straightened" oppose this decision. Similarly, we plot examples from other datasets as well.

Experimental results show that our method of contrastive adversarial training performs better than the baselines and other methods in the literature. Our method acts as a regularization technique that improves the generalization of the model. However, the amount of perturbation and weightage of the contrastive loss should be chosen carefully as perturbation distorts the embedding matrix, and overuse of perturbation may hurt the performance of the model.

VI. CONCLUSION

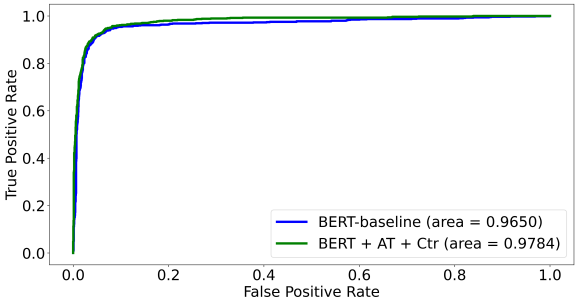
In this paper, we utilized contrastive adversarial training for the health mention classification task as a regularizer. We experimented with two transformers models, i.e., $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ as baselines, and incorporated contrastive adversarial training mechanisms in these models as well. We evaluated the performance of these methods on the three public datasets. Results show that contrastive adversarial training as a regularization technique significantly improves the HMC performance over the baseline methods. We visualized some of the examples from the test set that were correctly classified by the best-performing model of contrastive adversarial training and misclassified by its baseline version to understand the classification decisions made by these models.

REFERENCES

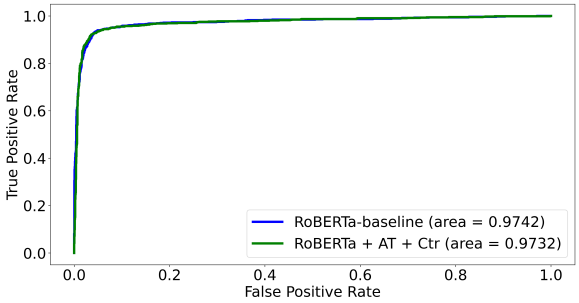
[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

FIGURE 3: ROC curves of baseline and contrastive adversarial training for both $\text{BERT}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Large}}$ models for Health Mention Classification. These embeddings are plotted for the validation set of the extended PHM2017 dataset.

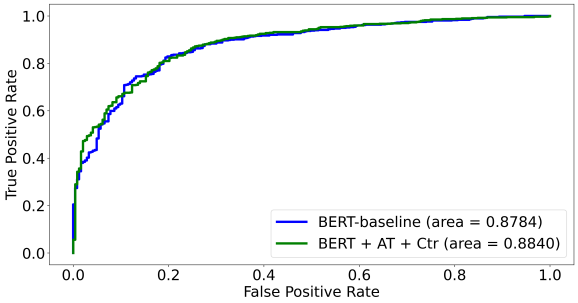
(a) ROC Curve for BERT baseline and BERT + AT + Ctr on PHM2017 dataset.



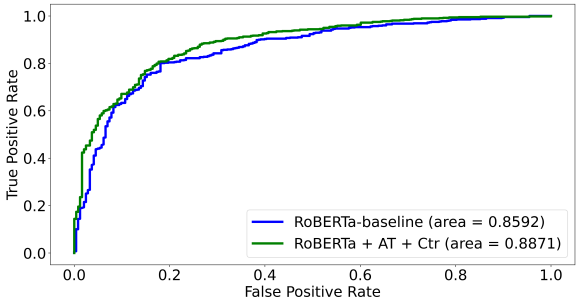
(b) ROC Curve for RoBERTa baseline and RoBERTa + AT + Ctr on PHM2017 dataset.



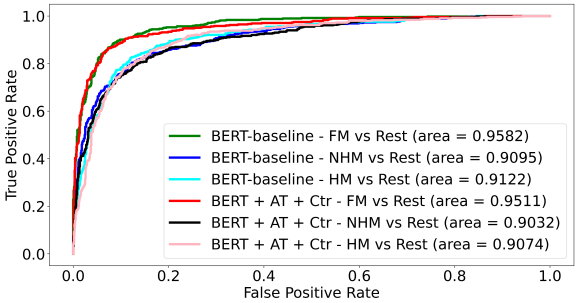
(c) ROC Curve for BERT baseline and BERT + AT + Ctr on PHM-COVID-19 dataset.



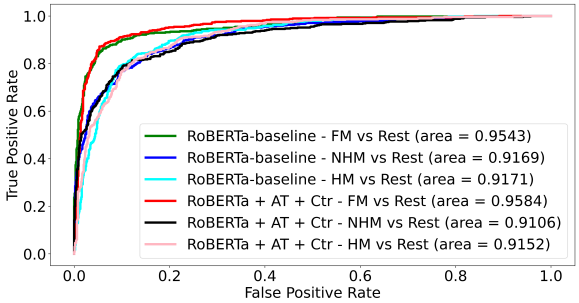
(d) ROC Curve for RoBERTa baseline and RoBERTa + AT + Ctr on PHM-COVID-19 dataset.



(e) ROC Curve for BERT baseline and BERT + AT + Ctr on RHMD dataset.



(f) ROC Curve for RoBERTa baseline and RoBERTa + AT + Ctr on RHMD dataset.



[2] 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.

[3] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In China National Conference on Chinese Computational Linguistics, pages 194–206. Springer, 2019.

[4] 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.

[5] Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. In Proceedings of The Web Conference 2020, pages 1217–1227, 2020.

[6] Keyuan Jiang, Shichao Feng, Qunhao Song, Ricardo A Calix, Matrika Gupta, and Gordon R Bernard. Identifying tweets of personal health experience through word embedding and lstm neural network. BMC bioinformatics, 19(8):67–74, 2018.

[7] Payam Karisani and Eugene Agichtein. Did you really just have a heart attack? towards robust detection of personal health mentions in social media. In Proceedings of the 2018 World Wide Web Conference, pages 137–146, 2018.

[8] Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, and Cecile Paris. Figurative usage detection of symptom words to improve personal health mention detection. arXiv preprint arXiv:1906.05466, 2019.

[9] Pervaiz Iqbal Khan, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. Improving personal health mention detection on twitter using permutation based word representation learning. In International Conference on Neural Information Processing, pages 776–785. Springer, 2020.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[11] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725, 2016.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

[13] Lin Pan, Chung-Wei Hang, Avirup Sil, Saloni Potdar, and Mo Yu. Improved text classification via contrastive adversarial training. arXiv preprint arXiv:2107.10137, 2021.

[14] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". AI magazine, 38(3):50–57, 2017.

[15] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Robust physical adversarial attack on faster r-cnn object detector. corr abs/1804.05810 (2018). arXiv preprint arXiv:1804.05810, 2018.

[16] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In 12th USENIX workshop on offensive technologies (WOOT 18), 2018.

[17] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 1369–1378, 2017.

[18] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 888–897, 2018.

[19] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In 2016 IEEE European symposium on security and privacy (EuroS&P), pages 372–387. IEEE, 2016.

[20] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5):828–841, 2019.

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[22] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069, 2018.

[23] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. arXiv preprint arXiv:1507.00677, 2015.

[24] Shunsuke Kitada and Hitoshi Iyatomi. Attention meets perturbations: Robust and interpretable attention with adversarial training. IEEE Access, 9:92974–92985, 2021.

[25] Shunsuke Kitada and Hitoshi Iyatomi. Making attention mechanisms more robust and interpretable with virtual adversarial training for semi-supervised text classification. arXiv preprint arXiv:2104.08763, 2021.

[26] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. arXiv preprint arXiv:1909.11764, 2019.

[27] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! arXiv preprint arXiv:1904.12843, 2019.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

[29] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning, pages 12310–12320. PMLR, 2021.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[31] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. arXiv preprint arXiv:2005.12766, 2020.

[32] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466, 2020.

[33] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. Coco-lm: Correcting and contrasting text sequences for language model pretraining. arXiv preprint arXiv:2102.08473, 2021.

[34] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403, 2020.

[35] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. arXiv preprint arXiv:2004.11362, 2020.

[36] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821, 2021.

[37] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326, 2015.

[38] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426, 2017.

[39] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[41] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[42] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec, volume 10, pages 2200–2204, 2010.

[43] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 174–184, 2018.

[44] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146, 2018.

[45] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing, pages 6645–6649. IEEE, 2013.

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5754–5764, 2019.

[47] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. Identification of disease or symptom terms in reddit to improve health mention classification. In Proceedings of the ACM Web Conference 2022, pages 2573–2581, 2022.

[48] Linkai Luo, Yue Wang, and Hai Liu. Covid-19 personal health mention detection from tweets using dual convolutional neural network. Expert Systems with Applications, 200:117139, 2022.

[49] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pages 80–89. IEEE, 2018.

[50] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896, 2020.

[51] Charles Pierse. Transformers Interpret, 2 2021.

[52] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR, 2017.

PERVAIZ IQBAL KHAN received his Bachelors degree in Computer Engineering and Masters degree in Computer Science from university of Engineering and Technology, Lahore, Pakistan. Currently, he is pursuing his Ph.D. in Computer Science at German Research Center for Artificial Intelligence (DFKI GmbH) under the supervision of Prof. Dr. Prof. h.c. Andreas Dengel. His research focus lies on improving the healthcare services using Artificial Intelligence.

SHOAIB AHMED SIDDIQUI received the B.S.degree in computer science from the National University of Sciences and Technology (NUST), Pakistan, and the M.S. degree in computer sciencefrom TU Kaiserslautern, Germany. He is a Junior Researcher at the German Research Center for Artificial Intelligence (DFKI GmbH) under the supervision of Prof. Dr. Prof. H. C. Andreas Dengel. His research interests include interpretability and robustness of deep learning models (including robustness against adversarial attacks as well as common image degradations), document understanding, and time series analysis. He is also a Reviewer of ICES Journal of Marine Science, IEEE ACCESS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and AISTATS.

IMRAN RAZZAK (SENIOR MEMBER,IEEE) is a Senior Lecturer of Computer Science in the School of Information Technology at Deakin University since November 2019. Razzak has published more than 120 papers in reputed journals and conferences. He is author of one book and inventor of one patent on Face Recognition. He has attracted research grant of 1.2 million AUD and has successfully delivered several research projects. His area of interest includes machine learning with its application spans a broad range of topics. He has applied machine learning methods with emphasis to natural language processing and image analysis to solve real world problems related to health, finance and social media.

ANDREAS DENGEL is Scientific Director at DFKI GmbH in Kaiserslautern. In 1993, he became Professor in Computer Science at TUK where he holds the chair Knowledge-Based Systems. Since 2009 he is appointed Professor (Kyakuin) in Department of Computer Science and Information Systems at Osaka Prefecture University. He received his Diploma in CS from TUK and his PhD from University of Stuttgart. He also worked at IBM, Siemens, and Xerox Parc. Andreas is member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is coeditor of international computer science journals and has written or edited 12 books. He is author of more than 300 peer-reviewed scientific publications and supervised more than 170 PhD and master theses. Andreas is an IAPR Fellow and received many prominent international awards. His main scientific emphasis is in the areas of Pattern Recognition, Document Understanding, Information Retrieval, Multimedia Mining, Semantic Technologies, and Social Media.

SHERAZ AHMED is Senior Researcher at DFKI GmbH in Kaiserslautern, where he is leading the area of Time Series Analysis and Life Science. He received his MS and PhD degrees in Computer Science from TUK, Germany under the supervision of Prof. Dr. Prof. h.c. Andreas Dengel and Prof. Dr. habil. Marcus Liwicki. His PhD topic is Generic Methods for Information Segmentation in Document Images. Over the last few years, he has primarily worked on development of various systems for information segmentation in document images. His research interests include document understanding, generic segmentation framework for documents, pattern recognition, anomaly detection, Gene analysis, medical image analysis, and natural language processing. He has more than 80 publications on the said and related topics including three journal papers and two book chapters. He is a frequent reviewer of various journals and conferences including Pattern Recognition Letters, Neural Computing and Applications, IJDAR, ICDAR, ICFHR, and DAS.