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

PERVAIZ IQBAL KHAN1,2, SHOAIB AHMED SIDDIQUI1,2, IMRAN RAZZAK3, (Senior Member, IEEE), ANDREAS DENGEL1,2, AND SHERAZ AHMED1

1German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
2Department of Computer Science, TU Kaiserslautern, 67663 Kaiserslautern, Germany
3School of Computer Science and Engineering, UNSW Sydney, Sydney, NSW 2052, Australia
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 BERT Large and RoBERTa 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" contains a disease of "depression", but this is used figuratively. Another tweet "Hearing people cough makes me angry. I cannot explain it" 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. The other 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 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], [6], [7], [8], [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 library² 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 BERT and RoBERTa 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 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.

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], [16], [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]. 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], [25], [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 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., \( x_i = [CLS, t_1, t_2, \ldots, t_T, SEP] \) as input to M that outputs contextual token representations \( [h^\text{CL}_{CLS}, h^1_T, h^2_T, \ldots, h^T_T, h^\text{SEP}_T] \), where ‘L’ denotes number of layers in ‘M’.

To fine-tune pre-trained model ‘M’, a softmax classifier is added as a final layer that takes the hidden representation \( h^\text{CLS} \) of the ‘CLS’ token. A model ‘M’ is trained by minimizing cross entropy loss:

\[
L_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p(y_{i,c} | h^\text{CLS}))
\]

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:

\[
\max L(f_0(x_i + r), y_i), s.t. \|r\|_{\infty} < \epsilon, \text{ where } \epsilon > 0
\]
The equation for the Barlow Twins is given as

$$Z_{bontar} et al.$$ To this end, we employ the Barlow Twins loss proposed by age contrastive learning as a part of fine-tuning process.

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 between two terms. 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 between two terms. 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.

where $L(f_0(x_i + r), y_i)$ is the loss function and $f_0$ is the neural network parameterized by $\theta$.

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

$$r = -\epsilon \times \text{sign}(\nabla_x L(f_0(x_i), y_i))$$

(3)

To generate adversarial examples, similar to [11] we perturb the embedding matrix $E \in \mathbb{R}^{d_e \times d_v}$ 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.

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, diabetes, 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 library2. 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 BERT\textsubscript{Large} and RoBERTa\textsubscript{Large} as baseline models. Then we apply contrastive adversarial training using these models. For all the experiments, we set a fixed learning rate of $1 \times 10^{-3}$ and fine-tune models for 10 epochs. For BERT\textsubscript{L} and RoBERTa\textsubscript{L} 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, $\lambda$, and $\epsilon$, 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 BERT\textsubscript{Large} and RoBERTa\textsubscript{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
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    | 79.78*    | 80.0*     |

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    |

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, 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 existing methods on an extended version of 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 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
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          |          | Base            | #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 |
| NHHM             |          | Base            | #s I just straight ed my hair out of depression wow look at me #/s               |
|                  | ✓        | Ctr             | #s I just straight ed my hair out of depression wow look at me #/s             |
| RIMD             |          | Base            | #s A British sk vid ever 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 vid ever plunged a terrifying 2 000 feet when his parachute malfunction ed yet survived with a stroke of nearly incredible luck #/s |
| COVID-19 PHM     |          | Base            | #s so my dad calls to ask me to write my well in case I died from corona #/s   |
|                  | ✓        | Ctr             | #s so my dad calls to ask me to write my well in case I died from corona #/s   |

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., BERT_{Large} and RoBERTa_{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] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[3] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?” in Proc. China Nat. Conf. Chin. Comput. Linguistics, Cham, Switzerland: Springer, 2019, pp. 194–206.

[4] 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,” 2019, arXiv:1907.11692.

[5] R. Bidlde, A. Joshi, S. Liu, C. Paris, and G. Xu, “Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter,” in Proc. Web Conf., Apr. 2020, pp. 1217–1227.

[6] K. Jiang, S. Feng, Q. Song, R. A. Calix, M. Gupta, and G. R. Bernard, “Identifying tweets of personal health experience through word embedding and LSTM neural network,” BMC Bioinf., vol. 19, no. S8, pp. 67–74, Jun. 2018.

[7] P. Karisani and E. Agichtein, “Did you really just have a heart attack? Towards robust detection of personal health mentions in social media,” in Proc. World Wide Web Conf., 2018, pp. 137–146.

[8] A. Iyer, A. Joshi, S. Karimi, R. Sparks, and C. Paris, “Figurative usage detection of symptom words to improve personal health mention detection,” 2019, arXiv:1906.05466.

[9] P. I. Khan, I. Razzak, A. Dengel, and S. Ahmed, “Improving personal health mention detection on Twitter using permutation based word representation learning,” in Proc. Int. Conf. Neural Inf. Process. Cham, Switzerland: Springer, 2020, pp. 776–785.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, arXiv:1412.6572.

[11] T. Chen, S. Komblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Mach. Learn. Res. (PMLR), 2020, pp. 1597–1607.

[12] L. Pan, C.-W. Hang, A. Sil, and S. Potdar, “Improved text classification and LSTM neural network,” Proc. World Wide Web Conf., 2018, pp. 137–146.

[13] T. Chen, S. Komblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Mach. Learn. Res. (PMLR), 2020, pp. 1597–1607.

[14] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a ‘right to explanation,’” AI Mag., vol. 38, no. 3, pp. 50–57, 2017.

[15] S.-T. Chen, C. Cornelius, J. Martin, and D. Horng Chau, “ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector,” 2018, arXiv:1804.05810.

[16] D. Song, K. Eykholt, I. Evtimov, E. Fernandez, B. Li, A. Rahmati, T. Tramer, A. Prakash, and T. Kohno, “Physical adversarial examples for object detectors,” in Proc. 12th USENIX Workshop Offensive Technol. (WOOT), 2018, pp. 1–10.

[17] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Vuille, “Adversarial examples for semantic segmentation and object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1369–1378.

[18] A. Arnab, O. Miksik, and P. H. S. Torr, “On the robustness of semantic segmentation models to adversarial attacks,” in Proc. IEEE/CFVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 888–897.

[19] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP), Mar. 2016, pp. 372–387.

[20] J. Su, D. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” IEEE Trans. Evol. Comput., vol. 23, no. 5, pp. 828–841, Oct. 2019.

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” 2017, arXiv:1706.06083.
ANDREAS DENGEL received the Diploma degree in CS from TUK and the Ph.D. degree from the University of Stuttgart. He is currently a Scientific Director of DFKI GmbH, Kaiserslautern. In 1993, he became a Professor in computer science at TUK, where he holds the Chair Knowledge-Based Systems. Since 2009, he has been appointed as a Professor (Kyakukin) with the Department of Computer Science and Information Systems, Osaka Prefecture University. He also worked at IBM, Siemens, and Xerox Parc. He is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is the co-editor of international computer science journals and has written or edited 12 books. He is the author of more than 300 peer-reviewed scientific publications and supervised more than 170 master’s and Ph.D. theses. He is a fellow of IAPR 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 received the M.S. and Ph.D. degrees in computer science from TUK, Germany, under the supervision of Prof. H. C. Andreas Dengel and Dr. Habil. Marcus Liwicki. His Ph.D. 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. He is currently a Senior Researcher at DFKI GmbH, Kaiserslautern, where he is leading the area of time series analysis and life science. 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 articles 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.