IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 July 2022, accepted 25 July 2022, date of publication 29 July 2022, date of current version 10 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195030

RESEARCH ARTICLE

Adversarial Training for Fake News Classification

ABDULLAH TARIQ[1], ABID MEHMOOD[2], (Member, IEEE), MOURAD ELHADEF[2], (Member, IEEE), AND MUHAMMAD USMAN GHANI KHAN[1]

[1]National Center for Artificial Intelligence, University of Engineering and Technology Lahore, Lahore 54890, Pakistan
[2]Computer Science and information Technology Department, College of Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates

Corresponding author: Abdullah Tariq (ab.sheikh909@gmail.com)

This work was supported by the Office of Research and Sponsored Program, Abu Dhabi University, United Arab Emirates.

ABSTRACT News is a source of information to know about progress in the various areas of life all across the globe. However, the volume of this information is high, and getting benefits from the available information becomes difficult. Moreover, the frequency of fake news is increasing significantly and used to fulfill a particular agenda. This led to research on the classification of news to prevent the spread of disinformation. In this work, we use Adversarial Training as a means of regularization for fake news classification. We train two transformed-based encoder models using adversarial examples that help the model learn noise invariant representations. We generate these examples by perturbing the model's word embedding matrix, and then we fine-tune the model on clean and adversarial examples simultaneously. We train and evaluate the models on the Buzzfeed Political News and Random Political News datasets. Results show consistent improvements over the baseline models when we train models using adversarial examples. Experiments show that Adversarial Training improves the performance by 1.25% over the BERT baseline, 2.05% over the Longformer baseline for the Random Political News dataset, 1.25% over the BERT baseline and 0.9% over the Longformer baseline for Buzzfeed Political News dataset in terms of F1-score.

INDEX TERMS Fake news classification, political news mining, adversarial training, transformers, longformer, BERT.

I. INTRODUCTION

Nowadays, the internet has become the most common medium for seeking information. The blowout of false news and misleading information is triggering severe problems in the world, partially because most of us only focus on headlines of the news rather than carefully paying heed to its detail. Viewers are misled by intentionally distributing false information because it may contain fabricated or fake information [1]. Increasing internet penetration has made digital media networking a hub for distributing misleading propaganda, inaccurate facts, fraudulent evaluations, rumors, and parodies [2]. Furthermore, massively deceptive communication chains have increasingly harmful implications in various industries, including the stock market. In 2013, for example, the stock market lost 130 billion dollars when false claims on Facebook spread that the US president had been injured in an event [3]. False reporting is accused of having contributed significantly to increasing political polarization during the recent election for president seat in the United States.

The imitations shaped by news headlines to readers are insistent and meaningfully contribute to becoming a news story viral on the social media platform. Therefore, detecting in-congruent news is vital to fight social media misinformation. Researchers have currently exploited different methods for detecting fake news, ranging from simple n-gram features based methods [4], hierarchical encoding based models [5], summarization based models [6] to artificially intelligent systems [7]–[9]. Normally, a system based on artificial intelligence encounters a bottleneck when optimization and tuning of different parameters [10] are essential.

In 2017, Vaswani *et al.* [11] introduced a new neural network architecture known as the transformer. The author stated that due to the fundamental design property of Recurrent Neural Networks (RNNs), it does not perform parallel

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

computing of the learning process. Transformers can tackle this limitation as they consist of multiple sequential attention layers, catch long-range correlations in a sentence, and be computationally efficient [11]. BERT takes the encoder part of the original transformer model and learns the representations for the given input text by masking 15% of the input text randomly. Besides predicting the masked words, BERT also uses the next sentence prediction (NSP) objective function for learning word representation. For predicting the following sentence, BERT jointly takes two sentences, X and Y, as input and learns to classify whether Y genuinely follows X or is only a random text to learn the relationship between sentences.

There are several Transformer models, namely BERT [12], RoBERTa, ALBERT, XLNet, DistilBERT, and Reformer. Other models such as RoBERTa and DeBERTa are built on top of the BERT model. Nevertheless, the major drawback of these models is: that they cannot perform well in longer sentences, i.e., 512 tokens are addressed by BERT at a time. To tackle the difficulties mentioned above, many approaches thrived. An example of such a model is the Longformer [13], a pre-trained model based on the transformer that simplifies the self-attention computation and, thus, reduces model complexity. Longformer can operate on long documents, thus eliminating the shortcomings of transformer models. Longformer can be used to perform several tasks other than language modeling. It can be done as Longformer comprises of the following attention patterns:

- Dilated Sliding Window: The first layers of an attention block of attention layers are regular sliding window layers, and the following layers are dilated sliding window layers. Consider it this way: shallow layers better capture local attention information, whereas top layers aim to move from local to global representations faster, minimizing the overall number of layers required.
- Global Attention: On a few pre-selected input points, we apply ''global attention.'' We also make this attention operation symmetric: a token with global attention pays attention to all tokens in the sequence, and all tokens in the series pay attention to it.
- Sliding Window: The Longformer mimics the convolution process in specific ways. The issue that led the memory needs in traditional attention layers to be quadratic is that it permits each query node to pay attention to both its peers in the keys and all of the key nodes, resulting in n attention weight per query node.

Self-attention is used in both a ''local'' and a ''global'' setting with Longformer self-attention. The majority of tokens interact with each other ''locally,'' which means that each token only interacts with half of its **v** preceding tokens and half of its **v** following tokens, where **v** is the window length.

It is worth noting that the query, key, and value matrices for ''locally'' and ''globally'' attending tokens are different. Additionally, each ''locally'' attending token does not just attend to tokens within its window v, as well as to the global attending tokens, ensuring symmetric global attention. Longformer self-attention can reduce query key matmul operation memory and computation from $O(m_l \times m_l)$ to $O(m_l \times v)$, where m_l be the sequence length and ''v'' is the average window size, the memory and time bottleneck associated with this operation. When compared to the number of ''locally'' attending tokens, it is thought that the numbers of ''globally'' attending tokens are negligible.

Adversarial Training (AT) helps in increasing the model's robustness to adversarial attacks by acting as a regularizer. The key concept is to perturb both clean data as well as disturbed data using a gradient-based perturbation procedure. Unlike images, text data is not directly suitable for this Adversarial Training technique. For the goal of text classification, Miyato *et al.* [14] used perturbations on word embeddings. Chen *et al.* [15] explained how contrastive loss can be used to learn features in computer vision. When this model is trained, the input picture is perturbed with the help of augmentation, and contrastive loss aids toward clean and augmented samples being drawn closer together during training while pushing the other examples apart from them. In model learning, contrast loss facilitates the representation of noise invariant visual features. Pan *et al.* [16] offered a contrastive adversarial strategy for text categorization that outperforms existing algorithms. Pan *et al.* [16] showed that contrastive Adversarial Training improves the text classification performance far better than the baseline approaches.

In this paper, we target the task of fake news classification using two transformer models, BERT and Longformer, and compare their performance. The fake news dataset we use for experimentation consists of the title and the body, where the title of the news is much smaller than the body. Hence, it is worth evaluating the performance of models like BERT and Longformer on smaller and more extended versions of the text. Further, we analyze the impact of Adversarial Training(AT) on these two models by training them on both the clean and perturbed data. The main insights of this research are:

- We compare the performance of two transformer models, namely BERT and Longformer, for classifying fake news.
- We analyze the noise as means of regularizer for fake news classification.

The remaining sections of the paper are organized in the following manner: Section II consists of the related work, followed by section III. In this section, we present the paper's methodology. In section IV, we provide some details about the experimental settings. Moreover, in section V, we elaborate the results of the paper. Finally, the paper's conclusion is discussed in section VI.

II. RELATED WORK

In this part, first, we described some work related to Adversarial Training. Then, we go over the work that has already been done in the field of identifying fake news.

A. ADVERSARIAL TRAINING

A number of supervised classification tasks in the Computer Vision (CV), such as object identification [17]–[19], object segmentation [19], [20], and image classification [21]–[23], have studied the impact of Adversarial Training (AT) on the performance of the tasks. AT uses the concept of using adversarial attacks. In these attacks, input (clean) samples are manipulated so that the system has to predict the incorrect class label [24] for them. As Goodfellow *et al.* suggested in [23], the Fast Gradient Sign Method (FGSM) can be used to produce adversarial examples of images. However, due to the nature of the text, FGSM cannot be implemented directly to the input text to generate adversarial samples. Therefore, Miyato *et al.* [14] applied FGSM to NLP tasks by perturbing word embeddings rather than real text input, which is applicable in both supervised and semi-supervised scenarios, as it uses Virtual Adversarial Training (VAT) [25] in the latter. [26]–[28] proposed different works to add the perturbation to the attention mechanism of transformer-based methods instead of the word embeddings. To generate adversarial examples, Madry *et al.* [24] adopted the multi-step approach in contrast to the single-step FGSM. Hiriyannaiah *et al.* [29] used GAN to better categorize fake news from online platforms. Zhu *et al.* [28] also achieved a larger effective batch by adding gradient accumulation in the free AT algorithm. To update model parameters, Shafahi *et al.* [30] proposed the free Adversarial Training in which the inner loop is responsible for the calculation of perturbation concerning model parameters. Chandra *et al.* [31] presented a methodology for the detection of offensive text. They utilize the adversarial training for making their model more robust.

B. FAKE NEWS CLASSIFICATION

Various AI researchers have proposed different methods in order to classify the news as real or fake using deep learning (DL) and machine learning (ML) tools and approaches. In this section, we present some of the conventional or touchstone techniques for classifying fake news. Social media platforms remain a significant source of news creation and propagation. Facebook, Twitter, WhatsApp, and many other platforms have declared that they are developing their algorithms for the classification of fake news in real-time in order to limit the spread of fake news [32].

Bhatt *et al.* [33] presented a state-of-art solution for stance classification of news. The proposed model is a hybrid of the deep recurrent model's neural encoding, the weighted n-gram bag-of-words model's features, and hand-crafted external features that use feature engineering methodologies. The method is evaluated on real-world data of fake news detection challenge FNC-1. It gained a weighted accuracy score of 83.08%, beating the baseline score of 82.05%.

Kaliyar *et al.* [34] utilized different ML classifiers, namely Gradient Boosting, Multinomial Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and Linear SVM for fake news classification problems. The study suggested that Gradient Boosting produced cutting-edge results with an 86% accuracy on the fake news classification dataset named as FNC-Challenge. Similarly, Jain *et al.* [35] proposed a methodology for fake news detection. The methodology consisted of three modules: an aggregator to extract news from websites, a news authenticator that predicts whether the predicted news is real or fake, and a recommender system. The fake news detection module employed a combination of Support Vector Machines, semantic analysis, and Naive Bayes. The proposed system gained an accuracy of 93.6% on the evaluation dataset.

Umer *et al.* [9] presented a hybrid algorithm that merges the properties of CNN and LSTM. Two dimensionality reduction techniques, chi-square and the Principal Component Analysis (PCA) were used to minimize the features of text articles that helped increase the processing speed. It helped in determining whether the features of news stories were consistent with the content article. The proposed algorithm was evaluated on the FNC dataset and attained an accuracy of 97.8% using the PCA technique. Ahmad *et al.* [36] explored different textual properties of fake news corpus to classify fake and real news. In this study, the dataset used was ISOT Fake News Dataset and two publicly available datasets on Kaggle. The main contribution of this paper was the introduction of an ensemble of ML algorithms using Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Multi-layer perceptron. Ensemble techniques used are Random Forest, Decision Trees as bagging classifiers, AdaBoost and XGBoost as boosting classifiers, and voting classifiers using above mentioned algorithms. The highest accuracy on the ISOT dataset was 99%, and on the combined three datasets was 99% using a Random Forest classifier. The main point of this research was to identify the key elements in classifying fake news. Hardalov *et al.* [37] came up with an end-to-end framework named as Mixture-of-Experts with Label Embeddings (MoLE). Based on unsupervised domain adaptation for the pre-trained RoBERTa model and label embeddings, it was able to learn different labels. The proposed model was evaluated on 16 different stance detection datasets with an average F1-score of 65.55%, while on the FNC-1 dataset, it obtained the same metric with a score of 75.82%.

Vaibhav *et al.* [38] came up with a graph neural network in order to classify fake news by looking up at the sentence level relation of the text. Slovikovskaya *et al.* [39] reported refined stance detection results for the FNC-1 dataset. First, the author tested the power of these embeddings using the Facebook InferSent encoder and BERT based features separately on the featMLP classifier. These findings prompted further research into fine-tuning the BERT, XLNet, and RoBERTa transformers on the FNC-1 extended dataset. Experiments revealed that transformer-based models produced better results, with an accuracy of 91.32%, 92.10%, and 93.19% produced by BERT, XLNet, and RoBERTa, respectively. Briskilal *et al.* [40] proposed an ensemble

TABLE 1. Samples from Buzzfeed and political news dataset.

Description	Type	Fake News	Real News
Buzzfeed Political News Dataset	Title	Pope Francis Endorses Bernie Sanders for President!!	Trump Hired Me As An Attorney. Please Don't...
		President Obama Confirms He Will Refuse To Leave...	Trump is a unique threat to American democracy
		Putin Hacked Emails Reveal That Clinton Threatened...	Trump Is Going To Be Elected
		RAGE AGAINST THE MACHINE To Reunite And..	Trump Voters Just Hear Me Out
		Robertson Said He Had Vision of Trump Seated...	Trump's campaign manager says rape would not...
	Body	experience also made us better realize that the true...	Donald Trump give a humble concession speech if he...
		current president claims he is "fully prepared" to ignore...	The real estate tycoon is uniquely unqualified to...
		Hacked Emails Reveal That Clinton Threatened Sander...	The American people voted for him a long time ago...
		Rage Against The Machine announced they would be...	after the election, so I'd like to address the people...
		Christians are beginning to hold presumed Republican...	Kellyanne Conway was discussing the roles of...
Random Political News Dataset	Title	ASSASSINATION OF RUSSIAN AMBASSADOR JUSTIFIED...	picks budget 'hawk' Mick Mulvaney to lead budget...
		HILLARY'S POPULAR VOTE WIN CAME ENTIRELY...	Trump has no idea how to run a superpower, say...
		Breaking Bombshell: NYPD Blows Whistle on New Hillary...	Trump again claims he 'would have won' popular...
		WikiLeaks – Obama Racist Recruiting – No White,.....	Trump presidency: What his new team...
		Pro-Cruz Robocall Warns Nevada Voters of Impending...	Report: White House preparing response to Russian...
	Body	justice was served" following the assassination of Russian..	shortly after Trump tweeted an inflammatory comment.
		Trump's electoral victory, supporters of presidential loser...	China's Global Times newspaper on Monday morning....
		Preexisitng conditions and republican plans to replace...	if that is possible, if the winner was based on popular...
		many naive Americans still gave him the benefit of the...	the incoming president has to fill more than 4,000...
		hundreds of thousands of conservative voters on Tuesday..	report says that details are still being finalized...

TABLE 2. Dataset statistics.

Metadata						
Dataset-ID	Description	Domain	Type	Satire News	Fake News	Real News
1	Buzzfeed News Data	Politics	Title	0	48	53
			Body	0	48	53
2	Random News Data	Politics	Title	75	75	75
			Body	75	75	75

method that combined the weights generated from two transformer models, BERT and RoBERTa. This proposed architecture was fine-tuned with idioms and literal, namely TroFiIn. In addition, the authors presented a new dataset containing 1470 idioms and literal expressions. The overall accuracy of the ensemble model was 90.4%, compared to 85% and 88% for the standalone BERT and RoBERTa models, respectively.

So far, we have seen how transformer-based models can capture semantic and protracted correlations in sentences. However, Kaliyar *et al.* [41] developed fakeBERT, a BERT based DL model for detecting fake news. Input vectors generated by BERT after word embedding were passed into three convolutional layers, each preceded by a pooling layer and stacked in parallel blocks. The model was evaluated using a fake news dataset provided by Kaggle with true and false labels only with an accuracy of 98.90%. Furthermore, this paper provided a detailed overview of BERT embeddings used with various state-of-the-art methods. However, this model only detected true or false labels and can be extended to classify multi-class real-world datasets.

In order to classify long textual news, In this paper, we fine-tuned BERT and Longformer with Adversarial Training. Adversarial Training helps the models classify the fake news with more accuracy.

III. METHODOLOGY

In methodology, first, we deliberate the fundamentals of transformers for text classification, followed by fine-tuning BERT and Longformer on our datasets with/without noise for the classification of fake news. Then, we describe the Adversarial Training. At last, we merge these ideas to improve the overall score.

A. TRANSFORMER ENCODERS

To get the statistical representations of the input text, we utilize two transformer models, BERT and Longformer, as transformer models. Let $x = \{[CLS], x_1, x_2, \ldots x_n, [SEP]\}$ be the input token, where $[CLS]$, and $[SEP]$ are the unique tokens representing the beginning and conclusion of the sequence. $[CLS]$ token contains the hidden representations of the whole input sequence x. Let *encoder* be a transformer encoder representing either BERT or Longformer at a time. Given a sequence x as an input, the encoder producers hidden representation H for each input token:

$$H = encoder(x) \tag{1}$$

where $H \in \mathbb{R}^{n \times d}$ where d represents number of hidden units and n is the maximum sequence length. We fine tune the *encoder* and add a softmax classifier that takes the hidden representation of $[CLS]$ token. The aim of the training function is to reduce the cross entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} log(p(y_i, c|h^i_{[CLS]})) \tag{2}$$

where '$\mathcal{L}$' represents the training loss, 'N' be the number of training samples in a batch, and 'C' represents the number of classes in a dataset.

B. ADVERSARIAL TRAINING

Inputs are being perturbed in Adversarial Training so that model misclassifies the given inputs. Fast Gradient Sign Method (FGSM) is one of the methods proposed in [23] to generate the perturbed examples. These perturbed samples are called adversarial samples. Once adversarial examples are generated, the model is trained on a pair of perturbed and clean examples. Let 'p' represents the perturbation, then the

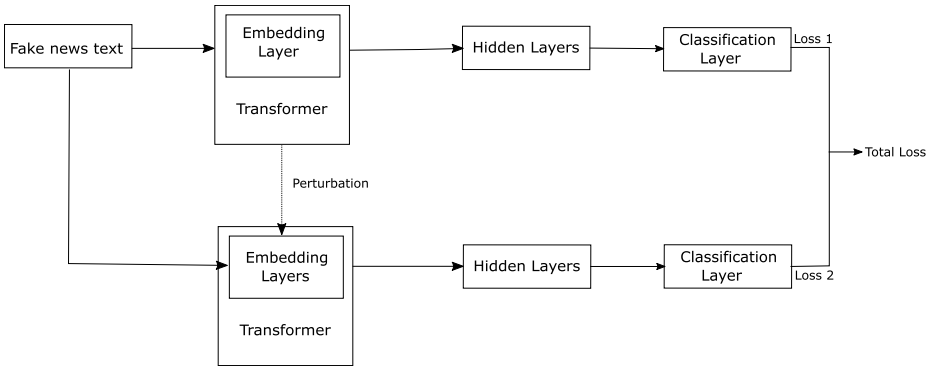


FIGURE 1. Workflow of Adversarial Training for fake news classification. We add perturbation to the embedding matrix to generate adversarial samples and then train the model using clean and adversarial samples simultaneously.

amount of perturbation generated using the FGSM method can be represented as follows:

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

where $\mathcal{L}(f_\theta(x_i + r), y_i)$ represents the cross entropy loss function in our work and f_θ represents neural network parameterized by θ, and ϵ is the hyperparameter to control the amount of perturbation. To generate adversarial examples, instead of adding perturbation to the input itself, we add perturbation 'p' to the embedding matrix for every input text. Then, we train the model jointly on adversarial and clean examples. The total loss is the sum of both losses. i.e., one for the clean example and the other for the perturbed example given as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{clean} + \mathcal{L}_{adv} \tag{4}$$

where $\mathcal{L}_{clean}$ and $\mathcal{L}_{adv}$ represent losses for the clean and adversarial examples, respectively.

C. ADVERSARIAL TRAINING FOR FAKE NEWS CLASSIFICATION

This study employs the transformer models coupled with Adversarial Training for classifying the text related to news and then classifying it as fake or real news. Input news text is passed to the transformer model that comprises an embedding layer and a series of hidden layers. The hidden state of $[CLS]$ token representing the whole input sequence is then passed to the classification layer, and then the loss is computed as given in equation 2. In order to estimate the perturbation, we compute the gradient of the loss in terms of the embedding matrix, as shown by equation 3. The adversarial example is created by adding this perturbation to the transformer model's embedding matrix. Then, this adversarial example also goes through a series of hidden layers, then the classification layer, and finally, the loss is computed. The total loss becomes the sum of losses of clean and adversarial examples, as shown in equation 4. Then the backward step is taken, and the model's parameters are updated. In this way, the fake news classification model is trained. The overflow of the training procedure is shown in Figure 1.

IV. EXPERIMENTAL SETTING

A. DATASET

The detail of dataset [42] we use for the experiments is given in Table 2. It contains news datasets from two independent sources, i.e., "Buzzfeed News Data" and "Random News Data". "Buzzfeed News Data" contains 48 samples for fake news and 53 samples for real news. "Random News Data" contains 75 samples for fake news, real news, and satires. In this work, we only use fake news and real news data. Both datasets contain the title as well as the body of the news. We classify the dataset using title and body separately. Some samples of these datasets are shown in Table 1.

B. BASELINE METHODS

As a baseline method, we use two transformer-based models. The detail of both models is given as follows:

1) BERT-LARGE

We fine-tune the BERT large model for the purpose of fake-news classification that consists of 24 encoder blocks. We consider this model as a baseline model. For adversarial training, we use the same architecture; however, now, instead of training on clean training examples, We use both clean and perturbed samples to train the model. This makes the one forward pass of the model training consist of two examples, i.e., clean and adversarial examples.

2) LONGFORMER-4096

We fine-tune the other model for fake news classification is Longformer-4096, which consists of 4096 hidden units in the final hidden layer. Like BERT, we also use this model for classification based on both the title and the news body and treat it as a baseline for its counter Adversarial Training. We use the same strategy of Adversarial Training for Longformer we used for BERT. However, the value of the noise parameter differs from the BERT.

C. EVALUATION MEASURE

We employed accuracy, precision, recall, and the F1-score as evaluation measures to assess our model's performance. The definitions of accuracy, precision, recall, and F1-score are given in equations 5, 6, 7 and 8 respectively.

$$Accuracy = TP + TN/TP + FP + FN + TN \quad (5)$$

$$Precision = TP/(TP + FP) \quad (6)$$

$$Recall = TP/(TP + FN) \quad (7)$$

$$F1 - score = 2\frac{Precision * Recall}{Precision + Recall} \quad (8)$$

where TP represents the true positive news, the news which was real, and the model also predicts that news as real, TN represents the true negative, and the actual fake news is predicted as fake by the model. FP represents the false positive; the actual fake news predicted real by the model. FN represents the false negative; the actual real news predicted fake by the model.

D. HYPERPARAMETERS

For both BERT and Longformer, we use a fixed learning rate of $1e^{-5}$. For the classification of news titles, we use a maximum of 50 sequences, both on the BERT and the Longformer. For classification based on the body of the news, we use a maximum sequence length of 512 for BERT and 1000 for Longformer. For the value of noise parameter ϵ, we use 0.001 and 0.0001 for BERT and Longformer, respectively. Due to computational constraints, we use a fixed batch size of 1 for both baseline and Adversarial Training on the news body using BERT and RoBERTa. For classification of the title of the news, as the maximum sequence length is small, we use a batch size of 8 for both baseline and Adversarial Training. However, we still use a batch size of 1 for Longformer to make a classification based on the title. The reason for this, the Longformer model does not fit in GPU memory even though the title length is small. We use Adam [43] as an optimization algorithm for both the models and linear weight decay. Models are trained for 10 epochs, and early stopping is used to prevent overfitting.

V. RESULTS AND DISCUSSION

For the classification of news as fake and real, we use two transformer models, BERT and Longformer. We use these two models as baselines. Moreover, we employ noise as means of regularization for fake news classification using Adversarial Training. The results on the body and title of the news for both Buzzfeed and Random Political News datasets are presented in Table 3.

A. DATASET 1: BUZZFEED POLITICAL NEWS DATASET

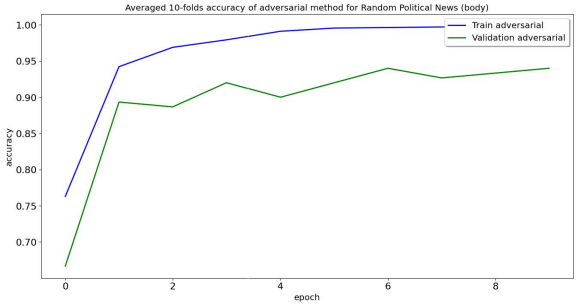
As shown in Table 3, we attained a precision, recall, and F1- score of 86.85%, 85.75%, and 84.95% respectively for ''Buzzfeed Political News'' dataset using BERT without adversarial training for the title of the news. However, when we include adversarial training for the BERT model on the title, precision, recall, and F1-score values, we get 84.85%, 83.85%, and 83.2%, respectively. This means the performance of the model degrades. For Longformer without Adversarial Training on the title of the news, we get a precision, recall, and F1-score of 87.75%, 87%, and 86.15%, respectively. On the other hand, for Longformer with Adversarial Training on the news title, we see the precision, recall, and F1-score values of 87.15%, 86.4%, and 86.0%, respectively. We again notice the performance degradation for the classification of news on the basis of the title when using Adversarial Training. On the other hand, in the case of classifying fake news on the basis of the body of the news, we get a precision, recall, and F1-score of 81.35%, 80%, and 79.0% for the BERT baseline model. However, for the Adversarial Training of the BERT model, based on news body precision, recall, and F1-score values increase by 3.25%, 0.75%, and 1.25%, respectively. Similarly, for the Adversarial Training of Longformer model over the body of the news, we see that the precision, recall, and F1-score increase by 2.15%, 0.65%, and 0.9%, respectively.

B. DATASET 2: RANDOM POLITICAL NEWS DATASET

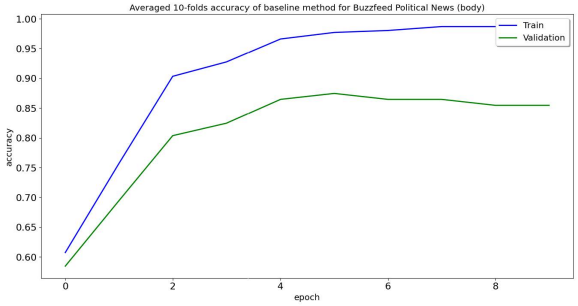
We discuss the performance of the baseline and adversarial training models of the ''Random Political News'' dataset in this subsection. Results show that for the title of the news, the BERT baseline gives a precision, recall, and F1-score of 83.2%, 81.95%, and 81.1%, respectively. On the other hand, BERT adversarial model performs with precision, recall, and F1-score of 81.6%, 80.85%, and 80.55%, respectively. In the case of the Longformer baseline model, we get a precision, recall, and F1-score of 89.05%, 87.55%, and 87.25%, respectively. For the Longformer with Adversarial Training, we had a precision of 89%, a recall of 87.55%, and an F1-score of 87.1%.

The BERT baseline method performs with precision, recall, and F1-score of 89.05%, 88.35%, and 88.05% when classifying fake news based on the body of the news, whereas BERT with Adversarial Training performs with precision, recall, and F1-score of 90.6%, 89.85%, and 89.3%, respectively. Longformer baseline performs with precision, recall, and F1-score of 94.25%, 94.15%, and 93.95%. However, Longformer with Adversarial Training outperforms the baseline with precision, recall, and F1-score of 96.25%, 96.1%, and 96%, respectively. BERT with Adversarial Training for the news body gains a performance improvement of 1.55%, 1.5%, and 1.25% for the precision, recall, and F1-score, respectively, over the BERT baseline method. Similarly, Longformer with Adversarial Training for the news body gains a performance improvement of 2.0%, 1.95%, and 2.05% in terms of precision, recall, and F1-score, respectively.

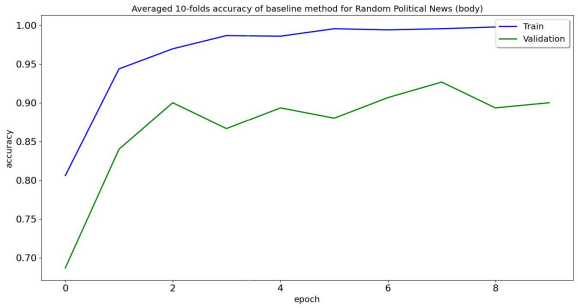
As Longformer performs the best on the body of news for both datasets, we plot accuracy and loss for it in Figure 2 and 3. Although validation loss increases after certain epochs, we save the model with the highest F1-score on the validation fold. Similarly, we plot the confusion matrix for both training methods in Figure 4.



(a) Accuracy plot of Longformer model on Buzzfeed Political News dataset using baseline training method.



(b) Accuracy plot of Longformer model on Buzzfeed Political News dataset using adversarial training method.



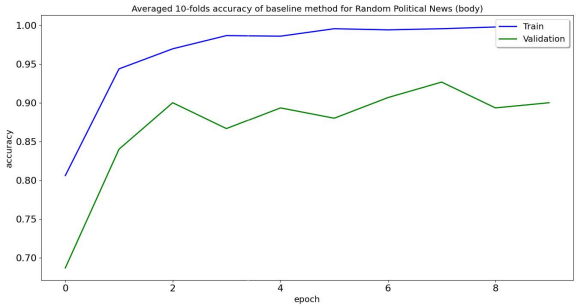
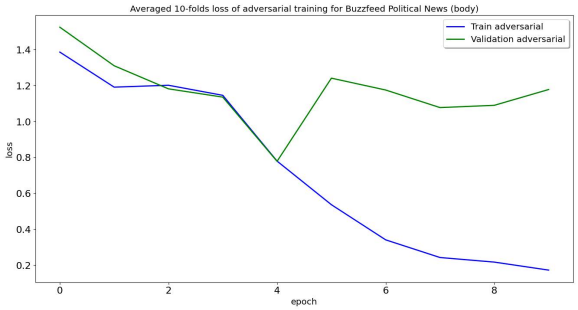
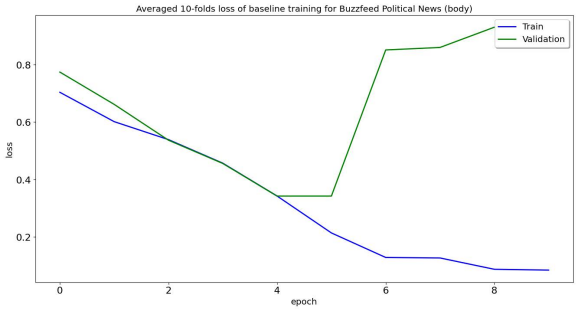


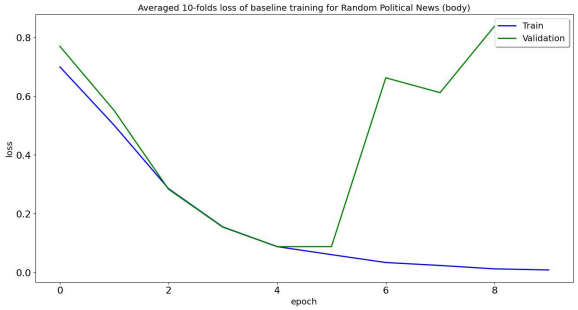
FIGURE 2. **Averaged train and validation accuracy plots of both datasets for the baseline and adversarial methods on 10-folds cross-validation.**



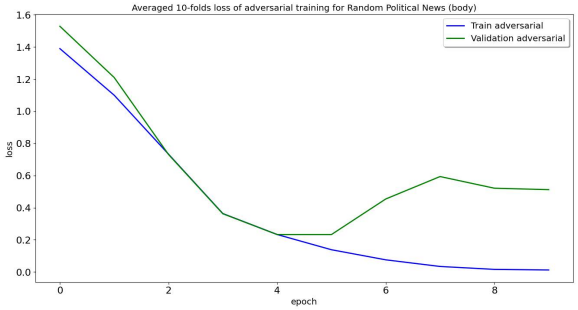
(a) Loss of Longformer on Buzzfeed Political News Dataset using baseline training method.



(b) Loss of Longformer on Buzzfeed Political News Dataset using adversarial training method.



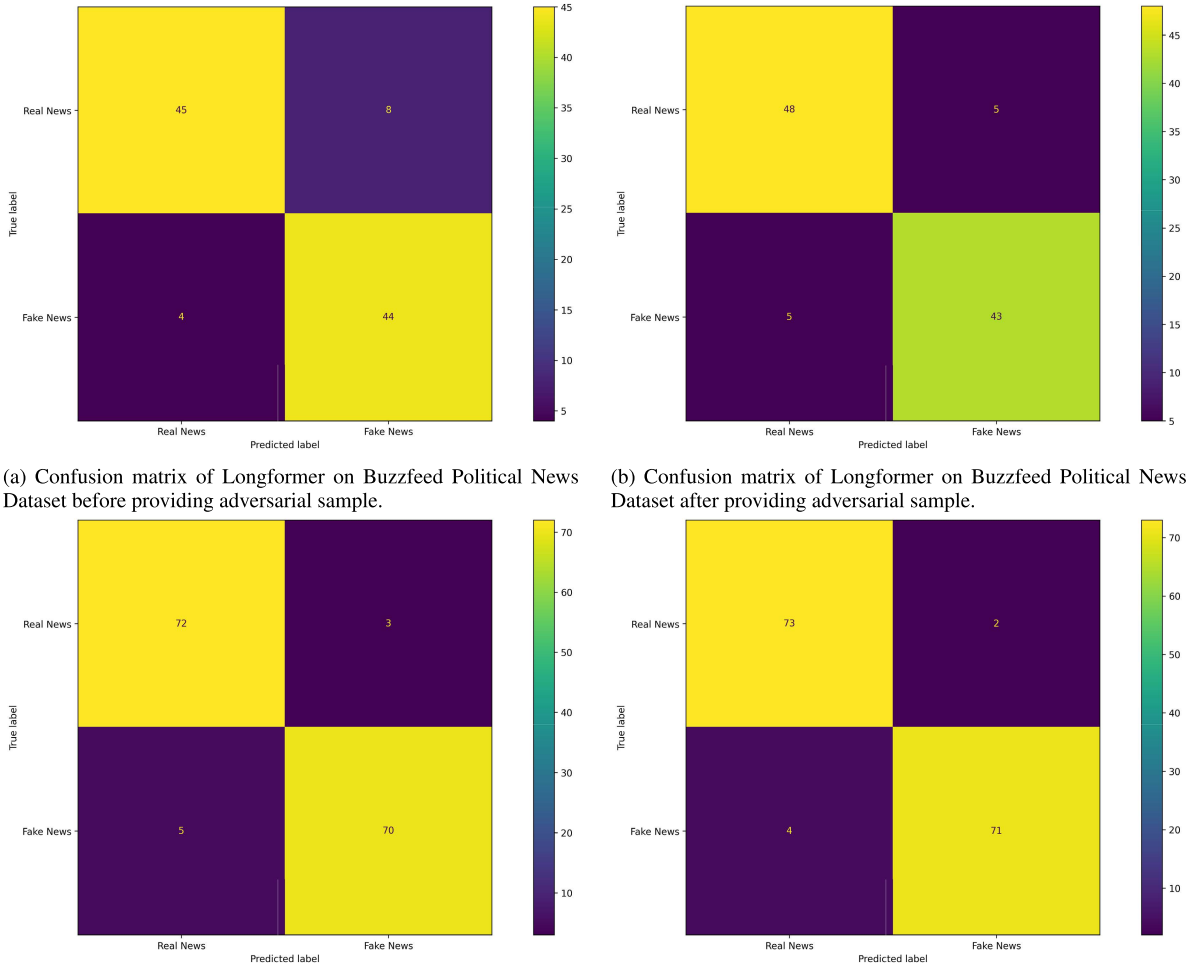
(c) Loss of Longformer on Random Political News Dataset using baseline training method.



(d) Loss of Longformer on Random Political News Dataset using adversarial training method.

FIGURE 3. **Averaged train and validation loss plots of both datasets for the baseline and adversarial methods on 10-folds cross-validation.**

According to the experimental results, the Longformer model outperforms the BERT on the title and the body of the news on both datasets. Moreover, the model's performance degrades when adding noise to the title of the news. This makes sense because the news title already contains short text, and adding noise further reduces the useful information.



(a) Confusion matrix of Longformer on Buzzfeed Political News Dataset before providing adversarial sample.

(b) Confusion matrix of Longformer on Buzzfeed Political News Dataset after providing adversarial sample.

(c) Confusion matrix of Longformer on Random Political News Dataset before providing adversarial sample.

(d) Confusion matrix of Longformer on Random Political News Dataset after providing adversarial sample.

FIGURE 4. Confusion matrix of 10-folds cross-validation for the baseline and adversarial methods.

TABLE 3. Averaged precision, recall and F1-score for the Buzzfeed political news and random political news datasets on 10-fold cross validations. Table shows the results for BERT and Longformer models with and without FGSM noise for both the title and body of the news. Results show that adding FGSM noise for body of the news consistently improves F1-score on both the models.

Dataset	Type	BERT						Longformer					
		Baseline			Adversarial			Baseline			Adversarial		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Buzzfeed Political News Dataset	Title	**86.85**	**85.75**	**84.95**	84.85	83.85	83.2	**87.75**	**87**	**86.15**	87.15	86.4	86.0
	Body	81.35	80	79.3	**84.6**	**80.75**	**80.55**	89.4	88.2	87.75	**91.55**	**88.85**	**88.65**
Random Political News Dataset	Title	**83.2**	**81.95**	**81.1**	81.6	80.85	80.55	**89.05**	87.55	**87.25**	89	87.55	87.1
	Body	89.05	88.35	88.05	**90.6**	**89.85**	**89.3**	94.25	94.15	93.95	**96.25**	**96.1**	**96**

On the other hand, model performance significantly improves if we employ Adversarial Training for the longer text, such as the body of the news.

VI. CONCLUSION

In this paper, we analyzed the impact of Adversarial Training as means of regularizer for the fake news classification task. To this end, we utilized two transformer models, BERT and Longformer. We measured the performance of the models on two publicly available datasets, namely, Random Political News and Buzzfeed Political News. Evaluation results show that Adversarial Training for the classification of news on the basis of long text such as the body of the news increases the models' performance significantly over the baseline in terms of precision, recall, and F1-score on both datasets. However, Adversarial Training for the short text, such as classification using the title of the news, degrades models' performance. Moreover, the Longformer model performs better than the BERT for the fake news classification using both the title and the body of the news. In addition, for future work, we would consider adding more algorithms for adding perturbed samples other than using FGSM and exploring the way to add the required noise value in order not to add much noise that the accuracy would fall rather than increase. Moreover, we would

try to explore how effective our technique for the detection of offensive text and present the outcomes in the graphical form.

REFERENCES

[1] X. Jose, S. D. M. Kumar, and P. Chandran, "Characterization, classification and detection of fake news in online social media networks," in *Proc. IEEE Mysore Sub Sect. Int. Conf. (MysuruCon)*, Oct. 2021, pp. 759–765.
[2] A. E. Fard and T. Verma, "A comprehensive review on countering rumours in the age of online social media platforms," in *Causes Symptoms Socio-Cultural Polarization*. Springer, 2022, pp. 253–284.
[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, May 2018.
[4] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," 2017, *arXiv:1707.03264*.
[5] R. Mishra, P. Yadav, R. Calizzano, and M. Leippold, "MuSeM: Detecting incongruent news headlines using mutual attentive semantic matching," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 709–716.
[6] S. Yoon, K. Park, M. Lee, T. Kim, M. Cha, and K. Jung, "Learning to detect incongruence in news headline and body text via a graph neural network," *IEEE Access*, vol. 9, pp. 36195–36206, 2021.
[7] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.
[8] S. Girgis, E. Amer, and M. Gadallah, "Deep learning algorithms for detecting fake news in online text," in *Proc. 13th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2018, pp. 93–97.
[9] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (CNN-LSTM)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020.
[10] L. Abualigah, A. Diabat, S. Mirjalili, M. A. Elaziz, and A. H. Gandomi, "The arithmetic optimization algorithm," *Comput. Methods Appl. Mech. Eng.*, vol. 376, Apr. 2021, Art. no. 113609.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[13] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.
[14] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," 2016, *arXiv:1605.07725*.
[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2020, pp. 1597–1607.
[16] L. Pan, C.-W. Hang, A. Sil, and S. Potdar, "Improved text classification via contrastive adversarial training," 2021, *arXiv:2107.10137*.
[17] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, 2018, pp. 52–68.
[18] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno, "Physical adversarial examples for object detectors," in *Proc. 12th USENIX Workshop Offensive Technol. (WOOT)*, 2018, pp. 1–10.
[19] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1369–1378.
[20] A. Arnab, O. Miksik, and P. H. S. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 888–897.
[21] 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.
[22] 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.
[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
[25] T. Miyato, S.-I. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," 2015, *arXiv:1507.00677*.
[26] S. Kitada and H. Iyatomi, "Attention meets perturbations: Robust and interpretable attention with adversarial training," *IEEE Access*, vol. 9, pp. 92974–92985, 2021.
[27] S. Kitada and H. Iyatomi, "Making attention mechanisms more robust and interpretable with virtual adversarial training for semi-supervised text classification," 2021, *arXiv:2104.08763*.
[28] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "FreeLB: Enhanced adversarial training for natural language understanding," 2019, *arXiv:1909.11764*.
[29] S. Hiriyannaiah, A. Srinivas, G. K. Shetty, G. Siddesh, and K. Srinivasa, "A computationally intelligent agent for detecting fake news using generative adversarial networks," in *Hybrid Computational Intelligence*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 69–96.
[30] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[31] S. Chandra and B. Das, "An approach framework of transfer learning, adversarial training and hierarchical multi-task learning—A case study of disinformation detection with offensive text," *J. Physics, Conf.*, vol. 2161, no. 1, Jan. 2022, Art. no. 012049.
[32] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 811–824, Nov. 2012.
[33] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "On the benefit of combining neural, statistical and external features for fake news identification," 2017, *arXiv:1712.03935*.
[34] R. K. Kaliyar, A. Goswami, and P. Narang, "Multiclass fake news detection using ensemble machine learning," in *Proc. IEEE 9th Int. Conf. Adv. Comput. (IACC)*, Dec. 2019, pp. 103–107.
[35] A. Jain, A. Shakya, H. Khatter, and A. K. Gupta, "A smart system for fake news detection using machine learning," in *Proc. Int. Conf. Issues Challenges Intell. Comput. Techn. (ICICT)*, vol. 1, Sep. 2019, pp. 1–4.
[36] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, pp. 1–11, Oct. 2020.
[37] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "Cross-domain label-adaptive stance detection," 2021, *arXiv:2104.07467*.
[38] V. Vaibhav, R. M. Annasamy, and E. Hovy, "Do sentence interactions matter? Leveraging sentence level representations for fake news classification," 2019, *arXiv:1910.12203*.
[39] V. Slovikovskaya, "Transfer learning from transformers to fake news challenge stance detection (FNC-1) task," 2019, *arXiv:1910.14353*.
[40] J. Briskilal and C. N. Subalalitha, "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa," *Inf. Process. Manage.*, vol. 59, no. 1, Jan. 2022, Art. no. 102756.
[41] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021.
[42] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, 2017, pp. 759–766.
[43] N. Landro, I. Gallo, and R. La Grassa, "Mixing Adam and SGD: A combined optimization method," 2020, *arXiv:2011.08042*.

ABDULLAH TARIQ is currently pursuing the M.S. degree in computer science with the University of Engineering and Technology Lahore. He is also working as a Research Officer with the Intelligent Criminology Research Laboratory, National Center of Artificial Intelligence. His research interests include computer vision, ML, and DL.

ABID MEHMOOD (Member, IEEE) received the Ph.D. degree in computer science from Deakin University, Australia. He is currently an Assistant Professor with Abu Dhabi University. His research interests include ML, privacy, information security data mining, and cloud computing.

MOURAD ELHADEF (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the Institute Supérieur de Gestion in Tunis, Tunisia, and the Ph.D. degree in computer science from the University of Sherbrooke, Sherbrooke, QC, Canada. He is currently a Computer Science Professor at the College of Engineering, Abu Dhabi University, United Arab Emirates. He has over 50 peer-reviewed articles and conference proceedings to his credit. His current research interests include failure tolerance and fault diagnosis in distributed, wireless, *ad-hoc* networks, cloud computing, artificial intelligence, and security. He is on the Editorial Boards of several major conferences and journals, including IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS and the *Journal of Parallel and Distributed Computing*.

MUHAMMAD USMAN GHANI KHAN is currently the Director of the Intelligent Criminology Laboratory under the Center of Artificial Intelligence. He is also the Director and the Founder of five research laboratories including, the Computer Vision and ML Laboratory, the Bioinformatics Laboratory, the Virtual Reality and Gaming Laboratory, the Data Science Laboratory, and the Software Systems Research Laboratory. He has over 18 years of research experience specifically in the areas of image processing, computer vision, bioinformatics, medical imaging, computational linguistics, and ML. A Well-Groomed Teacher and a Mentor for subjects related to artificial intelligence, ML, and DL. Recorded freely available video lectures on youtube for courses of bioinformatics, image processing, data mining and data science, and computer programming.