All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustness of Fake-news Detectors Under Black-Box Settings

Hassan Ali*, Muhammad Suleman Khan†, Amer AlGhadhban‡, Meshari Alazmi§,
Ahmad Alzamil‡, Khaled Al-utaibi ¶, Junaid Qadir‖

*IHSAN Lab, Information Technology University, Lahore, Pakistan.
†Dept. of Computer Science, Information Technology University (ITU), Lahore, Pakistan.
‡Dept. of Electrical Engineering, College of Engineering, University of Ha'il, Saudi Arabia
§Dept. of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, Saudi Arabia.
¶Dept. of Computer Engineering, College of Computer Science and Engineering, University of Ha'il, Saudi Arabia.
‖Dept. of Electrical Engineering, Information Technology University (ITU), Lahore, Pakistan.
*hassanalihassan3093@gmail.com †{msds19011}@itu.edu.pk ‡{a.alghadhban,aa.alzamil}@uoh.edu.sa
§ms.alazmi@uoh.edu.sa ¶alutaibi@uoh.edu.sa ‖junaid.qadir@itu.edu.pk (Corresponding author)

***Abstract*—With the hyperconnectivity and ubiquity of the Internet, the fake news problem now presents a greater threat than ever before. One promising solution for countering this threat is to leverage deep learning (DL)-based text classification methods for fake-news detection. However, since such methods have been shown to be vulnerable to adversarial attacks, the integrity and security of DL-based fake news classifers is under question. Although many works study text classification under the adversarial threat, to the best of our knowledge, we do not find any work in literature that specifically analyzes the performance of DL-based fake-news detectors under adversarial settings. We bridge this gap by evaluating the performance of fake-news detectors under various configurations under black-box settings. In particular, we investigate the robustness of four different DL architectural choices—multilayer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN) and a recently proposed Hybrid CNN-RNN trained on three different state-of-the-art datasets—under different adversarial attacks (Text Bugger, Text Fooler, PWWS, and Deep Word Bug) implemented using the state-of-the-art NLP attack library, Text-Attack. Additionally, we explore how changing the detector complexity, the input sequence length, and the training loss affect the robustness of the learned model. Our experiments suggest that RNNs are robust as compared to other architectures. Further, we show that increasing the input sequence length generally increases the detector's robustness. Our evaluations provide key insights to robustify fake-news detectors against adversarial attacks.**

***Index Terms*—fake news detection, deep neural networks, adversarial attacks, adversarial robustness.**

I. Introduction

Recent advances in information and communication technology including the rise of social media, artificial intelligence (AI), computational bots, and ubiquitous connectivity has resulted in an information ecosystem that is awash with low-quality, partisan, or even outright fake-news [1]. Advances such as deep learning and generative adversarial networks (GANs) have made it easy for any motivated entity to create fake-news and use it for large-scale opinion manipulation [2]. One such example is the US 2016 presidential elections where fake-news generated for personal gains were believed and shared by 37 million Facebook users [1], [3].

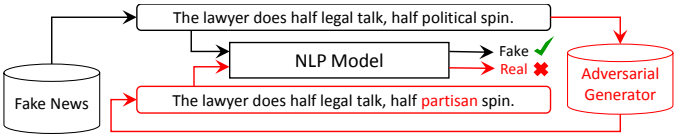


Fig. 1: Illustration of adversarial attack on an NLP model. Illustration of adversarial attack on an NLP model. *By flipping just one word of sample fake news, an adversary can make an ML-based NLP method fail in correctly recognizing the fake-news.*

The future well-being of our society is contingent on combating the fake-news malaise effectively. In recent times, the use of AI and machine learning (ML) have been proposed for developing algorithmic fake-news detectors that can flag false information. In particular, researchers are leveraging deep neural networks (DNNs)-based text-classification methods to meet the fake-news detection challenge [1]–[4]. Although DNNs provide a general solution to many intelligent tasks of diverse nature, e.g., object recognition and localization, scene understanding, paragraph generation, and summarizing (to name a few), the performance of DNNs highly depends on large training datasets [5] which makes them vulnerable to attacks at both the training [6] and inference stages [7]. For example, recent works show that DNNs are easily fooled when a slightly perturbed input, a so-called *adversarial input* is input. Deep learning-based methods in all their popular incarnations including convolutional neural networks (CNNs), recurrent neural networks (RNNs), multilayer perceptrons

(MLP) have been found to be vulnerable to these adversarial attacks [8], [9]. The adversarial ML threat is applicable to various application domains including image recognition, speech recognition, networking devices, and even natural language processing (NLP) models. We illustrate the working of an adversarial attack on an NLP-based ML-model in Fig. 1.

Although we find a range of works studying the robustness[1] of DL-based text-classifiers against the adversarial attacks [10], there is very limited work in the literature that explores the adversarial ML threat for ML-based fake-news detection methodologies. To bridge this gap, we evaluate a recently proposed Hybrid CNN-RNN based fake-news detector [1], generalizable to different datasets under the adversarial setting. For this purpose, we utilize the state-of-the-art library *Text-Attack*[2] [10], which implements 16 different state-of-the-art attack strategies to benchmark the robustness of DNNs on several Natural Language Processing (NLP) tasks. Further, we analyze the adversarial threat surface of different detector architectures for several hyper-parameters under the *black-box threat model*, a threat model in which knowledge of the detector and its parameters are not assumed which makes this model more practical and adaptive [11]. Our motivation for adopting the black-box assumption is also based on the observation that black-box attacks are considered more reliable compared to the white-box attacks when used for benchmarking since black-box attacks more effectively counter the gradient-obfuscation problem exhibited by many defenses and models [12].

Although a recent work [13] evaluates fake-news detector under the adversarial threat, our work differs in a number of ways. Unlike the approach adopted in current models [13], which used a manual method for generating adversarial examples, we automatically generate adversarial examples using four different approaches, i.e. Text-Bugger, Text-Fooler, PWWS and Deep Word Bug, from a state-of-the-art library, Text-attack [10]. The main goal of this study is to answer the following key questions about the robustness of fake-news detectors under several engineering choices.

- Which architecture provides the most robust solution to the fake-news detection problem?
- How does changing the number of learnable parameters of the detector (detector complexity) affect its robustness?
- How do different training-time design choices, i.e. input sequence length, the training loss and the regularization affect the robustness of the final detector?

Our findings highlight key insights for a generic defense against adversarial attacks. Specifically, our experiments suggest that RNNs are relatively robust as compared to the CNN, MLP, and other hybrid architectures. We experiment with different input sequence lengths and discover that large input sequences increase the robustness of the detector by increasing the Attack Success Rate (ASR)[3] and the number of queries required to achieve a successful attack. To the best of our knowledge, we are the first to validate the *accuracy-robustness trade-off* [14] in specific regards to the fake-news detection task noting that the fake-news detectors should be trained with appropriate regularization to increase the robustness. We discover that the detectors trained with the "binary cross-entropy" loss are slightly more robust. We also note that increasing the detector complexity slightly increases its robustness. Our contributions are summarized next.

- We are the first to study the adversarial robustness of different deep learning architectures and model sizes in specific regards to the fake-news detection. Specifically, we conduct our study on MLP, CNN, RNN, and a recently proposed Hybrid CNN-RNN architecture trained on three distinct datasets (Kaggle fake-news dataset, ISOT dataset, and LIAR dataset).
- To the best of our knowledge, we are the first to study the fake-news detectors against several adversarial attacks under the black-box settings on popular state-of-the-art fake-news datasets. We are also the first to analyze the robustness of fake-news detectors for different input sequence lengths.
- We use the Local Interpretable Model-agnostic Explanations (LIME) explainable AI method to provide a preliminary analysis of why current fake-news detectors are adversarially vulnerable, and discuss key insights for possible future defenses.
- We identify a need for more comprehensive fake-news datasets and more adaptive detection mechanisms scalable to different domains and geographical regions.

We note that we use the term "fake-news" to broadly refer to all types of false information including disinformation (the spread of false information with explicit intent to deceive) or misinformation (the naive spread of false information without explicit malintent). Even though there are various types of Fake-news and the fact that the term has also been politicized, we use the term to refer to "false information" and use further qualification where necessary [15].

The rest of the paper is organized in the following way. Section II provides background and a discussion on related work. Our methodology is introduced in Section III and the results are presented in Section IV. Finally, the paper is concluded in Section V.

II. Background and Related Work

In the recent past, a plethora of adversarial attacks have been proposed that demonstrate the vulnerability of ML-based models in different applications ranging from malware analysis [16], object recognition [17], intrusion detection [18], traffic classification [19], emotion recognition [20], networking applications [21] [22], and self-driving cars [23]. Our focus in this paper is on the adversarial robustness of ML-based

[1]We define robustness as the ratio of correctly classified adversarial inputs to the total number of adversarial inputs.

[2]TextAttack is a Python framework for adversarial attacks, data augmentation, and model training in NLP, textattack.readthedocs.io/en/latest/

[3]ASR is defined as the ratio of incorrectly classified adversarial inputs to the total number of adversarial inputs.

fake news detectors that use ML-based NLP techniques. In this section, we will provide relevant background and discuss salient works related to our work.

A. Threat Model

Current literature on adversarial attacks identifies two different threat models/settings [10].

1) **White-box Threat Model**—in which a powerful adversary is assumed to be fully aware of the model architecture and parameters such as updated weights.
2) **Black-box Setting**—in which an adversary is assumed who is not aware of the model architecture and its weights but is able to query a given model with some input and gets its response. The adversary however is assumed to have a clear understanding of different machine learning architectures and the training methodologies.

We choose the attacks assuming a black-box threat model for our experiments because such attacks are considered less dependant on the model being attacked and more practically applicable [12].

B. Adversarial Attacks

We specifically focus on the recently proposed four different attacks implemented in the Text-attack library. We choose these attacks based on their efficiency, relevance, and recency [10]. Let us assume that an input sequence, X is composed of n words, represented as $\{x_1, x_2, ..., x_i, ..., x_n\}$. Given a classifier, F, the goal of an attacker is to compute a perturbed sequence, $P = \{p_1, p_2, ..., p_i, ..., p_n\} \approx \{x_1, x_2, ..., x_i, ..., x_n\}$, such that $F(X) \neq F(P)$. To achieve this, the attacker usually identifies a set, S_x of important words sorted in the descending order based on how significantly they contribute to an output decision. The attacker then replaces these words, one-by-one in the descending order of their influence, with the perturbed words such that the grammatical and semantic similarity is retained.

Text-Bugger. For a given input sequence, X, such that $F(X) = y$, Text-bugger [24] first identifies key words (S_x) by computing the Jacobian matrix of the classifier for X. For each word in S_x, the attacker generates five perturbations by (a) randomly inserting a *space* in the word (e.g., "stated" becomes *stat ed*), (b) randomly deleting a character, (c) swapping any two unique characters, (d) replacing a character with a visually similar one (e.g., "o" becomes 0, "l" becomes 1, "a" becomes "@") and (e) replacing the word by another semantically similar word. The attacker then chooses the optimal perturbation for each word in S_x based on the maximum reduction in the output score of class, y.

Probability Weighted Word Saliency Attack (PWWS). For a given input sequence, X, such that $F(X) = y$, PWWS [25] first identifies key words (S_x) by substituting an "unknown" word for each word in X, and measuring the change in the output probability of the classifier. Each word in X is then substituted with several synonyms computed using "Word2Net" and the optimal synonym is chosen based on the maximum reduction in the output score of class y. Each word in S_x is then substituted with its optimal synonym (one at a time; best first) until X gets misclassified.

Deep Word Bug. For a given input sequence, X, such that $F(X) = y$, Deep-Word-Bug [26] first identifies key words (S_x) by replacing each word in X by an unknown word. The attack then applies four perturbations, i.e., character swapping, substitution, deletion, and insertion, to each word in S_x (one at a time; best first) until X gets misclassified.

Text-Fooler. For a given input sequence, X, such that $F(X) = y$, Text-Fooler [27] first identifies key words (S_x) by computing the difference between the classifier's prediction score before and after deleting a word from the input. For each word in S_x, the attacker generates "N" perturbations by replacing the word with "N" different words closest to the actual word in a pre-defined Embedding space. The best perturbation is then selected based on the maximum reduction in the output score of class y.

Jin et al. [27] comprehensively study the efficacy of the attack against various deep learning classifiers with varying architectures for different datasets. We note that our work is significantly different in the following ways.

- We attack fake-news detectors using various attack methodologies (including "Text-Fooler") and compare how different attack methodologies affect the adversarial vulnerabilities of the detectors.
- We study how certain architectural and engineering choices, e.g. the input sequence length, the complexity of a detector, the regularization and the loss function used for training can affect the adversarial robustness of the detectors. In doing so, we develop guiding insights for the future researchers to train more robust fake-news detectors which can withstand small adversarial perturbations without affecting their final decision.
- We study the transferability of adversarial news and establish that similar detector architectures show similar adversarial vulnerabilities.
- We do not evaluate the robustness of our fake-news detectors based only on the Attack Success Rate (ASR) of an attacker. Instead, for each attack method that we use, we also compare the robustness of a detector based on the number of queries required and the number of words perturbed by an attacker to fool the network.
- We exploit LIME to explain why the adversarial examples are misclassified by our detectors and show that the strength of adversarial attacks-—minimally perturbing the words to achieve misclassification-—can prove a potential weakness that may be exploited by future researchers for devising effective defense mechanisms. Specifically, we observe that since the aim of an attacker is to fool a detector by perturbing as minimum number of words as possible, for an adversarial input, the decision of a detector is largely contributed to by a small number of words in contrast to a clean input.

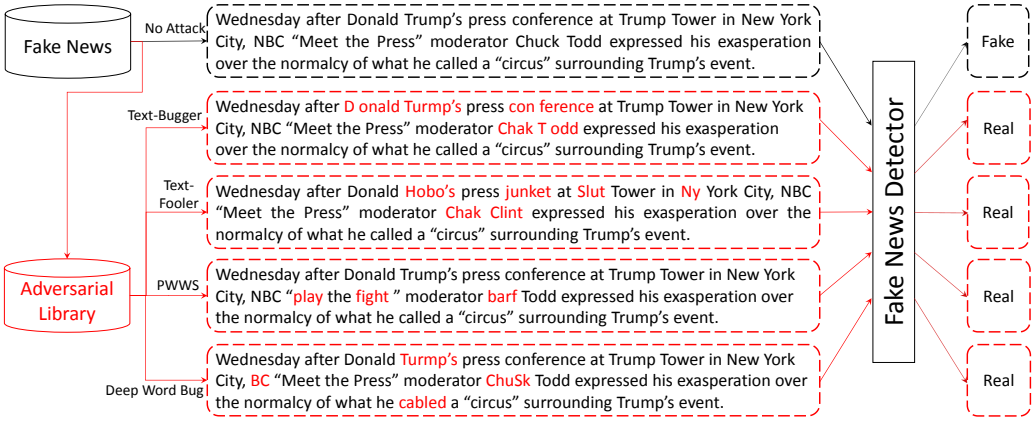


Fig. 2: Adversarial examples generated using different attacks along with the original input for comparison. Perturbed words are highlighted in red.

For illustration, Fig. 2 shows typical adversarial examples generated using the above-mentioned attacks from the Text-attack library. Text highlighted in red shows the changes/perturbations made to the original text by the attack algorithm. Words changed by the PWWS attack are considerably different than those perturbed by other attacks. This is because of different methodologies used to measure the importance of a word.

C. Fake-news Detection

Shu et al. [28] define "fake news" as a verifiably false piece of information shared intentionally to mislead the readers. Currently, there are two popular approaches to fake-news detection, i.e., information propagation-based detection and content-based detection, which are discussed next.

Propagation based techniques exploit the fact that generally public reacts differently to the fake-news than to legit one. Zhao et al. [29] explore these propagation trends in specific regards to how they exhibit themselves at different times since the time of the generation of fake news. They show that the fake news can be identified at about five-hours from the first re-posting in a content- and user-agnostic manner. Monti et al. [3] leverage geometric deep learning to capture the fake-news propagation trends. Although propagation-based schemes provide a general content-agnostic approach to the fake-news detection task, such approaches require large annotated data and several pre-processing techniques to duly model the propagation trends [30].

Content-based detection schemes target either the lexical or the semantic features of the news to identify it as fake or credible. Lexical fake-news indicators include the absence of source URLs, lengthy articles, exaggerated words, emotional patterns, and first/second pronouns [31]–[33].

Utilizing semantic features for fake-news detection is generally considered a better approach due to its universality. Recent works show that fake news can be distinguished based on their semantic features such as topic style, writing style, sentiment analysis, and topic modeling [34]. More recently, Ghanem et al. [4] attempt to model the flow of the article and its relevance to the topic using a CNN and a bidirectional GRU. The CNN performs the topic modeling while the bidirectional GRU extracts the semantic and contextual information from a given input. The outputs of CNN and bidirectional GRU are then concatenated and fed into a classifier for prediction. Another very recent work [1] propose a Hybrid CNN-RNN approach, leveraging both CNN to encode contextual representations and RNN to encode temporal representations for the fake-news detection. The authors show that such a simple CNN-RNN combination provides state-of-the-art classification accuracy on the fake-news detection task. Further, unlike many previous works, the model is generalizable to several datasets [1]. Therefore, we use the same model for analyzing the effect of different hyper-parameters on the robustness of a classifier.

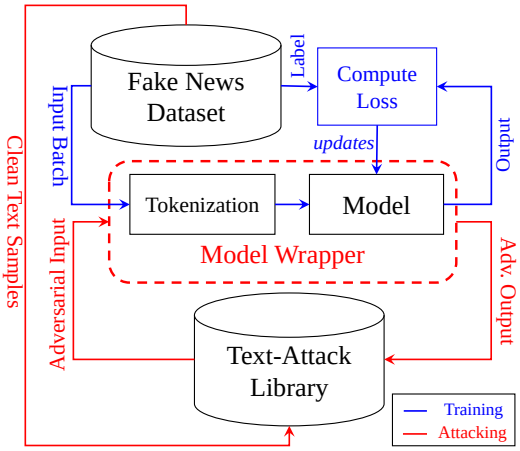


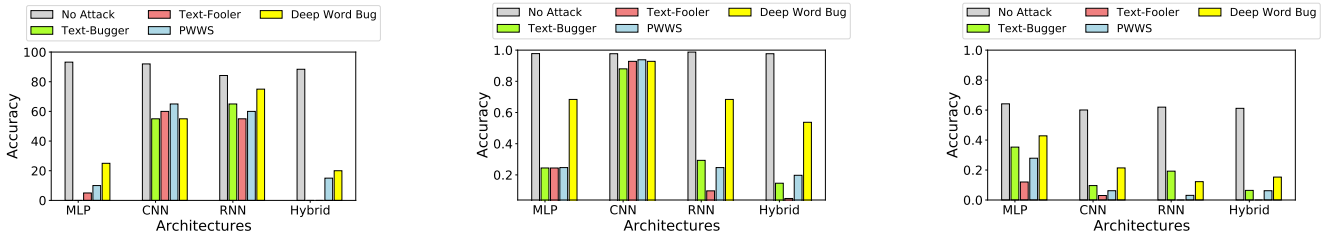
Fig. 3: Experimental setup used in our experiments. For each case, the detector is first trained on some data and then provided to the Text-Attack library in a model-wrapper function along with the original dataset for robustness evaluation. *Blue arrows represent the standard flow, while red arrows represent the adversarial flow.*

III. Methodology

Our experimental setup is illustrated in Fig. 3. To summarize, during training, each input from the dataset is first

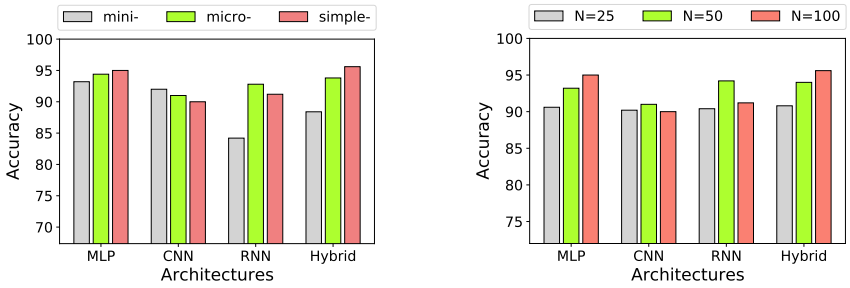
TABLE I: Details different detector architectures used in our experiments. For each detector architecture, we train three variants of different sizes to analyze the effect of detector complexity on the adversarial robustness.

Layer	simple-Hybrid [1]	mini-Hybrid	micro-Hybrid	simple-CNN	mini-CNN	micro-CNN	simple-RNN [35]	mini-RNN	micro-RNN	simple-MLP	mini-MLP	micro-MLP
Embedding	300	300	300	300	300	300	300	300	300	300	300	300
Conv	128×5	128×5	**64**×5	32×3×3	32×3×3	**16**×3×3	-	-	-	-	-	-
Activation	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	-	-	-	-	-	-
MaxPool	2	2	2	-	-	-	-	-	-	-	-	-
Normalize	-	-	-	YES	YES	YES	-	-	-	-	-	-
Dropout	-	-	-	0.4	0.4	0.4	0.3	0.3	0.3	0.8	0.8	0.8
Conv	-	-	-	32×3×3	**16**×3×3	**8**×3×3	-	-	-	-	-	-
Activation	-	-	-	ReLU	ReLU	ReLU	-	-	-	-	-	-
Normalize	-	-	-	YES	YES	YES	-	-	-	-	-	-
Dropout	-	-	-	0.4	0.4	0.4	-	-	-	-	-	-
Conv	-	-	-	32×3×3	**16**×3×3	**8**×3×3	-	-	-	-	-	-
Activation	-	-	-	ReLU	ReLU	ReLU	-	-	-	-	-	-
Dropout	-	-	-	0.4	0.4	0.4	-	-	-	-	-	-
Normalize	YES	YES	YES	YES	YES	YES	-	-	-	-	-	-
LSTM	32	**16**	**16**	-	-	-	**100**	**50**	**50**	-	-	-
Dense	-	-	-	-	-	-	64	**32**	**16**	20	**10**	**5**
Dense+Softmax	2	2	2	2	2	2	2	2	2	2	2	2



(a) Kaggle Fake-News Dataset (b) ISOT Fake-News Dataset (c) LIAR Fake-News dataset

Fig. 4: Comparing the accuracy of different detector architectures before and after the attack (i.e., **"No Attack" vs. four different attack algorithms**) for three state-of-the-art datasets [1]. (Settings: N=100 words, loss: binary cross-entropy.) *CNN and RNN detectors are comparatively more robust than other architectures. MLP architectures give better performance on LIAR dataset because of the smaller input lengths.*



(a) Different detector architectures and complexity. (Settings: N = 100)

(b) Different input lengths (N) and detector architectures.

Fig. 5: Comparison of accuracy of fake news detectors **with different architectures under different (non-adversarial) settings**. *BCE, MSE and CCE denote the binary cross-entropy, mean-squared error and categorical cross-entropy respectively.*

converted into a numerical form (tokenization) using the Keras tokenizer. The output of the tokenizer is fed to the subsequently deployed deep learning model. While attacking, a few samples are chosen from the dataset and provided to the text attack library which perturbs the input and queries the model for various perturbations until either the input is misclassified or the attack algorithm issues a failure.

A. Datasets

We use three different datasets commonly used in literature [1], i.e. Kaggle fake-news dataset, ISOT dataset, and LIAR dataset.

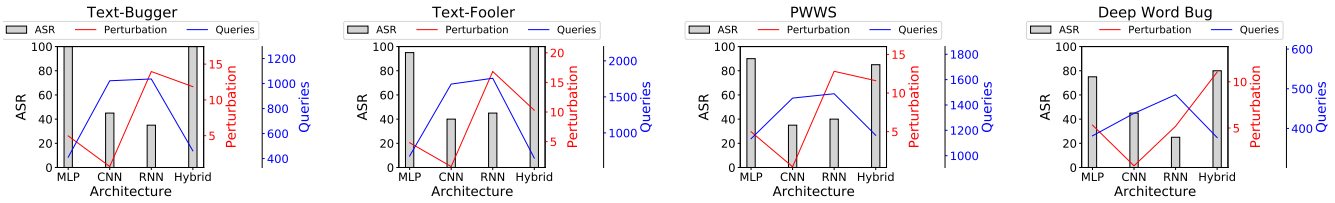


Fig. 6: Evaluating the efficacy of four different attack methods (in terms the Attack Success Rate or ASR, avg. % words perturbed and avg. query counts) **for different deep learning architectures**. *CNN and RNN detectors are more robust than MLP and Hybrid detectors shown by low ASRs and high query counts. CNNs can learn contextual features because of their structure while RNNs are temporally deeper and therefore demonstrate greater robustness.*

1) Kaggle fake-news dataset: We use an openly available dataset[4], "fake news". The dataset contains 26000 sample articles divided into 20800 training samples and 5200 test samples. Each sample is further comprised of different fields, i.e., *id, title, author, text* and *label*. *id* denotes the index of the article from 0 to 26000. *title, author* and *text* denote the topic, writer and the content of the article. The field *label* says whether it is a fake (1) or credible (0).

*2) **ISOT dataset**:* ISOT dataset is an openly available[5] rich dataset containing a total of 44898 samples of which 21417 represent legit news obtained from reuters.com and 23481 represent fake news collected from various sources. Each sample further comprises different fields, i.e. *article, title, text, type* and *date*.

*3) **LIAR Dataset**:* LIAR dataset [36] contains 12836 samples collected from Politifact.com and manually labeled for their truthfulness score. Each sample in the LIAR dataset comprises several fields, i.e. *statement, speaker, context* and *label*.

B. Different DNN Architectures.

We use four different deep learning architectures for robustness evaluation. Specifically, we use the state-of-the-art Hybrid CNN-RNN detector [1] along with simpler CNN, RNN, and MLP architectures. All our detectors, though significantly diverse in architecture, achieve an accuracy comparable to the state-of-the-art Hybrid CNN-RNN detector. The details of each architecture are given in Table I. Following [1], we initialize the embedding layer with the openly available Global Vectors for Word Representations (GloVe) embedding for all cases. Additionally, we experiment with changing the number of learning parameters of the detector and observe how this affects its accuracy and adversarial robustness. Specifically, for each DNN architecture, we introduce two variants (a micro-detector and a mini-detector) ad described in Table I and provide these to the Text-Attack library for adversarial evaluation.

IV. Results

For illustration, we compare the accuracy of different detector architectures—MLP, CNN, RNN and Hybrid CNN-RNN—before and after the attacks in Fig. 4. We observe an alarmingly sharp drop in the accuracy of different detectors, especially for the MLP and the Hybrid architectures. We also observe that CNN and RNN detectors are more robust to adversarial perturbations as compared to other architectures. We attribute the robustness shown by an MLP detector in Fig. 4(c) to significantly smaller length inputs in the case of the LIAR dataset.

In what follows, we specifically discuss the attack results for different settings in greater detail for Kaggle dataset. We choose the Kaggle dataset for research expediency—Kaggle dataset being smaller than ISOT dataset—and cross-dataset generalizability—detectors trained on Kaggle dataset give more than a random-guess performance on ISOT and LIAR datasets (more details in Sec IV-G).

A. Performance Evaluation

The performance of the various ML-based fake news detectors with different architectures under different settings is shown in Figure 5 and is briefly discussed next.

1) Different Detector Architectures: The test accuracy of different detectors used in our experiments on unperturbed inputs is presented in Fig. 5(a). It can be seen that all our detectors perform comparably to the state-of-the-art Hybrid CNN-RNN detector despite their diverse architectures. Note that increasing the number of variables/parameters of a detector generally increases its performance for the task.

2) Different Input Lengths: We experiment with different input lengths. Specifically, we carry out several experiments for a range of architectures by setting the maximum number of words in $\{25, 50, 100\}$. Intuitively, reducing the maximum number of words should decrease the attacker's search space, thus increasing the robustness of the model. However, this may cause a significant drop in the accuracy of the model.

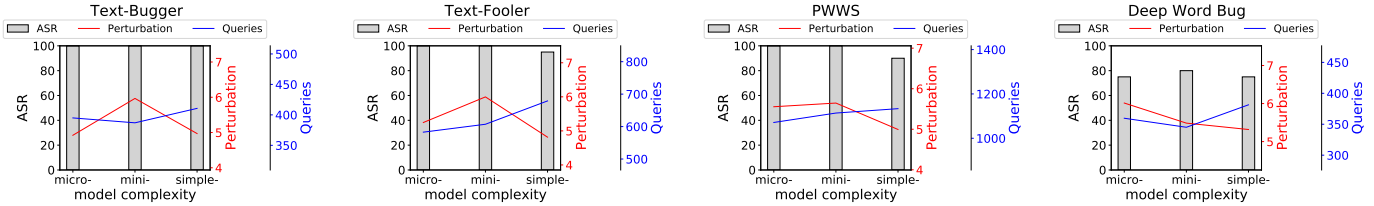
Fig. 5(b) shows that the accuracy of a model increases as the input sequence length is increased. This is intuitive as a longer input should contain more information, thus, allowing the detector to identify fake news relatively better.

B. Adversarial Robustness Evaluation

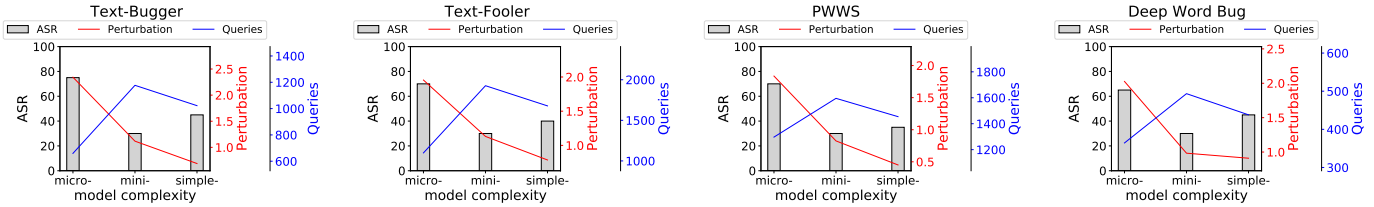
1) Impact of DNN Architecture: Fig. 6 reports results of adversarial attacks on different architectures. Specifically, we sample 20 input sequences that are correctly classified by the trained model from the test. Each input is then clipped so that it only contains the first 100 words. We then adversarially perturb each clipped input using the Text-Attack library against

[4]https://www.Kaggle.com/c/fake-news/data

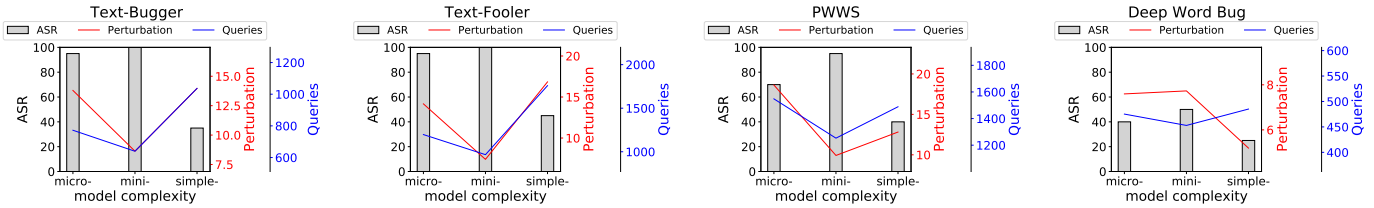
[5]https://www.uvic.ca/engineering/ece/isot/datasets/



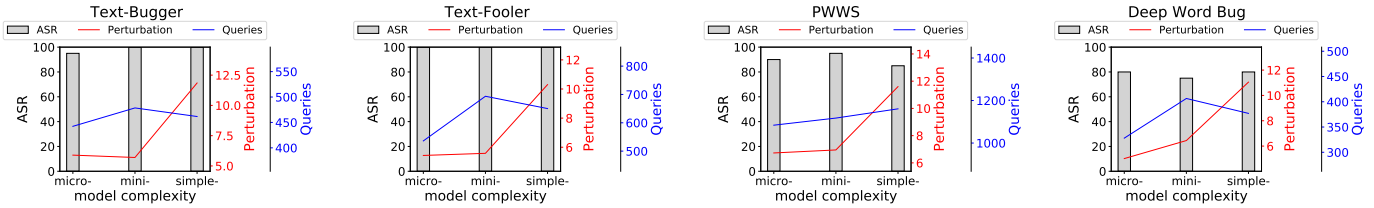
(a) Attack efficacy **with varying MLP detector complexity**. *Increasing the complexity slightly increases the number of queries required.*



(b) Attack efficacy **with varying CNN detector complexity**. *Increasing the complexity slightly increases the robustness of CNN detector.*



(c) Attack efficacy **with varying RNN detector complexity**. *Increasing the complexity slightly increases the robustness of RNN detector.*



(d) Attack efficacy **with varying Hybrid CNN-RNN detector complexity**. *Increasing the complexity slightly increases the robustness of Hybrid CNN-RNN detector.*

Fig. 7: Evaluating the efficacy of attack methods (in terms of ASR, average perturbation, and average query counts) for different deep learning architectures. (Settings: N = 100 words, loss = binary cross-entropy). *Generally, increasing detector complexity increases the robustness as shown by low ASRs and higher number of queries.*

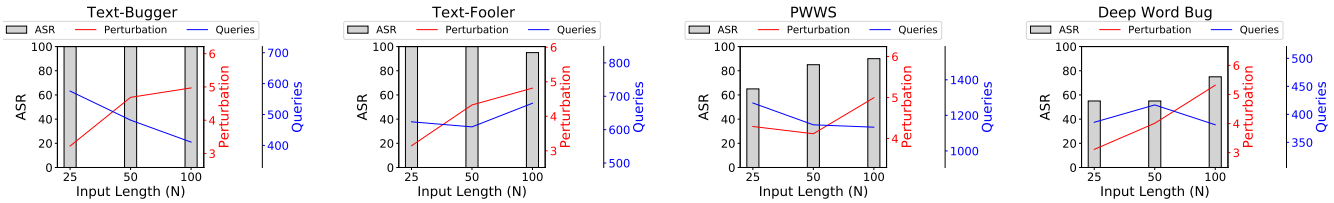
our detectors, i.e., MLP, CNN, RNN, and Hybrid CNN-RNN, and report the average number of queries, the average percent of words perturbed in a given input, and the accumulative Attack Success Rate (ASR).

We observe significantly low ASRs and high query counts for both the CNN and the RNN detectors, suggesting that they are relatively more robust to adversarial perturbation as compared to the MLP and Hybrid CNN-RNN detectors. We also observe that the ratio of words perturbed to perform a successful attack is considerably greater for the RNN-detector as compared to the CNN-detector, suggesting that an RNN-detector is more robust.

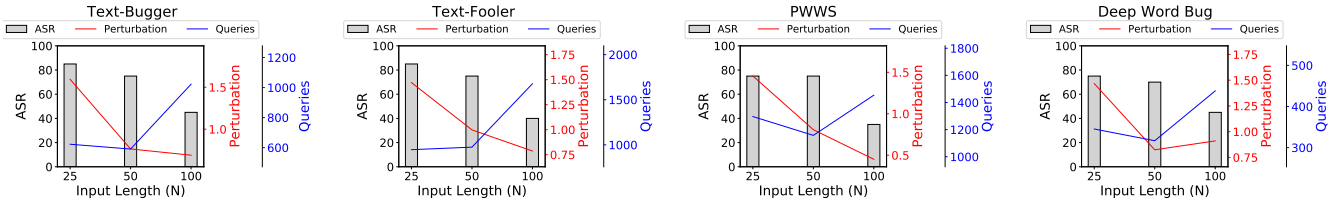
We attribute this to the *accuracy-robustness trade-off*—Su et al. [14] demonstrates how the increased accuracy of DNNs also results in its reduced robustness. Note that MLP-based and Hybrid CNN-RNN-based detectors provide slightly better accuracy as compared to the CNN- and RNN-based detectors. Additionally, the filters of the CNN detector are shared among all the words of an input sequence. Consequently, a CNN detector can contextually learn better features from the input data which are more generalizable to the detection task. This makes it hard for an attacker to change an output decision without drastically changing the input (which may significantly hurt the semantic information and thus is infeasible for the attacker). RNN detectors are temporally deeper and can model long-range temporal features, thus, are more robust to adversarial attacks [37].

2) Impact of the Detector Complexity: Fig. 7 shows how changing the model size may impact the robustness of a detector. We observe that generally increasing the detector parameters also increases the robustness of a detector. For example, in Fig. 7, the ASR decreases as the detector complexity increases while the required number of queries and perturbation rate increases. This effect exhibits itself more strongly for RNN- and CNN-detectors, as compared to other architectures in Fig. 7. However, as observed in [14], the robustness is not critically related to the detector's complexity, as evident by a number of exceptions from the general trend.

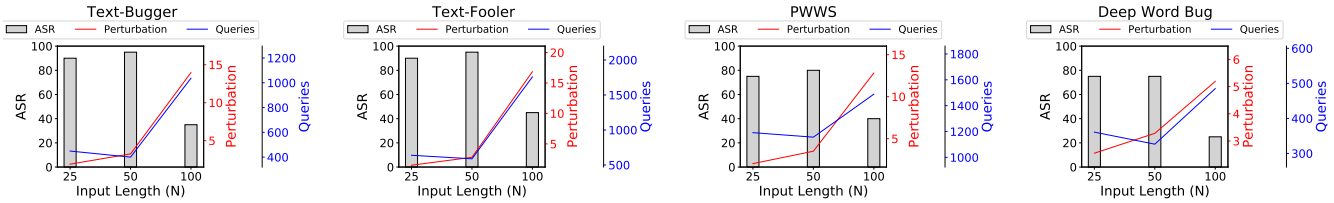
3) Impact of Changing the Input Length: Fig. 8 shows that a longer input results in a more robust model, as evident by a reduction in ASR, and an increase in the perturbation ratio and required number of queries. This is because a



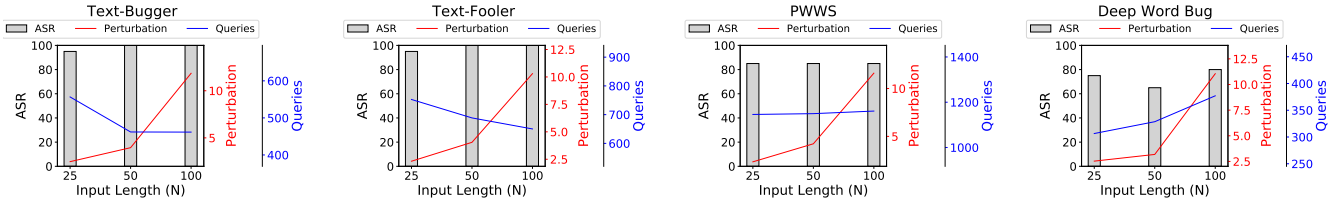
(a) **MLP detectors** with different input lengths.



(b) **CNN detectors** with different input lengths.



(c) **RNN detectors** with different input lengths.



(d) **Hybrid CNN-RNN detectors** with different input lengths.

Fig. 8: Evaluating the efficacy of attack methods (in terms of ASR, average perturbation, and average query counts) for different deep learning architectures ***with different input lengths*** **assuming** ***binary cross-entropy loss***. *Increasing input length increases the robustness, as evident by low ASRs and high perturbation, because a longer sentence contains more information.*

longer input sequence is easier for a detector to identify as either fake or credible due to a more effective information representation. The increase in query counts can however be partially attributed to larger search space for an attacker caused by longer inputs. Again, we observe that RNN- and CNN-detectors are considerably more robust as compared to other architectures.

4) Different Loss Functions: For this experiment we train a Hybrid RNN-CNN for three commonly used loss functions in literature, i.e., *Mean Square Error (MSE), Binary Cross-entropy (BSE)* and *Categorical Cross-entropy (CCE)*. We find out that *binary cross-entropy* based training gives the best performance in terms of the detector's accuracy.

We present the models to the Text-Attack library for adversarial evaluation. Results are given in Fig. 9. The figure suggests that training a detector with *binary cross-entropy* results in a slightly more robust model.

C. BERT Adversarial Example (BAE) Attack

BERT Adversarial Example (BAE)-attack [38] works by first identifying important words in a given input. The importance of each word is estimated by deleting the word from a given input sequence and measuring the decrease in the probability of the original class. The most important word is then masked and given to BERT Masked-Language Model (BERT-MLM) which fills in the mask by generating alternative words while the semantic structure of the sentence. Universal Sentence Encoder (USE) [39] is then used to select the most optimal of these alternates based on their cosine similarity with the original input.

We use BAE-attack to attack the four detectors of different architectures as given in Table I. For this experiment, we set the input sequence length to be 100 and train the network using "binary cross-entropy" (BCE) loss. The results of our experiments are shown in Fig. 11.

Although BAE-attack generates more natural adversarial examples, we observe that the BAE-attack has a lower ASR

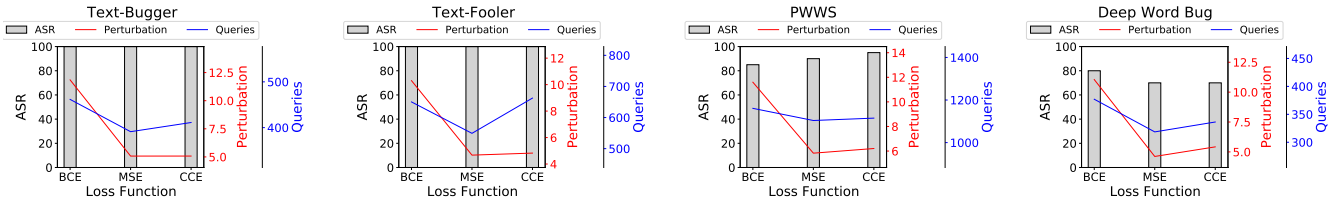
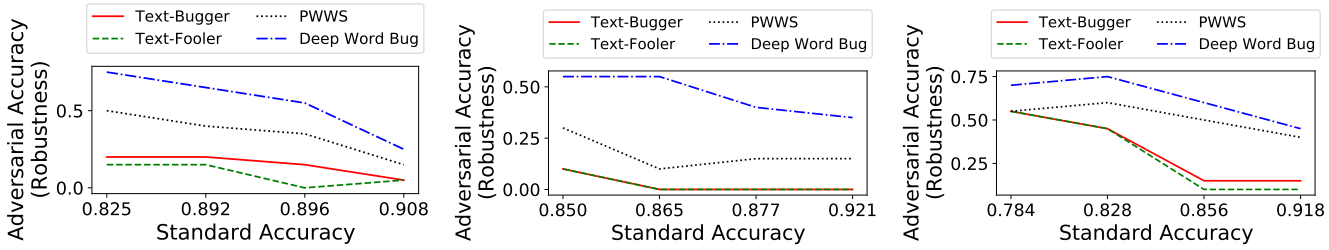


Fig. 9: Evaluating the efficacy of four different attack methods (in terms of ASR, average perturbation, and average query counts) on Hybrid CNN-RNN **with different loss functions** for different deep learning architectures. (Settings: Model: Hybrid CNN-RNN, N = 100 words). *Binary cross-entropy results in a more robust model being trained as compared to other losses.*



(a) Input length, N = 25 (b) Input length, N = 50 (c) Input length, N = 100

Fig. 10: Accuracy robustness trade-off of a Hybrid CNN-RNN fake-news detector for different input lengths for l_2-regularization varying from the strongest(left-most) to the weakest(right-most). (Settings: loss=binary cross-entropy). *The robustness of the fake-news detector increases as the accuracy increases.*

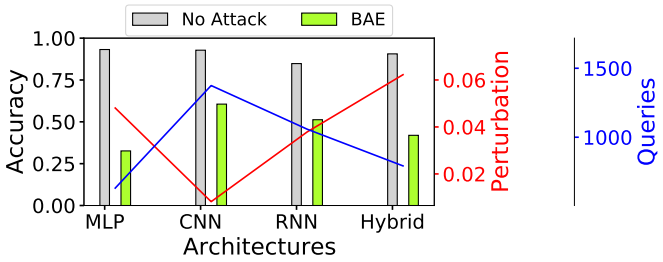


Fig. 11: Evaluating the efficacy of BAE (in terms of ASR, average perturbation, and average query counts) for different deep learning architectures. (Settings: N = 100 words, loss = binary cross-entropy). *Increasing input length increases the robustness, as evident by low ASRs and high perturbation, because a longer sentence contains more information.*

as compared to other attacks, i.e. PWWS, Text-Fooler, Text-Bugger. We attribute the reduced ASR of BAE-attack to a stronger constraint of preserving the natural structure of a sentence as compared to other attacks. *Therefore, we recommend future researchers to evaluate future models/defenses against Text-Fooler, Text-Bugger and PWWS attacks.* As observed previously, we note that CNN-based detectors are more robust than other architectures. This is evident by their higher accuracy against adversarial examples and larger query counts as compared to other architectures.

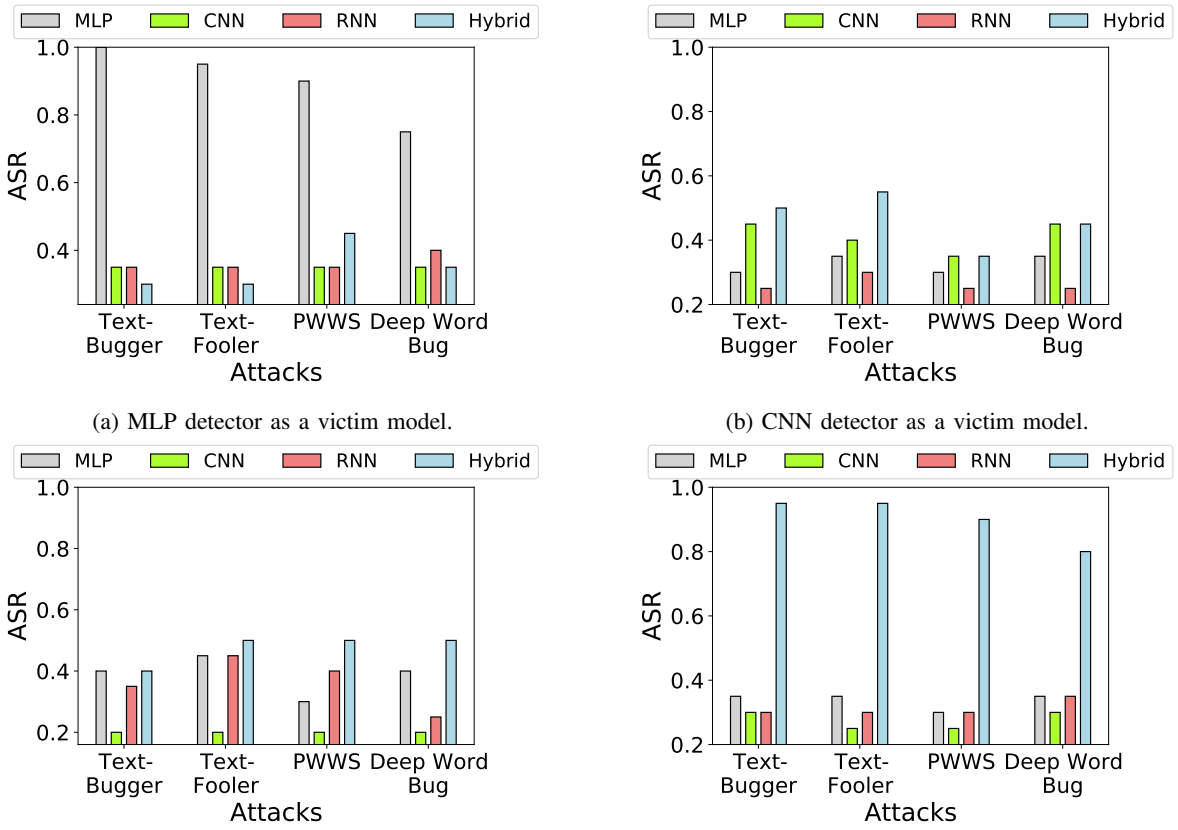
D. Accuracy-Robustness Tradeoff Evaluation

Many previous works analyze the accuracy-robustness trade-off on visual tasks. However, to the best of our knowledge, we do not find any work validating the phenomenon, specifically for fake-news detection. Following [14], [40], we experiment with different l_2-regularization strengths and report the robustness-accuracy curve in Fig. 10 for Hybrid CNN-RNN detectors of varying input lengths—a stronger regularization causes the accuracy of a detector to drop slightly while more effectively resisting the over-fitting. Evidently, we observe that as the accuracy of a detector increases, the robustness decreases as illustrated in Fig. 10 by a decrease in the adversarial accuracy. Additionally, we note that although the trade-off phenomenon generally exhibits itself irrespective of the input length chosen, the decrease in the robustness is comparatively more drastic for larger input lengths. For example, for N = 25, the maximum decrease in the adversarial accuracy of the detector against Text-Bugger is 10%, contrary to N = 100, where the maximum decrease in the adversarial accuracy is around 30%. We attribute this to a larger adversarial space—the number of words an adversary may perturb while attacking—available to the adversary for larger input lengths.

E. Transfer Adversarial Attacks

It has been observed that adversarial examples generated against one machine learning model can be transferred to fool a different ML model. Such attacks are more commonly known as Transfer adversarial attacks. Transfer attacks have been effectively used to analyze the robustness of ML models, specifically under black-box settings. Additionally, transfer attacks can also be used to evaluate the similarity between different ML models.

Here, we analyze the vulnerabilities of different fake-news detectors under transfer attack settings. Specifically, for each



(a) MLP detector as a victim model.

(b) CNN detector as a victim model.

(c) RNN detector as a victim model.

(d) Hybrid CNN-RNN detector as a victim model.

Fig. 12: Evaluating various detectors against **adversarial examples generated for different victim detector architectures** using different attacks from the Text-Attack library. (Settings: N=100 words). *Due to similar architectures, adversarial examples generated against CNN- and RNN- detectors transfer well to the Hybrid CNN-RNN detector.*

detector architecture, we generate adversarial examples using Text-Attack library and see if these adversarial examples are also misclassified by other detector architectures.

Results are shown in Fig. 12. We observe that compared to the clean samples, different detectors are, on average, less accurate against the adversarial examples transferred from a different architecture. However, the attack success rate may significantly be reduced while transferring. For example, in Fig. 12, adversarial examples generated for MLP detector show significantly lower attack success rates for other architectures. We say that adversarial examples for MLP classifiers do not transfer upon different architectures effectively.

We also note that the adversarial examples generated for CNN and RNN detectors are highly effective when used against the Hybrid RNN-CNN architecture. For example, in Fig. 12, adversarial examples generated for CNN-detector show high ASR against Hybrid CNN-RNN detector. We attribute this to the similarity of architectures, i.e. because Hybrid CNN-RNN architecture comprises both the convolutional and the recurrent layers, the similarity in architectures allows the adversarial examples to transfer more effectively.

TABLE II: The frequency of occurence of various combination of words in the dataset.

Real News		Fake News	
the united states	29	the white house	12
the U.S.	22	the united states	11
one of the	21	president donald trump	10
donald j trump	11	donald j trump	7
new york times	11	according to the	7
of the united	8	october 27 2016	7
the new york	8	one of the	6
in the united	8	out of the	6
of the most	8	a lot of	6
to be a	7	it is a	6
president of the	7	pic twitter.com	5

F. Discussion

1) Bias in the dataset: Adversarial examples may also be caused by natural bias in the dataset. We illustrate this with a simple technique by analyzing the "Fake-News" dataset based on the frequency of occurrences of bag-of-words. Specifically, we identify all possible sets of three words in the training set and report the frequency of their occurrences in Table II. The table shows that many word combinations occur much

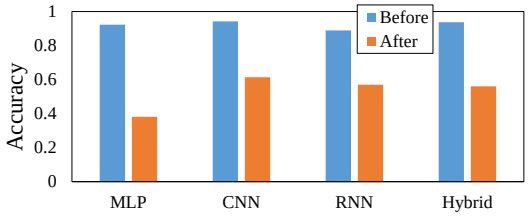


Fig. 13: Accuracy of different fake-news detectors on fake-news from the test set **before and after concatenating a manually constructed sentence—"President of the United States Donald J. Trump told the New York Times"**. (Settings: N=100 words, loss = binary crossentropy). *Bias in the dataset can lead to adversarial behavior at test time*.

frequently in the real news as compared to the fake news. For example, "the United States" occurs 29 times in real news articles in contrast with 11 times in the fake news, for the dataset provided. Similarly, "New York Times" occurs 11 times in the real news while less than 5 times in the fake news. Such biases may readily be learned by the detectors, thus, poisoning the resulting model.

To illustrate the adversarial effects of bias, we manually generate a sentence using a combination of those words which frequently occur in the real news for the provided dataset. More specifically, we use the following sentence; "President of the United States Donald J. Trump told the New York Times". This sentence is concatenated with all the fake news in the test set, which are then provided to different detectors for predictions. Results are shown in Fig. 13.

It is evident from the figure that the accuracy of a fake-news detector is highly compromised by such a simple, yet effective, perturbation method. The effect is more considerable for the MLP detectors. We believe that ensuring the fairness of the dataset and the training algorithm should result in more robust fake-news detectors.

2) Explaining Adversarial Examples using LIME: Local Interpretable Model-Agnostic Explanations (LIME) is a popular model-agnostic explainability technique [41]. To explain the predictions of any model (classifier or regressor), it highlights the words in input text by understanding the relationship between input text and prediction by learning a local linear interpretable model around the prediction. A local linear model is approximated by using sparse linear estimation and performing the search using input text perturbations. LIME explains the decision of the model for a particular input by representing the local importance of interpretable components of an input.

In order to further explain and analyze the adversarial example phenomena, we use LIME to generate explanations for the decisions made by the state-of-the-art Hybrid CNN-RNN detector. More specifically, we use the Text-Attack library to generate the adversarial examples by perturbing a correctly classified inputs such that the perturbed input is misclassified. The original inputs and their corresponding adversarial inputs are provided to LIME which explains why the detector is making a particular decision.

Fig. 14 illustrates a typical case where an original input (correctly classified by the model as fake) along with its corresponding adversarial examples are explained using LIME. Words suggesting potentially fake content are highlighted in "orange", while those indicating that the news is real are highlighted in "blue". We observe that a major strategy of different attacks is to identify important words and replace them with other words, preferably those unknown to the dictionary. As this can be achieved by simple character substitution or space insertion, the adversary can successfully launch an attack in an inconspicuous manner.

Additionally, we observe that for a clean input, the number of words contributing to the final decision is significantly larger than the number of words causing an incorrect decision. For example, in Fig. 14(a), 9 out of 10 most contributing words agree with the final decision, i.e. fake. For an adversarial input; however, this number is much smaller—for Text-Bugger, only 3 out of the 10 most contributing words agree with the final decision, i.e. real.

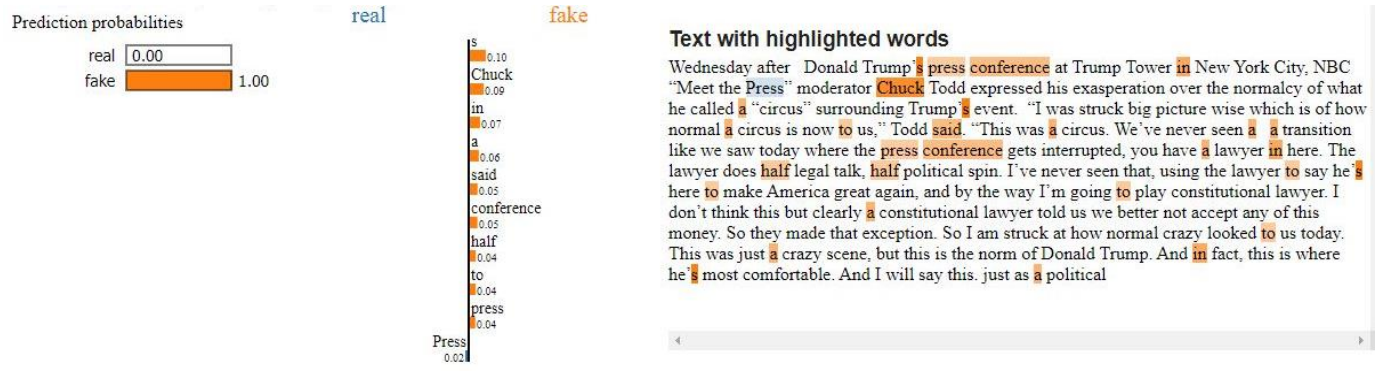
3) Future Defense Techniques: Depending on the attack algorithm, the optimal perturbations may vary (as was illustrated in Fig. 2). In the case of the Text-Bugger attack, the perturbations introduced by the attacker usually cause misspellings, which can simply be detected by a spell-checker. Alternatively, one may intentionally perturb the training set at random to include misspellings as a data augmentation methodology to robustify an NLP model.

Our experiments suggest the effectiveness of developing future adversarial defenses based on RNN architectures. Also, longer input lengths allow the detector to learn better representations, thus, increasing both the accuracy and the robustness of a fake-news detector. Additionally, using strong regularization techniques—e.g., the l_2-regularization used in Fig. 10—can significantly robustify a fake-news detector, though, with a decrease in accuracy on original inputs. However, such a decrease in the accuracy of the detector can be addressed by strong data augmentation techniques as studied in recent works for the visual task [42], [43]. We leave exploring appropriate data augmentation techniques as future work.

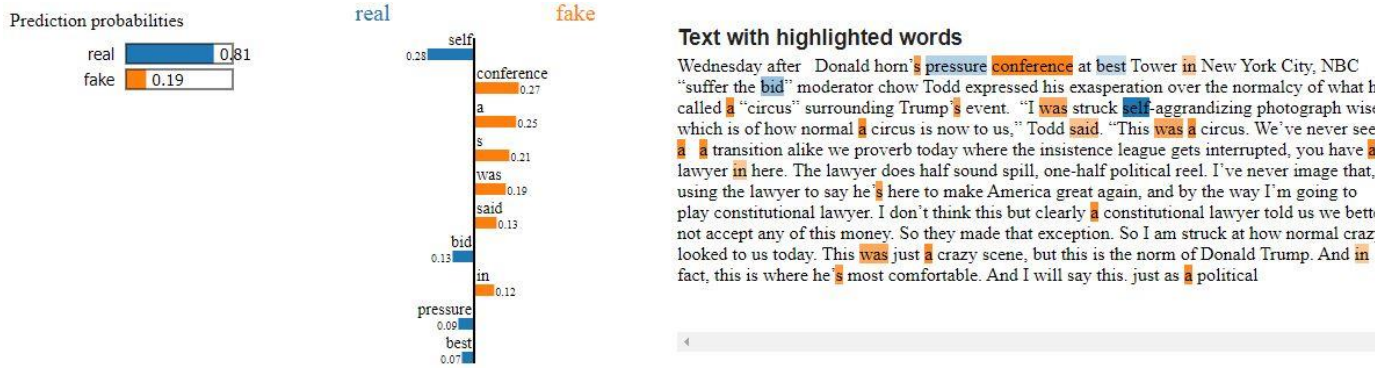
We also observe that for clean/uncompromised inputs, the final decision is in agreement with a number of contributing words, contrary to the case with adversarial inputs, where the final decision is usually credited to a few out of many input words. This insight can be used to counter adversarial example threat in an efficient and effective manner. We strongly recommend that future defenders should follow the guidelines suggested by Carlini et al. [44] to fairly evaluate the defense mechanism.

G. Caveats of Current Fake-News Detectors

1) Cross dataset analysis: We analyze how accurately a fake-news detector trained on one dataset generalizes to other datasets. Results are shown in Fig. 15. In Fig. 15(a), we train detectors of different architectures on Kaggle dataset and report the accuracies against ISOT and LIAR datasets. We observe that detectors trained on LIAR dataset give a random

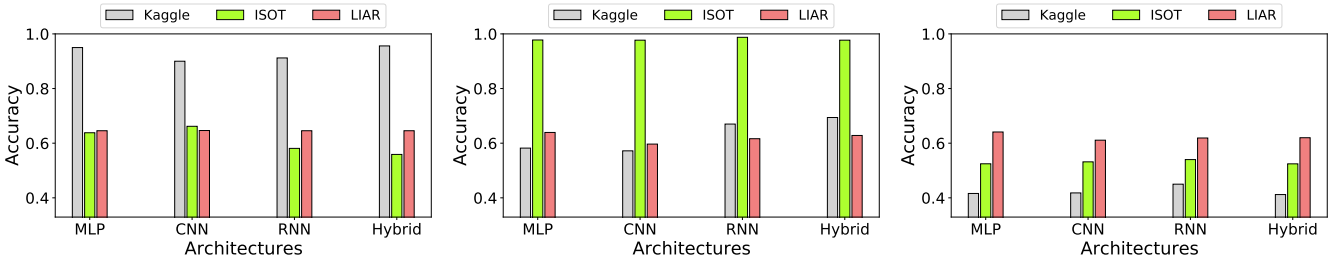


(a) Original input.



(b) Adversarial input.

Fig. 14: Comparing LIME explanations for a clean and the corresponding adversarial input generated using PWWS attack against the Hybrid CNN-RNN detector. (Settings: N=100 words, detector = Hybrid CNN-RNN). *For an adversarial input, the number of words agreeing with the final decision is significantly less than that for a clean input.*



(a) Kaggle fake-news dataset (b) ISOT dataset (c) LIAR dataset

Fig. 15: Cross-dataset accuracy of different detector architectures trained for only one dataset. (Settings: loss=binary cross-entropy, N=100 words). *Detectors trained for kaggle dataset also, to some extent, generalize on ISOT dataset and vice versa. Detectors trained on LIAR dataset do not generalize well to other datasets.*

guess performance ($\approx 50\%$) when evaluated on Kaggle and ISOT datasets. This is because the LIAR dataset contains short sentences (smaller input lengths) and a far lesser number of samples as compared to other datasets. However, detectors trained on Kaggle and ISOT dataset somehow generalize to other datasets with performance slightly better than a random guess performance.

2) News from different geographical regions/domains: All the datasets used in this paper include the news related to the USA. To see how these detectors respond to news from a different domain or geographical region, we present the state-of-the-art Hybrid CNN-RNN detector with authentic news from two countries—Pakistan and Saudi Arabia. We find that the Hybrid CNN-RNN detector labels all of these news (collected from authentic sources) as fake. We observe that, unlike information propagation-based methods, current *content-based fake-news detectors* are fragile and fail to correctly classify news from different geographical regions and backgrounds. This potential direction is identified for future researchers to develop robust detectors scalable to a number of domains and

regions.

3) Adaptive fake-news detectors: The definition of information critically depends on specific scenarios and geography, and rapidly changes with time. We believe that there is a need to develop more generic fake-news detection methodologies capable of adapting to such changing scenarios in an effective way—e.g., by modeling the information provided in input and validating it on some dictionary/encyclopedia updated on daily basis. Studying such detectors under the adversarial threat should provide a better understanding of the vulnerabilities of current fake-news detectors.

V. Conclusions

In this work, we analyze the robustness of fake-news detectors to black-box adversarial attacks. For this purpose, we use four different architectures—multi-layer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and a recently proposed Hybrid CNN-RNN fake news detector—and multiple datasets—Kaggle fake-news dataset, ISOT dataset and LIAR dataset. We vary the complexity of detectors and experiment with different input lengths and loss functions. Our findings suggest that CNNs provide the most robust solution closely followed by RNNs. Further, training the detector for lengthy inputs using binary cross-entropy loss can significantly robustify it against adversarial attacks. In the future, we plan to propose a robust defense against adversarial attacks based on our findings in this paper.

VI. Acknowledgement

This research has been funded by Deputy for Research & Innovation, Ministry of Education through Initiative of Institutional Funding at University of Ha'il-Saudi Arabia through project number IFP-2004.

References

[1] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, 2021.

[2] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018* (Y. Guo and F. Farooq, eds.), pp. 849–857, ACM, 2018.

[3] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *CoRR*, vol. abs/1902.06673, 2019.

[4] B. Ghanem, S. P. Ponzetto, P. Rosso, and F. Rangel, "Fakeflow: Fake news detection by modeling the flow of affective information," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021* (P. Merlo, J. Tiedemann, and R. Tsarfaty, eds.), pp. 679–689, Association for Computational Linguistics, 2021.

[5] F. Khalid, H. Ali, H. Tariq, M. A. Hanif, S. Rehman, R. Ahmed, and M. Shafique, "Qusecnets: Quantization-based defense mechanism for securing deep neural network against adversarial attacks," in *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, pp. 182–187, IEEE, 2019.

[6] H. Ali, S. Nepal, S. S. Kanhere, and S. Jha, "Has-nets: A heal and select mechanism to defend dnns against backdoor attacks for data collection scenarios," *arXiv preprint arXiv:2012.07474*, 2020.

[7] A. C. Serban and E. Poll, "Adversarial examples - A complete characterisation of the phenomenon," *CoRR*, vol. abs/1810.01185, 2018.

[8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387, IEEE, 2016.

[9] H. Ali, F. Khalid, H. A. Tariq, M. A. Hanif, R. Ahmed, and S. Rehman, "Sscnets: Robustifying dnns using secure selective convolutional filters," *IEEE Design & Test*, vol. 37, no. 2, pp. 58–65, 2019.

[10] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020* (Q. Liu and D. Schlangen, eds.), pp. 119–126, Association for Computational Linguistics, 2020.

[11] F. Khalid, H. Ali, M. A. Hanif, S. Rehman, R. Ahmed, and M. Shafique, "Fadec: A fast decision-based attack for adversarial machine learning," in *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pp. 1–8, IEEE, 2020.

[12] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (J. G. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 274–283, PMLR, 2018.

[13] Z. Zhou, H. Guan, M. M. Bhat, and J. Hsu, "Fake news detection via NLP is vulnerable to adversarial attacks," in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence, ICAART 2019, Volume 2, Prague, Czech Republic, February 19-21, 2019* (A. P. Rocha, L. Steels, and H. J. van den Herik, eds.), pp. 794–800, SciTePress, 2019.

[14] D. Su, H. Zhang, H. Chen, J. Yi, P. Chen, and Y. Gao, "Is robustness the cost of accuracy? - A comprehensive study on the robustness of 18 deep image classification models," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), vol. 11216 of *Lecture Notes in Computer Science*, pp. 644–661, Springer, 2018.

[15] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe report*, vol. 27, pp. 1–107, 2017.

[16] M. Usama, J. Qadir, and A. Al-Fuqaha, "Adversarial attacks on cognitive self-organizing networks: The challenge and the way forward," in *2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops)*, pp. 90–97, IEEE, 2018.

[17] 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 *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[18] M. Usama, M. Asim, S. Latif, J. Qadir, *et al.*, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *2019 15th international wireless communications & mobile computing conference (IWCMC)*, pp. 78–83, IEEE, 2019.

[19] M. Usama, A. Qayyum, J. Qadir, and A. Al-Fuqaha, "Black-box adversarial machine learning attack on network traffic classification," in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 84–89, IEEE, 2019.

[20] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," *arXiv preprint arXiv:1811.11402*, 2018.

[21] M. Usama, J. Qadir, A. Al-Fuqaha, and M. Hamdi, "The adversarial machine learning conundrum: Can the insecurity of ml become the achilles' heel of cognitive networks?," *IEEE Network*, vol. 34, no. 1, pp. 196–203, 2019.

[22] M. Usama, R. Mitra, I. Ilahi, J. Qadir, and M. Marina, "Examining machine learning for 5g and beyond through an adversarial lens," *IEEE Internet Computing*, 2021.

[23] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 998–1026, 2020.

[24] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*, The Internet Society, 2019.

[25] S. Ren, Y. Deng, K. He, and W. Che, "Generating natural language adversarial examples through probability weighted word saliency," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (A. Korhonen, D. R. Traum, and L. Màrquez, eds.), pp. 1085–1097, Association for Computational Linguistics, 2019.

[26] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pp. 50–56, IEEE Computer Society, 2018.

[27] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8018–8025, AAAI Press, 2020.

[28] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor.*, vol. 19, no. 1, pp. 22–36, 2017.

[29] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin, "Fake news propagates differently from real news even at early stages of spreading," *EPJ Data Sci.*, vol. 9, no. 1, p. 7, 2020.

[30] L. Tian, X. Zhang, Y. Wang, and H. Liu, "Early detection of rumours on twitter via stance transfer learning," in *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I* (J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, eds.), vol. 12035 of *Lecture Notes in Computer Science*, pp. 575–588, Springer, 2020.

[31] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011* (S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, eds.), pp. 675–684, ACM, 2011.

[32] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017* (M. Palmer, R. Hwa, and S. Riedel, eds.), pp. 2931–2937, Association for Computational Linguistics, 2017.

[33] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, pp. 1–18, 2020.

[34] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu, "The future of false information detection on social media: New perspectives and trends," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 68:1–68:36, 2020.

[35] "Best submission." https://www.kaggle.com/khanrahim/fake-news-classification-easiest-99-accuracy. Accessed: 2021-02-27.

[36] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers* (R. Barzilay and M. Kan, eds.), pp. 422–426, Association for Computational Linguistics, 2017.

[37] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pp. 5695–5699, IEEE, 2016.

[38] S. Garg and G. Ramakrishnan, "Bae: Bert-based adversarial examples for text classification," *arXiv preprint arXiv:2004.01970*, 2020.

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

[40] A. Bietti, G. Mialon, and J. Mairal, "On regularization and robustness of deep neural networks," *CoRR*, vol. abs/1810.00363, 2018.

[41] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[42] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk, "Improving robustness without sacrificing accuracy with patch gaussian augmentation," *arXiv preprint arXiv:1906.02611*, 2019.

[43] E. Borgnia, V. Cherepanova, L. Fowl, A. Ghiasi, J. Geiping, M. Goldblum, T. Goldstein, and A. Gupta, "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff," *arXiv preprint arXiv:2011.09527*, 2020.

[44] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.