Evaluating the Robustness of Fake News Detectors to Adversarial Attacks with Real User Comments (Extended Abstract)

Annat Koren
City College of San Francisco
San Francisco (CA), USA
koren@mail.ccsf.edu

Chandler Underwood
Boise State University
Boise (ID), USA
chandlerunderwoo@u.boisestate.edu

Edoardo Serra
Boise State University
Boise (ID), USA
edoardoserra@boisestate.edu

Francesca Spezzano
Boise State University
Boise (ID), USA
francescaspezzano@boisestate.edu

2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA) | 979-8-3503-6494-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/DSAA61799.2024.10722837

***Abstract*—The widespread use of social media has led to an increase in false and misleading information presented as legitimate news, also known as fake news. This poses a threat to societal stability and has led to the development of fake news detectors that use machine learning to flag suspicious information. However, existing fake news detection models are vulnerable to attacks by malicious actors who can manipulate data to change predictions. Research on attacks on news comments is limited, and current attack models are easily detectable. We propose two new attack strategies that instead use real, pre-existing comments from the same dataset as the news article to fool fake news detectors. Our experimental results show that fake news detectors are less robust to our proposed attack strategies than existing methods using pre-existing human-written comments, as well as a malicious synthetic comment generator.**

***Index Terms*—misinformation, adversarial machine learning, machine learning robustness**

I. Introduction

The widespread use of social media has led to the rise of fake news, or false and misleading information presented as if it is legitimate. Misinformation and disinformation erode trust in the press, promote conspiracy, exacerbate polarization, and can also lead to poor policy decisions. The increased prevalence of fake news poses a threat to societal stability and has led to the development of fake news detectors. These detectors use machine learning to mitigate the problem of fake news by using news content and in some cases user comments to flag suspicious information. While many fake news detection models have been developed, it is possible for malicious actors to manipulate data to change the predictions of fake news detectors [1]–[3]. Several attacks have been developed to fool detectors into misclassifying news pieces, for instance by modifying the news content [3]–[5].

Research that considers attacks on comments made by readers of the news is extremely limited. To the best of our knowledge, only one such attack model, MALCOM [6], has been developed, where malicious comments are synthetically generated based on headlines and user comments to fool fake news detectors. However, these machine-generated attacks are easily detectable. Our goal is therefore to study the effectiveness of attacks that use pre-existing, human-generated comments on fake news detection models.

In this paper, we propose two new attack strategies by using comments from the same dataset as the target article, rather than generating text to do so. Our proposed attack strategies retrieve comments from the dataset that are either relevant to the content of the targeted news article (topic-specific comments) or generic. In both cases, we take into account the influence of the selected comments in the fake news detection process. Because the comments are genuine, such attacks are less likely to be flagged as adversarial examples. In our attack scenario, the attacker aims to change the classification result of the victim fake news detection model, by adding a single user comment. The attack comment should be selected while limiting the number of queries made on the victim model. Hence, in our work, we also perform experiments using surrogate models, to eliminate the need for querying the target model directly.

Experimental results show that fake news detectors are less robust to our proposed attack strategies than existing methods that consider attacks based on pre-existing human-written comments and machine-generated comments.

II. Methodology

We use the GossipCop and PolitiFact datasets from the FakeNewsNet [7] repository and attack three fake news detectors: dEFEND [8], TextCNN [9], and RoBERTa [10]. Our data preparation methods and model training are consistent with the work of Le et al. [6]. To understand the relationship comments have with the classification of news articles, we define the *influence* of each comment c, individually, on its

respective news article a, as the difference in the model's confidence that a is fake when a has no comments present and when a has only the comment c.

To fool the fake news detector, we use real comments sourced from the dataset itself, as opposed to generating synthetic comments, which can be identified with 99% accuracy. Attacking a classifier involves adding comments to correctly classified articles in the test set: this includes true positives (correctly identified fake news) and true negatives (correctly identified real news). An attack is successful if it causes the classifier to change its prediction from correct to incorrect, either from fake to real or real to fake. We consider two types of attacks: *fake news promotion* (i.e., forcing misclassifications of true positives as real news articles), and *real news demotion* (i.e., changing the prediction of true negatives).

We propose two approaches to retrieve comments to perform the attack: (a) retrieving **topic-specific comments**, i.e., comments that are thematically similar to the targeted news article; and (b) retrieving **generic comments**, which are highly dissimilar from their respective news articles. We selected h comments of the desired type (topic-specific or generic) from news articles with the opposite ground truth classification (real or fake) of the target (victim) article. This was done to avoid variation caused by selecting just one comment at random (as done in [6]), and increase the attack's chance of success.

In both approaches (topic-specific or generic comments) we first filter for comments with an individual influence in the desired direction of change. For instance, if we want to promote fake news, then the target article will have a ground truth of fake, and the set of attack comment candidates will consist of real comments that make their respective articles ***less fake***, a negative percent change in individual comment influence. We can also find comments with a more pronounced effect on the classification of their original article by selecting those with individual influence above a given threshold magnitude θ (i.e., within one tail of the histogram). We use this approach to create a list of candidate comments to use in an adversarial attack. Other existing methods for retrieving real comments to attack a fake news detector, i.e., CopyCat and its variants [6], do not consider individual comment influence and thus may be less effective.

III. Experimental Results

We compare the success of an attack of our proposed topic-specific and generic comments attack-based procedures against CopyCat and MALCOM [6] state-of-the-art methods. The success of an attack is measured by the **attack success rate**, defined as the percentage of articles correctly classified before the attack for which the classifier changes its prediction after the attack.

Results show that our proposed topic-specific or generic comment attacks match or outperform CopyCat and MALCOM in nearly all the considered cases. As expected, we found out that models with lower fake news detection accuracy (in particular, TextCNN with 69% accuracy) are easier to fool. Moreover, despite RoBERTa and dEFEND having comparable classification performance (more than 80% accuracy), results show that a classifier specially designed to detect fake news such as dEFEND is more robust to adversarial attacks than a generic model such as RoBERTa. We also observe that topic-specific comments are particularly effective for fake news promotion, and attacks are more effective when there are no initial comments (which is expected, as the classifier accuracy is lower in this case).

Another way to perform an attack is to perform a black box attack through a surrogate model. Hence, we performed further experiments considering RoBERTa, TextCNN, and an RNN (Recurrent Neural Network) based model as the surrogate fake news detection models and assume the adversary has access to the datasets. To perform the attack with a surrogate model, we generated comments assuming the surrogate as the fake news detection model and then measured the attack success rate of such an attack on the considered fake news detectors (dEFEND, TextCNN, and RoBERTa). Our experimental results clearly show that using the surrogate model when computing topic-specific or generic attack comments achieves an attack success rate higher or comparable to CopyCat and MALCOM in nearly all the considered cases.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014.

[3] C. Koenders, J. Filla, N. Schneider, and V. Woloszyn, "How vulnerable are automatic fake news detection methods to adversarial attacks?," 2021.

[4] H. Ali, M. S. Khan, A. Alghadhban, M. Alazmi, A. Alzamil, K. Al-Utaibi, and J. Qadir, "All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings," *IEEE Access*, vol. 9, pp. 81678–81692, 2021.

[5] J. 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 *ACL EMNLP*, pp. 119–126, 2020.

[6] T. Le, S. Wang, and D. Lee, "MALCOM: Generating malicious comments to attack neural fake news detection models," in *IEEE ICDM*, pp. 282–291, 2020.

[7] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media," *arXiv preprint arXiv:1809.01286*, 2018.

[8] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *ACM SIGKDD*, KDD 2019, pp. 395–405, 2019.

[9] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[10] K. Pelrine, J. Danovitch, and R. Rabbany, "The surprising performance of simple baselines for misinformation detection," in *The Web Conference 2021*, p. 3432–3441, 2021.