How Effectively Can Machines Defend Against Machine-Generated Fake News? An Empirical Study

Meghana Moorthy Bhat
The Ohio State University
bhat.89@osu.edu

Srinivasan Parthasarathy
The Ohio State University
srini@cse.ohio-state.edu

Abstract

We empirically study the effectiveness of machine-generated fake news detectors by understanding the model's sensitivity to different synthetic perturbations during test time. The current machine-generated fake news detectors rely on provenance to determine the veracity of news. Our experiments find that the success of these detectors can be limited since they are rarely sensitive to semantic perturbations and are very sensitive to syntactic perturbations. Also, we would like to open-source our code and believe it could be a useful diagnostic tool for evaluating models aimed at fighting machine-generated fake news.

1 Introduction

The advancement of language models (LM) in text generation has raised concerns over misusing LM in generating fake news, misleading reviews, spreading rumor and propaganda (Vosoughi et al., 2018; Solaiman et al., 2019; Varshney et al., 2019, 2020). Fact-checking is one approach that involves studying veracity of the news using external evidence (Popat et al., 2018; Nie et al., 2018). However, it remains a challenging task since the performance of the current automatic fact-checking models are not satisfactory (Thorne et al., 2018). Rashkin et al. (2017) studied automated fact-checking by examining the role of stylistic bias to help verify the truthfulness of an article. Another approach which has recently gained traction to combat mass-scale production of fake news is detecting stylistic differences in human-written and machine-generated news[1](Radford et al., 2019). Later, Grover (Zellers et al., 2019), a transformer based model (Vaswani et al., 2017) trained on news corpora was proposed to determine machine-generated fake news.

The detection of machine-generated fake news purely based on stylistic biases can be hard because: (1) legitimate human-written articles can be easily corrupted at scale by machines, (2) an attacker can overlay the distributional features of human-written text over machine-generated text to fool the style-based classifiers and vice-versa, (3) legitimate text can be generated with LM and the current machine-generated fake news detectors rely on similar distribution for generation of legitimate and fake news (Schuster et al., 2019). However, to the best of our knowledge, there has not been a systematic empirical evaluation to validate these claims. We devise six different perturbations to study the behavior of models[2]. In this study we do not cover (3) since generative models for applications like summarization (See et al., 2017; Nallapati et al., 2016), text completion (Vaswani et al., 2017; Radford et al., 2019) can be directly tested for veracity on detector models. From our experiments, we see that models are insensitive to semantic types of perturbations considered in this work, and moreover are extremely sensitive to grammatical perturbations, that do not change the semantics.

2 Related Work

Universal Attacks in NLP: Ribeiro et al. (2018) debugged models using semantic-preserving perturbations that forced changes in predictions for downstream tasks such as sentiment analysis, visual QA and machine comprehension. Behjati et al. (2019) crafted data-independent adversarial sequences that can fool text classifier when added to any input sample. Alternatively, Wallace et al. (2019) study triggers in the form of a word or a few words to analyze models and biases in datasets for LM, text

[1]https://openai.com/blog/gpt-2-1-5b-release/

[2]For easy convention, we refer machine-generated fake news detectors as models in this work

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 48–53

ISBN 978-1-952148-66-8

classification.

Machine-generated text detection: Bakhtin et al. (2019) study the generalization ability of models trained to detect real text from the machine-generated text, Gehrmann et al. (2019) show statistical distributional differences between human-written and machine-generated text and provide a tool to the readers to detect machine generated text. Zellers et al. (2019) proposed defense against machine-generated fake news, Grover, by building a linear classifier on top of the last hidden state of its controlled generator model trained on a large news corpora.

Fake news detection: Shu et al. (2017) study fake news detectors in social media from a data mining perspective. Similarly, Zhou et al. (2019) study vulnerabilities of the Fakebox tool, an open-source fake news detector and emphasized the necessity of crowd-source based knowledge graphs and fact checking based solutions to combat fake news. Automatic fact checking is another approach that is being studied actively with synthetic (Thorne et al., 2018) and real datasets (Wang, 2017; Augenstein et al., 2019). In parallel work, Schuster et al. (2019) discusses the limitations of stylistic based approaches for machine-generated fake news detection. They devised two benchmarks: (1) text completion using LM and, (2) negating the meaning of human-written articles by maintaining the distribution of human-written text as learned by the model.

Our work is orthogonal to these efforts in the following ways: (1) we study (in)sensitivity of multiple models for semantic perturbations by keeping the distribution of human-written text intact, (2) we also study the sensitivity of models to semantic-preserving syntactic perturbations with the goal of overlaying distribution of human-written text over machine-generated text, (3) our experiments can be used as a diagnostic evaluation tool for future machine-generated fake news detectors and, (4) our semantic perturbations can be used to evaluate and study fact-checking based solutions as well.

3 Methodology

We measure the performance of models by looking at accuracy with respect to perturbations introduced in this work. All models are trained without any perturbations and the behavior is studied only at test time. We consider the following models in our experiments - Grover Mega discriminator[3], GPT2 output detector[4] and Fakebox[5] and use RealNews dataset[6].

3.1 Types of Perturbations

We devise perturbations across real news (human-written) articles to test the model behavior. The perturbations in this work are broadly categorized in two main streams: semantic and syntactic. The semantic perturbations are sentence-level perturbations while the syntactic are word-level perturbations. At $N\%$ perturbation level, for any type of semantic perturbation, semantics of $N\%$ of sentences are changed; whereas in a syntactic perturbation, $N\%$ of the words are modified.

3.1.1 Semantic perturbations

The semantic perturbations are intended to turn a real news (human-written text) into a fake news. Our aim is to understand to what extent the content and factuality of text influences model decisions. The psychology studies show that people try to diverge as little as possible from the truth while lying (Mazar et al., 2008). Hence, we study perturbations at various levels to understand the sensitivity of models. Understandably, models find difficulty in spotting minor perturbations (except article shuffling) and the performance improves as we add more noise to the real articles. An ideal model will flip its decision to all the semantic perturbations introduced in this work. Below is a brief description of types of semantic perturbations we consider.

varying sentiment: We change the polarity of sentences within an article, by replacing positive, or comforting words to negative words and vice-versa; thereby changing the overall sentiment of the article. In order to reverse the polarity of sentences, we replace one randomly chosen word in a sentence with its antonym from Stanford NLTK[7].

source-target exchange: The source and the target entities[8] in a sentence are interchanged that do not have coordinating or correlating conjunctions.

article shuffling: We perturb a real article by randomly adding $N\%$ of sentences from a fake article where N is the perturbation level. We also

[3] https://grover.allenai.org/detect
[4] https://huggingface.co/openai-detector
[5] https://machinebox.io/docs/fakebox
[6] https://rowanzellers.com/grover/
[7] https://www.nltk.org/

Perturbation type	Real article	Fake article
varying sentiment	— Google said **last** year it spent more than $100 million on Content ID. They say the **automatic** filters are blunt — **Some** consumers worry that the new rules would bring an end. The EU **denies** this.	— Google said **first** year it spent more than $100 million on Content ID. They say the **manual** filters are blunt — **No** consumers worry that the new rules would bring an end. The EU **admit** this.
source-target exchange	— **Lokuhettige** had 14 days to respond to the new charges, the **ICC** added. **Sri Lanka Cricket** has been thrown into turmoil as the **ICC** continues to investigate corruption allegations in the island nation. —	— **ICC** had 14 days to respond to the new charges, the **Lokuhettige** added. **ICC** has been thrown into turmoil as the **Sri Lanka Cricket** continues to investigate corruption allegations in the island nation. —
article shuffling	— Rose feels not enough action is being taken and the disparity in the punishment highlights its ineffectiveness. **"Obviously, it is a bit sad (to feel like this) but when countries only get fined what I'd probably spend on a night out in London, what do you expect" he added.** "You see my manager get banned for two games for just being —"	— Rose feels not enough action is being taken and the disparity in the punishment highlights its ineffectiveness. **This would pave the way for Daenerys Targaryen to bring the Wall down.**"You see my manager get banned for two games for just being —"
entity replacement	**A New Jersey** bus driver's incredible note to the parents of two children who reached out to another student with a disability went viral.—	**Pribbernow, New Kilgore** bus driver's incredible note to the parents of two children who reached out to another student with a parents went viral.—
altering numerical facts	**Mueller Report: 10 Instances of Possible Obstruction** of Justice by Trump Special counsel Robert Mueller's report on Russian interference in the 2016 presidential election included instances of potential obstruction by President Donald Trump.	**Mueller Report: 67497 Instances of Possible Obstruction** of Justice by Trump Special counsel Robert Mueller's report on Russian interference in the 2016 presidential election included instances of potential obstruction by President Donald Trump.
syntactic perturbation	**There is** no way to fully understand **what is** going on in crypto **world. I am** not even sure anyone could even if you tried to. I can tell you that recent surge in BitCoin is an opportunity to buy **long term** real assets	**There's** no way to fully understand **what's** going on in **the** crypto **world - I am** not sure anyone could even if you tried to. I can tell you that **the** recent surge in BitCoin is an opportunity to buy **long-term** real assets.

Table 1: Examples of excerpts from articles subjected to perturbations. Humans in general find difficulty in detecting fake news articles without scrutiny.

remove an equal number of sentences from the real article to maintain the total article length. The fake article chosen for shuffling will not have entities present in the title of real article.

entity replacement: We replace entities[8] with another irrelevant entity of the same type. The irrelevant entities are picked from the fake articles which are not present in real articles.

alter numerical facts: The numerical facts are distorted in a real news article. A numerical figure (digits and words) less than hundred thousand will be scaled up to a random number in the range of 1 million to 1 trillion and vice-versa.

Table 1 represents examples of articles subjected to perturbations in this work. We use spacy[8] to identify entities and entity types in our experiments.

3.1.2 Syntactic perturbations

Ippolito et al. (2019) recently studied the influence of excerpt length for classification of machine-generated and human-written text. In the training dataset of Grover, we observe that machine-generated articles have shorter length but longer sentences than human-written articles. We perturb these features by: (i) breaking longer sentences, (ii) removing definite articles if they appear among the most repeated words in an article, (iii) using semantic-preserving rules (for example converting *that's →that is*) (Ribeiro et al., 2018), (iv) reformatting paragraphs of machine-generated text. These perturbations preserve semantics of articles, hence an ideal model should *not flip* its decision.

[8]https://spacy.io/

4 Results and Analysis

Table 2 summarizes the performance of models when subjected to different types of perturbations at test time. For our experiments, we pick 2K samples for every perturbation type (real for semantic and fake for syntactic perturbations) from the RealNews dataset that are classified correctly (100% accuracy) by all the models without any perturbations introduced in this work. We start with 25% perturbation level because very small perturbation levels may not be enough to change the overall semantics of the article. However, Grover identifies its own generated text even at 1% perturbation level (18% accuracy on article shuffling perturbation). On manual examination, we found that on an average 5% of the real articles did not change semantics on perturbing for varying sentiment and source-target exchange. Fakebox performance is not reported due to very low accuracy for all the perturbations introduced in this work. Since the details of Fakebox tool is not publicly available,

Perturbation levels (%)	Models	Semantic perturbations (accuracy (%))					Syntactic perturbations (accuracy (%))
		varying sentiment	source-target exchange	article shuffling	entity replacement	alter numerical facts	machine to human
25	Grover	0	$3.8_{(\pm 0.28)}$	$45.2_{(\pm 1.88)}$	$1.98_{(\pm 0.45)}$	$0.2_{(\pm 0.0)}$	$23.07_{(\pm 2.21)}$
	GPT2	$3.83_{(\pm 0.52)}$	0	$15.8_{(\pm 0.42)}$	$8.68_{(\pm 0.29)}$	$0.1_{(\pm 0.0)}$	$17.7_{(\pm 1.27)}$
50	Grover	0	$9.4_{(\pm 0.73)}$	$72.66_{(\pm 1.98)}$	$3.9_{(\pm 0.32)}$	$0.2_{(\pm 0.01)}$	$12.83_{(\pm 0.55)}$
	GPT2	$7.37_{(\pm 0.56)}$	0	$35.3_{(\pm 0.42)}$	$26.9_{(\pm 0.79)}$	$0.1_{(\pm 0.01)}$	$13.83_{(\pm 0.6)}$
75	Grover	0	$11.44_{(\pm 0.29)}$	$83.14_{(\pm 2.03)}$	$6.11_{(\pm 0.39)}$	$0.23_{(\pm 0.08)}$	$10.07_{(\pm 0.68)}$
	GPT2	$8.03_{(\pm 0.67)}$	0	$40.36_{(\pm 0.61)}$	$48.53_{(\pm 0.54)}$	$0.13_{(\pm 0.02)}$	$3.83_{(\pm 0.49)}$
100	Grover	0	$18.39_{(\pm 0.61)}$	NA	$8.68_{(\pm 0.19)}$	$0.28_{(\pm 0.06)}$	$6_{(\pm 1.45)}$
	GPT2	$6.11_{(\pm 0.21)}$	0	NA	$67.82_{(\pm 0.42)}$	$0.13_{(\pm 0.02)}$	$1.1_{(\pm 0.45)}$

Table 2: Performance of detectors for perturbations measured in accuracy. The cell values contain the mean and standard deviation across 5 runs of experiments. We choose real articles for devising semantic perturbations and fake articles for devising syntactic perturbations. For semantic perturbations, we see that model performance increases with level of perturbations while for syntactic perturbation, models tend to perform bad with increase in attributes of human-written text. We mark 'NA' in article shuffling for 100% perturbation level since 100% shuffling will be a full machine-generated text which was already classified correctly by the detectors.

we omit them from analysis. The code is publicly available[9]. From our experiments, we make the following observations:

- All the machine-generated fake news detectors considered in this work are vulnerable to semantic perturbations even under extreme perturbations indicating that actuality of the articles do not aid in model decision.
- Grover performs well on article shuffling indicating that it learns sentence structures of its own generated text pretty well.
- The machine-generated fake news detectors are also vulnerable to semantic preserving syntactic perturbations indicating they could possibly be learning sentence structures. Another reason could be due to data bias since the training dataset of machine-generated text has longer sentences, punctuation and definite articles when compared to human-written text.
- Grover fails to detect sentiment changes in articles indicating that it is insensitive to polarity between entities. From manual examination we found that 5% of real articles perturbed due to varying sentiment have uncommon phrases which would have aided the GPT2 detector. For example, *Police say Aranda told them he would **go** to the mall →Police say Aranda told them he **stay_in_place** to the mall*
- Current machine-generated fake news detectors rely on previously seen data without external resources for classification. This could possibly explain the performance drop of GPT2 in varying sentiment at 100% perturbation level since there will be no inconsistent polarity towards entities unlike perturbations at 50% or 75%.
- Transformers are insensitive to perturbations like word-level shuffling and possibly learn bag-of-word like distribution (Sankar et al., 2019). GPT2 fails to identify source-target exchange indicating they adhere to the above observations. The marginal gains of Grover probably indicates better understanding of linguistic cues in sentences.
- The better performance of GPT2 in entity replacement could be due to non-existence of replaced entities in articles labeled real in the training dataset of GPT2.

5 Conclusion

With the advances in language modeling for text generation, the detection of fake news becomes challenging. We find that success of style-based classifiers are limited when real articles are perturbed even under extreme modifications. We believe our experiments motivates to explore integration of multiple dimensions like examine source credibility, fact-checking via external resources, model robustness by adversarial training, commonsense reasoning to machine-generated fake news detectors. By open-sourcing our code, we believe our methodology of studying vulnerabilities in the

[9]https://github.com/meghu2791/evaluateNeuralFakenewsDetectors

fake news detectors can aid in creation of robust models in the future.

Acknowledgments

We thank our reviewers for their feedback and suggestions. This work is supported by the National Science Foundation grant EAR-1520870. All content presented represents the opinion of the authors, and is not necessarily endorsed by their sponsors.

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *CoRR*, abs/1906.03351.

M. Behjati, S. Moosavi-Dezfooli, M. S. Baghshah, and P. Frossard. 2019. Universal adversarial attacks on text classifiers. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7345–7349.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled.

Nina Mazar, On Amir, and Dan Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6):633–644.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining fact extraction and verification with neural semantic matching networks. *CoRR*, abs/1811.07039.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *CoRR*, abs/1809.06416.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. 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*, pages 2931–2937. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher J. Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *CoRR*, abs/1906.01603.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2019. Are we safe yet? the limitations of distributional features for fake news detection. *arXiv preprint arXiv:1908.09805*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher. 2019. Pretrained ai models: Performativity, mobility, and change.

Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher. 2020. Limits of detecting text generated by large-scale language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Empirical Methods in Natural Language Processing*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *ArXiv*, abs/1705.00648.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9051–9062.

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via NLP is vulnerable to adversarial attacks. *CoRR*, abs/1901.09657.