Robustness Analysis of Grover for Machine-Generated News Detection

Rinaldo Gagiano[1]
S3870806@student.rmit.edu.au
Xiuzhen Zhang[1]
xiuzhen.zhang@rmit.edu.au

Maria Myung-Hee Kim[2]
maria.kim@dst.defence.gov.au
Jennifer Biggs[2]
jennifer.biggs@dst.defence.gov.au

[1] School of Computing Technologies, RMIT University, Australia
[2] Defence Science and Technology Group, Australia

Abstract

Advancements in Natural Language Generation have raised concerns on its potential misuse for deep fake news. Grover is a model for both generation and detection of neural fake news. While its performance on automatically discriminating neural fake news surpassed GPT-2 and BERT, Grover could face a variety of adversarial attacks to deceive detection. In this work, we present an investigation of Grover's susceptibility to adversarial attacks such as character-level and word-level perturbations. The experiment results show that even a singular character alteration can cause Grover to fail, affecting up to 97% of target articles with unlimited attack attempts, exposing a lack of robustness. We further analyse these misclassified cases to highlight affected words, identify vulnerability within Grover's encoder, and perform a novel visualisation of cumulative classification scores to assist in interpreting model behaviour.

1 Introduction

Online disinformation has become a crucial issue in current society and has been the focus of extensive study in recent years (Buning, 2018; Fletcher, 2018; Zerback, 2020). Fake news, one form of online disinformation, can deceive people with intent of monetary gain, political slander, or entity discreditation (Quandt et al., 2019). While current sources of fake news are mainly derived from human hand, recent developments in Natural Language Generation (NLG) (Radford, 2018, 2019; Brown, 2020) have made it possible to produce neural fake news [1] at scale. The key problem with this technology is that it is harder for humans to distinguish machine-generated text from human-produced text (Heaven, 2020; Hao, 2020).

To counter the rising threat of neural fake news, an automatic discriminator has been developed that can serve as a defence mechanism. In 2019, Grover (Zellers et al., 2019) (**G**enerating a**R**ticles by **O**nly **V**iewing m**E**tadata **R**ecords), a neural fake news generator and discriminator, was released to the public. As a generator, it generates formal news articles, (including title, domain, authors, date) with given contextual metadata. As a discriminator, it detects the difference between machine and human-produced articles. By utilising articles produced by the generator, Grover's discriminator achieved 92% accuracy while detectors based deep contextual language models including GPT-2 and BERT achieved 73% (Zellers et al., 2019).

Grover can be misused to mass produce plausible disinformation by adversaries. For example, Grover generated propaganda articles were rated as more trustworthy than human-produced ones of the same context by human judges (Zellers et al., 2019). Given this alarming ability, the capability to auto-detect the differences between machine and human-produced articles can reduce the risk of neural fake news spreading online.

Following the establishment of text-based perturbations by Jia and Liang (2017), studies on robustness interpretability through adversarial examples have grown rapidly through the Natural Language Processing (NLP) community (Vadillo, 2021; Zafar, 2021; Yuan, 2021). Since then, there have been several attempts to manipulate NLP models by character-level alterations on its input text. For example, Belinkov and Bisk (2017) demonstrated that synthetic and natural noise can cause state-of-the-art language translation models

[1] From here on out, we will use 'neural fake news' and 'machine-generated fake news' interchangeably.

to fail. Gao (2018) also proposed DeepWord-Bug, a novel algorithm for small character perturbations causing drastic classification inaccuracies in tasks such as text classification, sentiment analysis, and spam detection. These studies conducted character-level perturbations to identify a lack of robustness within various mainstream language models.

In a similar manner, Grover, when acting as a defence mechanism against neural fake news, can face heavy adversarial scrutiny. Thus, following the direction of recent studies (Belinkov and Bisk, 2017; Gao, 2018), we conducted analyses through various adversarial attacks including character-level and token-level perturbations.

This paper presents an investigation of Grover to examine its performance change on various adversarial attacks. In our assessment, we find that Grover is highly susceptible to adversarial attacks with around 93% of target articles vulnerable to misclassification after alteration. Analysing the effects of successful perturbations, we identify a weakness within the model's encoding framework which influences Grover's classification scoring, with recorded score variations of 0.74 on average. In this work, we introduce our novel visualisation of cumulative classification score on various unaltered/altered articles and explore classification score polarity induced by adversarial attacks.

This paper is organised as follows. Section 2 accounts related work and section 3 reports a general summary of Grover. Section 4 presents the experiments of adversarial attacks. Section 5 conveys the results of the experiments along with error analysis. Section 6 presents cumulative classification score visualisation and analysis on extreme polarity change. Finally, section 7 presents our concluding discussion.

2 Related Work

Recent studies on adversarial attacks in NLP follow a white-box approach leveraging accessible information from within a model as surveyed by Zhang (2020). Many studies have utilised a white-box gradient-based approach for various attacks such as character-based alterations (Ebrahimi, 2017, 2018; Liang, 2017), word-based alterations (Cheng, 2020; Liang, 2017; Neekhara, 2018), and word-based concatenations (Wallace, 2019; Behjati, 2019). Blohm (2018) used white-box model attention to attack a reading comprehension model as well as a question answering model.

Contrary to the white-box approach, Wolff and Wolff (2020) adopted a black-box approach and performed homoglyph and misspelling attacks on a variety of neural text classifiers including GPT-2, GLTR, RoBERTa, and Grover. They conducted adversarial attacks on 20 samples of *Machine* articles to draw comparison between leading neural classifiers and Grover yet refrain from exploring the results of Grover's classification in detail. Our work includes the attack concepts from Wolff and Wolff's work (2020) but explore singular applications of the attacks, rather than multiple applications. We also focus our analysis solely on Grover, studying the effect of the attacks produced on Grover, and its potential fragile points within the framework.

Visualising a language model's outcome to increase a model's interpretability is another recent trend in NLP. Gehrmann (2019) introduced GLTR, a visualisation tool (using statistical methods) that can detect generation artifacts across a sample and display its findings through coloured annotation on the input to support a human's fake text detection. Stemming from this concept, we propose a novel visualisation approach through the plotting of cumulative classification scores. Our visualisation method aims to help a user to interpret how Grover is affected at each word vector and highlight key alteration artifacts within an article.

3 Grover

Grover consists of two components: a generator and a discriminator.

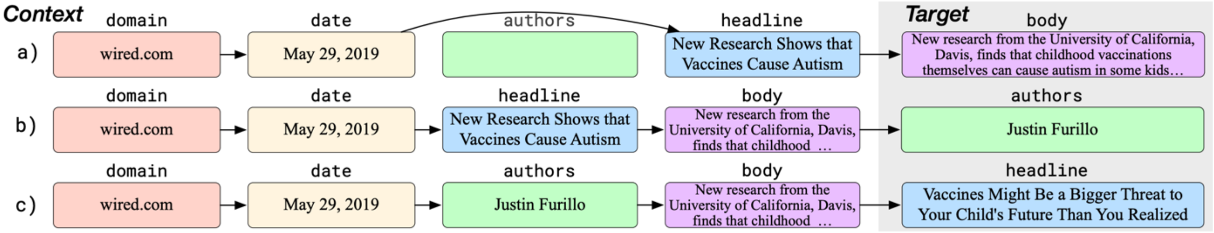
Context
Target
a)
domain
wired.com
date
May 29, 2019
authors
headline
New Research Shows that Vaccines Cause Autism
body
New research from the University of California, Davis, finds that childhood vaccinations themselves can cause autism in some kids…
b)
domain
wired.com
date
May 29, 2019
headline
New Research Shows that Vaccines Cause Autism
body
New research from the University of California, Davis, finds that childhood …
authors
Justin Furillo
c)
domain
wired.com
date
May 29, 2019
authors
Justin Furillo
body
New research from the University of California, Davis, finds that childhood …
headline
Vaccines Might Be a Bigger Threat to Your Child's Future Than You Realized

Figure 1: A diagram of Grover examples for article generation. Note ~ Fig 2 from 'Defending Against Neural Fake News' by Zellers et al., 2019.

3.1 Generator

The generator component of Grover comprises a novel architecture with adapted components of GPT-2. Grover, as shown in Figure 1, can generate the domain, date, headline, body, or author of a news article, given any subsetted combination of these fields. The generator comes in three versions – Grover-Base, consisting of 12 layers and 124 million parameters, Grover-Large, consisting of 24 layers and 355 million parameters, and Grover-Mega, with 48 layers and 1.5 billion parameters matching GPT-2's architecture; each trained on successively larger datasets (comprised of real news articles scraped from common crawl[2]).

3.2 Discriminator

The discriminator component of Grover acts as a detector of neurally generated articles. Utilising articles produced by the generator, the discriminator is trained to differentiate between machine-generated articles and human-produced articles. Articles can be classified on their own or with additional metadata such as domain, date, headline, and author, that aids prediction strength.

4 Experiments

The functionality of Grover's discriminator, given either machine-generated articles (labelled as *Machine*) or human-produced articles (labelled as *Human*), is to produce a classification label of '**Human**' or '**Machine**' on each article. Input articles contain the body of an article, with or without metadata (title, domain, date, or authors).

To assess Grover's robustness, we conducted experiments on the discriminator's classification accuracy when classifying altered (adversarial attacked) *Machine* articles. Minor alterations (altering only one character or one word in a whole news article) have been performed on a subset of *Machine* articles applying four methods of adversarial attacks including (1) upper/lower flip, (2) homoglyph, (3) whitespace, and (4) misspelling. After each attack, the altered articles were submitted to Grover's discriminator for reclassification and the classification results were investigated.

4.1 Discriminator Setup

For experiments, the publicly available pre-trained Grover Mega discriminator was used; the set-up contains Grover Mega config file and necessary checkpoints[3]. We ran the discriminator in its GPU configuration.

4.2 Dataset

Grover provides a dataset containing 12,000 articles with metadata[4]; it consists of 8,000 *Human* articles (RealNews dataset[5]), and 4,000 *Machine* articles, which were generated using Grover's generator (Grover-Mega). Submitting this dataset to Grover's discriminator, we gain the predictions seen in Table 1. From the prediction we obtain a total accuracy of 0.93, a precision score of 0.85, a recall score of 0.94, and a F1 score of 0.89.

		True Class	
		Machine n=4000	*Human* n=8000
Predicted Class	**Machine**	TP (3,751)	FP (649)
	Human	FN (249)	TN (7,351)

Table 1: Confusion Matrix of 12,000 articles classified by Grover Mega discriminator. True Positives (TP). False Positives (FP). False Negatives (FN). True Negatives (TN).

For our experiments, we sampled 100 articles with the highest true positive (TP) classification scores produced by the discriminator. This will be referenced as 100 *Machine* article subset. All articles selected have classification score over 0.49 where 0.5 is the maximum score an article could be assigned for a '**Machine**' classification.

[2] https://commoncrawl.org/

[3] https://github.com/rowanz/grover/tree/master/discrimination

[4] gs://grover-models/discrimination/generator=medium~discriminator=grover~discsize=medium~dataset=p=0.96/checkpoint

[5] https://github.com/rowanz/grover/tree/master/realnews

Original	“A Romanian hospital will face a fine for leaving a towel in a patient's stomach…”
Whitespace	“A Romanian hospital **willface** a fine for leaving a towel in a patient's stomach…”
Upper/Lower Flip	“A Romanian hospital will face a fine for leav**I**ng a towel in a patient's stomach…”
Misspelling	“A Romanian hospital will face a fine for leaving a towel in a patient's **stomache**…”
Homoglyph	“A Romanian hospital will face a fin**e*** for leaving a towel in a patient's stomach…”

Table 2: Adversarial attacks and their respective change on an article. *The word ‘Fine’ in the homoglyph example contains Cyrillic ‘e’ ~ Unicode: U+x0435 compared to the regular Latin ‘e’ ~ Unicode: U+0065.

Attack	**Alterations**	**Misclassifications (Proportion)**	**Affected Articles**
U/L Flip	212,224	4,295 (2.02%)	96%
Homoglyph	157,532	6,914 (4.39%)	97%
Whitespace	46,036	1,447 (3.14%)	85%
Misspelling	43,789	4,281 (9.78%)	94%

Table 3: Classification results of all adversarial examples. **Alterations** indicate how many iterations of the specified attack was conducted across the dataset. **Affected Articles** indicate how many articles, from the 100 *Machine target* articles, had one or more misclassifications resulting from an alteration.

4.3 Adversarial Attack Parameters

As news articles are written to a high level of coherency with minimal punctual mistakes or grammatical errors, an adversary would want to limit article alteration to preserve readability and ensure a human reader does not question the article’s credibility. To simulate this mindset, we limit the application of an attack to only a single change, such as one character or one-word alteration on an article, iterating the attack through the entirety of an article to assess all possible combinations for each attack’s relative application. As demonstrated in Table 2, the following four types of adversarial attacks were applied for the experiments:

(1) **Upper/Lower Flip:** Uppercasing or lowercasing of a letter originally lowercased or uppercased respectively.

(2) **Homoglyph:** Replacement of certain characters with their homoglyph equivalent from either the Greek or Cyrillic alphabet[6].

(3) **Whitespace:** Removal of a space between adjacent words.

(4) **Misspelling:** Replacement of certain words coinciding with a list of commonly misspelled English words on Wikipedia[7].

4.4 Adversarial Attack Results

We present the results from our adversarial attack experiments on Grover.

As shown in Table 3, character-level attacks (U/L Flip and Homoglyph) create a higher number of altered articles compared to word-level attacks (Whitespace and Misspelling). Based on the number of alterations, the Misspelling attack achieved the highest misclassification rates (nearly 10%) compared to the other three attacks which got a relatively lower rate of 2-4%.

Surprisingly, across the 100 *Machine* article subset, Homoglyph, U/L Flip and Misspelling attacks affected 97%, 96% and 94% of the target articles, respectively. Even the simplest attack, Whitespace attack, could affect 85% of the 100 target *Machine* articles. This suggests that Grover is highly susceptible to adversarial efforts.

Table 4 shows the ten most common words that affected (flipped the classification from ‘**Machine**’ to ‘**Human**’) Grover’s discriminator during adversarial attacks. Around 20% of

[6] We use 19 different Greek substitutions and 30 different Cyrillic substitutions. All substitutions can be found in the appendix.

[7] https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines

Affected Word	Frequency	Proportion	POS
that	1639	8.92%	IN
the	1533	8.34%	DT
to	516	2.81%	TO
and	334	1.82%	CC
with	321	1.75%	IN
in	298	1.62%	IN
of	279	1.52%	IN
for	257	1.40%	IN
from	236	1.28%	IN
The	202	1.10%	DT

Table 4: Statistics of affected words from all misclassified inputs. **POS** is the part-of-speech tag for that respective word obtained from NLTK[5]. **IN** ~ Preposition, **DT** ~ Determiner, **TO** ~ To, **CC** ~ Coordinating Conjunction. Note we only take the top 10 most occurring words within the misclassified subset.

misclassifications were caused by altering the words ‘that’, ‘the’ and ‘to’. Noticeably, the majority of the affected words are stop words.

4.5 Input Encoding

We observed in general which words were altered to elicit a misclassification. To assess how character-level perturbations affect Grover, we examined how the model interprets and scores a given input.

Grover uses a byte-pair encoder (BPE) to pre-process input data. BPE (Senrich et al., 2015) splits a given input into its largest subword units based on character co-occurrence frequency distribution and assigns each unit a pre-determined pairing ID. This turns a tokenised input into a vector of numbers.

Previously, BPEs have been found to be lacking in robustness when facing character-level perturbations (Heigold et al., 2017). In Table 5 we can see the effect that the upper/lower flip attack has on a particular sequence from one of the articles. The uppercasing of the letter ‘i’ in ‘hospital’ changes the subword unit allocation. Originally encoded as [4437], ‘hosp**I**tal’ gets broken down into ‘hosp’,’It’,’al’ then encoded into [10497, 1027, 283].

Original	Vector IDs		Altered
A	33		A
Romanian	34345		Romanian
hospital	**4437**	**10497**	**hosp**
		1027	**It**
		283	**al**
will	482		will
face	1987		face
a	258		a
fine	3735		fine
for	330		for

Table 5: An original encoding sequence compared to the same encoded sequence after a single character alteration.

5 Visual Analysis

Grover produces a classification score at each word vector, as it processes the input from left to right. If we successively and cumulatively feed Grover word vectors in sequential order, we can obtain a classification score at each step, allowing for a cumulative classification score to be recorded. Using the classification scores recorded at each increment as word vectors are appended to the accumulating input, we can visualise how these are perceived by Grover over the course of an entire input.

5.1 Cumulative Classification Score Visualisation

***Human* Articles:** Figure 2 illustrates the cumulative classification score of five randomly selected *Human* articles from the original 8,000 *Human* article dataset. At the initial processing of the sequence, all articles start at a strong ‘**Machine**’ classification. As more of the respective input is processed, we see the articles' classification scores increase toward ‘**Human**’ over time. It is observed that cumulative classification scores often plateau with greater encoded sequence lengths.

[5] https://www.nltk.org/

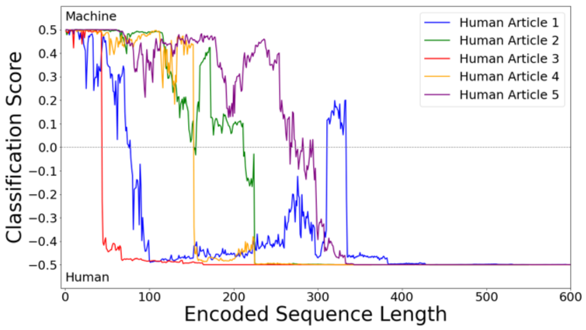

Figure 2: Comparison of cumulative classification scores between five *Human* articles.

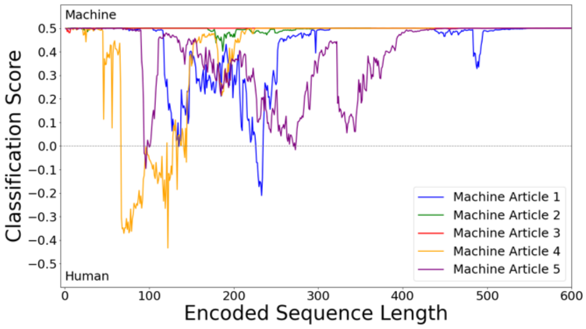

Figure 3: Comparison of cumulative classification scores between five *Machine* articles.

***Machine* Articles:** Figure 3 shows the cumulative classification scores of five randomly selected *Machine* articles from our target dataset. As seen in the visualisation of *Human* articles, the beginning of each sequence starts at a strong '**Machine**' classification. Over the early stages of the sequence, we see high classification score variance due to the limited word vectors processed. Over time, the selected *Machine* articles tend to return to a strong '**Machine**' classification, plateauing toward the end of the encoded sequence.

False Negative (FN) Case: Figure 4 presents the cumulative classification score of one of the misclassified articles from our experiments. The red line indicates the location of the adversarial attack within the encoded sequence. In this example, the input word 'that' was transformed into 'thaT' by U/L Flip attack which uppercased the second 't'. At the point where Grover processed the altered word vector, the classification score of the article dropped dramatically, falling a total of 0.98. This large variation in classification score due to alteration will be discussed in terms of 'Extreme Polarity Change' in section 5.2.

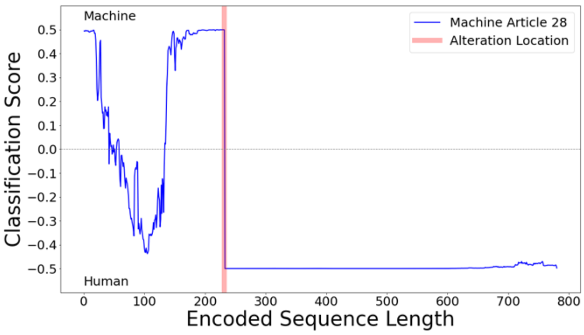

Figure 4: Cumulative classification scores of misclassified altered *Machine* article after the U/L Flip attack.

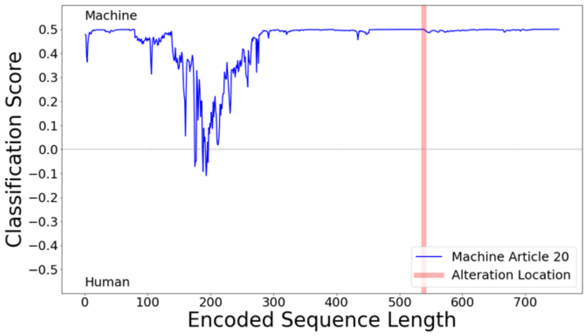

Figure 5: Cumulative classification scores of correctly classified altered *Machine* article after the U/L Flip attack.

True Positive (TP) Case: Figure 5 demonstrates the cumulative classification score of a *Machine* article that had its classification unaffected after an adversarial attack. Again, the red line indicates the location of the attack. In this example, the input word, 'These' was altered to 'these' by the U/L Flip attack which lowercased the first 'T'. This alteration causes a very minimal change in classification score at the site of alteration.

5.2 Extreme Polarity Change

From visualising a FN case's cumulative classification scores, we observed a large change in classification score at the point of an adversarial attack. To analyse whether all FN cases show a drastic variation in classification score, we took a random sample of 500 FN case articles and 500 TP case articles from each of the four adversarial attacks. In total, we examined the 4,000 articles' classification score at each point of the adversarial attack. The average score variation of each subset is shown in Table 6.

	Average Score Variation	
Attack	TP Subset	FN Subset
U/L Flip	0.12	0.76
Homoglyph	0.17	0.81
Whitespace	0.04	0.70
Misspelling	0.21	0.69
Average	0.14	0.74

Table 6: Average classification score variation at the point of an attack within an input.

The FN cases had a much higher average variation in classification score compared to the TP cases as shown in Table 7. This implies that particular alterations caused Grover's classification score to drop dramatically (at the site of an attack) ultimately affecting the final prediction produced by Grover.

6 Discussion

In this study, the robustness of Grover's discriminator was assessed through various adversarial attacks. We found that even a singular character change can cause the model to fail. Through analyses of successful perturbations, it was found that Grover's encoder is highly sensitive to selected perturbations, causing downstream effects in classification assignment.

We conducted a broad implementation of adversarial attacks and identified vulnerabilities in single alterations on certain types of words. These results outline potential dependencies within Grover's language modelling which could be potentially extorted by adversaries through implementation of multiple instances of an adversarial attack across an article or an adversary targeting and affecting more than one key word outlined in Table 4.

To the best of our knowledge, the proposed visualisation of cumulative classification scores are novel, allowing interpretation of model behaviour, as it gives a user the ability to visually understand the effects that each word vector has at its relative point of inference as well as the effects that alterations may produce on the classification prediction.

Our findings open various paths for further exploration. Our adversarial attacks' focus was exclusively directed onto the body of an article. One path for future work could consist of focussing adversarial attacks on the metadata of an article, further exploring Grover's robustness. Our visualisation of cumulative classification scores highlighted the effects some character-level alterations had on the classification score of an article. The large score variations noted could allow for work to be done in the field of adversarial attack detection. Finally, the nature of our assessment was broad and based on a black-box approach. Furthering our work, the undertaking of a white-box approach could be performed to explore model interpretability.

Acknowledgments

R. Gagiano is supported by Defence Science and Technology Group Graduate Industry Placement. X. Zhang is supported by the ARC Discovery Project DP200101441.

References

Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. "Universal adversarial attacks on text classifiers." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7345-7349. IEEE, 2019.

Yonatan Belinkov, and Yonatan Bisk. "Synthetic and natural noise both break neural machine translation." arXiv preprint arXiv:1711.02173 (2017).

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. "Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension." arXiv preprint arXiv:1808.08744 (2018).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

Madeleine de Cock Buning. "A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation." (2018).

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. "Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 3601-3608. 2020.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. "Hotflip: White-box adversarial examples for text classification." arXiv preprint arXiv:1712.06751 (2017).

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. "On adversarial examples for character-level neural machine translation." arXiv preprint arXiv:1806.09030 (2018).

Richard Fletcher, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. "Measuring the reach of" fake news" and online disinformation in Europe." Australasian Policing 10, no. 2 (2018).

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. "Black-box generation of adversarial text sequences to evade deep learning classifiers." In 2018 IEEE Security and Privacy Workshops (SPW), pp. 50-56. IEEE, 2018.

Sebastian Gehrmann , Hendrik Strobelt, and Alexander M. Rush. "GLTR: Statistical detection and visualization of generated text." arXiv preprint arXiv:1906.04043 (2019).

Will Douglas Heaven. 2020. "A GPT-3 Bot Posted Comments on Reddit for a Week and No One Noticed." MIT TECHNOLOGY REVIEW (blog). October 8, 2020.

Georg Heigold, Günter Neumann, and Josef van Genabith. "How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse?." arXiv preprint arXiv:1704.04441 (2017).

Karen Hao. 2020. "A College Kid's Fake, AI-Generated Blog Fooled Tens of Thousands. This Is How He Made It." MIT TECHNOLOGY REVIEW (blog). August 14, 2020.

Robin Jia, and Percy Liang. "Adversarial examples for evaluating reading comprehension systems." arXiv preprint arXiv:1707.07328 (2017).

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. "Deep text classification can be fooled." arXiv preprint arXiv:1704.08006 (2017).

Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, and Farinaz Koushanfar. "Adversarial reprogramming of text classification neural networks." arXiv preprint arXiv:1809.01829 (2018).

Thorsten Quandt, Lena Frischlich, Svenja Boberg andTim Schatto-Eckrodt. (2019). Fake News. 1-6. 10.1002/9781118841570.iejs0128.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8 (2019): 9.

Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).

Jon Vadillo, Roberto Santana, and Jose A. Lozano. "When and How to Fool Explainable Models (and Humans) with Adversarial Examples." arXiv preprint arXiv:2107.01943 (2021).

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. "Universal adversarial triggers for attacking and analyzing NLP." arXiv preprint arXiv:1908.07125 (2019).

Max Wolff, and Stuart Wolff. "Attacking neural text detectors." arXiv preprint arXiv:2002.11768 (2020).

Chaoran Yuan, Xiaobin Liu and Zhengyuan Zhang, "The Current Status and progress of Adversarial Examples Attacks." 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 2021, pp. 707-711, doi: 10.1109/CISCE52179.2021.9445917. (2021)

Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cédric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. "On the Lack of Robust Interpretability of Neural Text Classifiers." arXiv preprint arXiv:2106.04631 (2021).

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. "Defending against neural fake news." arXiv preprint arXiv:1905.12616 (2019).

Thomas Zerback, Florian Töpfl, and Maria Knöpfle. "The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them." New Media & Society 23, no. 5 (2021): 1080-1098.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. "Adversarial attacks on deep-learning models in natural language processing: A survey." ACM Transactions on Intelligent Systems and Technology (TIST) 11, no. 3 (2020): 1-41.

Supplementary Material

Appendix A: Full list of Latin characters with their respective Greek and Cyrillic substitutions and all respective character Unicode.

Original (Basic Latin) Letter ~ Unicode		Greek Letter ~ Unicode		Cyrillic Letter ~ Unicode	
A ~ U+0041	a ~ U+0061	Α ~ U+x0391		А ~ U+x0410	а ~ U+x0430
B ~U+0042	b ~ U+0062	Β ~ U+x0392		В ~ U+x0412	ь ~ U+x044C
C ~ U+0043	c ~ U+0063	Ϲ ~ U+x2CA3	ϲ ~ U+x03C2	С ~ U+x0421	с ~ U+x0441
E ~ U+0045	e ~ U+0065	Ε ~ U+x0395		Е ~ U+x0415	е ~ U+x0435
F ~ U+0046		Ϝ ~ U+x03DC			
H ~ U+0048	h ~ U+0068	Η ~ U+x0397		Н ~ U+x041D	һ ~ U+x04BB
I ~ U+0049	i ~ U+0069	Ι ~ U+x0399		І ~ U+x0406	і ~ U+x0456
J ~ U+004a	j ~ U+006a			Ј ~ U+x0408	ј ~ U+x0458
K ~ U+004b		Κ ~ U+x039A		К ~ U+x041A	
M ~ U+004d		Μ ~ U+x039C		М ~ U+x041C	
N ~ U+004e		Ν ~ U+x039D			
O ~ U+004f	o ~ U+006f	Ο ~ U+x039F	ο ~ U+x03BF	О ~ U+x041E	о ~ U+x043E
P ~ U+0050	p ~ U+0070	Ρ ~ U+x03A1		Р ~ U+x0420	р ~ U+x0440
S ~ U+0053	s ~ U+0073			Ѕ ~ U+x0405	ѕ ~ U+x0455
T ~ U+0054		Τ ~ U+x03A3		Т ~ U+x0422	
V ~ U+0056	v ~ U+0076		ν ~ U+x03BD	Ѵ ~ U+x0474	ѵ ~ U+x0475
	w ~ U+0077				ѡ ~ U+x0461
X ~ U+0058	x ~ U+0078	Χ ~ U+x03A7		Х ~ U+x0425	х ~ U+x0445
Y ~ U+0059	y ~ U+0079	Υ ~ U+x03A5		Ү ~ U+x04AE	у ~ U+x0443
Z ~ U+005a	z ~ U+007a	Ζ ~ U+x036			