ARE WE SAFE YET? THE LIMITATIONS OF DISTRIBUTIONAL FEATURES FOR FAKE NEWS DETECTION

Tal Schuster[1], **Roei Schuster**[2,3], **Darsh J Shah**[1], **Regina Barzilay**[1]

[1]Computer Science and Artificial Intelligence Lab, MIT
[2]Tel Aviv University
[3]Cornell Tech
{tals, darsh, regina}@csail.mit.edu, rs864@cornell.edu

ABSTRACT

Automatic detection of fake news — texts that are deceitful and misleading — is a long outstanding and largely unsolved problem. Worse yet, recent developments in language modeling allow for the automatic generation of such texts. One approach that has recently gained attention detects these fake news using stylometry-based provenance, i.e. tracing a text's writing style back to its producing source and determining whether the source is malicious. This was shown to be highly effective under the assumption that legitimate text is produced by humans, and fake text is produced by a language model.

In this work, we identify a fundamental problem with provenance-based approaches against attackers that auto-generate fake news: fake and legitimate texts can originate from nearly identical sources. First, a legitimate text might be auto-generated in a similar process to that of fake text, and second, attackers can automatically corrupt articles originating from legitimate human sources. We demonstrate these issues by simulating attacks in such settings, and find that the provenance approach fails to defend against them. Our findings highlight the importance of assessing the veracity of the text rather than solely relying on its style or source. We also open up a discussion on the types of benchmarks that should be used to evaluate neural fake news detectors.

1 Introduction

As the performance of language models improves and generated text becomes more realistic, we face rising concerns about the malicious use of these models (Vosoughi et al., 2018; Radford et al., 2019). An example of such misuse is *neural fake news* (Zellers et al., 2019), automatically generating fake articles en masse.[1]

One approach for automating fake news detection is fact-checking. These detectors are motivated by the way humans validate news reports (Nyhan, Reifler, 2015) and focus on identifying falsified information. However, despite a large body of work in this area, existing detectors are not yet sufficiently accurate to automate the detection task (Thorne, Vlachos, 2018; Thorne et al., 2018).

Recently, Zellers et al. (2019) proposed an alternative approach, that relies on text provenance, or source identification. This approach assumes that fake-ness is determined by the source that generated the text. For instance, we might assume that The Associated Press news articles are more accurate than posts from a propaganda website (Baly et al., 2018). Alternatively, the source may reflect whether the article was written by a human or generated by a machine (Hashimoto et al., 2019; Gehrmann et al., 2019; Bakhtin et al., 2019). In Zellers et al. (2019)'s setup, the language model itself is used to extract *language distributional* features from the article that can be traced to a particular text source. These features might implicitly include n-gram frequencies, sentence structures, coherency of text, among others (Pérez-Rosas et al., 2018).

[1] https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

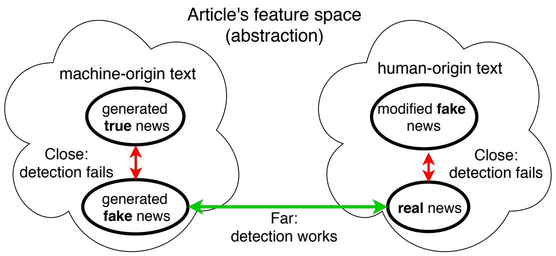


Figure 1: Hypothetical abstraction of the language distributional feature space.

In this paper, we argue that this approach does not fully solve the detection of fake news created by neural language models. Specifically, it fails in two important instances:

1. Discriminating between truthful and fake information that are both machine-generated. Neural language models are increasingly used to produce *any* text. Applications such as auto-completion, text modification, question answering, text simplification, summarization, and others, are growing in popularity. Those might be used both for producing legitimate[2] (Logan et al., 2019) and malicious text, resulting in "fake" and "real" text sources that are almost identical.[3] To illustrate this problem, we consider machine-generated text that is true or false in the context of a given article.
2. Detecting an attacker that uses text from a legitimate human source, but automatically corrupts it to alter its meaning by making only subtle, relatively small changes. Unfortunately, powerful language models designed for text generation, can also easily be used to corrupt existing text. To demonstrate this, we consider an attacker that inverts the negation of statements in a human-written article, guided by a language model's output probabilities on the attacker's modification.

The failure of the provenance-based approach is not surprising since both cases are similar in the distributional features that it relies on. In the first case, both real and fake texts are of an artificial source, as they were directly generated by a language model. In the second, both real and fake texts originate from a human author (though the attacker locally and subtly modifies text by an automated procedure). Figure 1 illustrates this.

Our work highlights the need to consider a much broader landscape of how fake news is generated, when designing automatic detectors. There are many different ways in which machines and humans can collaborate for writing true and false text pieces. Therefore, we recommend an approach that defines fake news based on text truthfulness rather than provenance. Under this approach, detector benchmarks would incorporates truthfulness as a primary indicator for fake-ness.

2 Background

Fact Verification The task of fact-checking is tightly tied to the task of detecting fake (and false) news (Vlachos, Riedel, 2014; Wang, 2017). Fact-checking involves retrieving a potential evidence for a claim and evaluating the stance between them (Popat et al., 2017; Hanselowski et al., 2018; Thorne et al., 2018; Mohtarami et al., 2018; Ma et al., 2019).

A fact-checking system could be applied for false news detection by validating each of the article's claims against a reliable source. However, the performance of current automatic models is still relatively low (Thorne, Vlachos, 2018; Schuster et al., 2019).

Machine-generated Text Detection The recent improvement in the quality of text generated by language models (LM) motivated several studies to examine their differences from human-written text. Hashimoto et al. (2019) combine the output's likelihood with human scores to improve the evaluation of the generations. Gehrmann et al. (2019) visualize the per-token probability to demonstrate that human texts contain less probable tokens, making them distinguishable from LM's outputs. Bakhtin et al. (2019) learn a dedicated provenance neural classifier. While their classifier achieves high in-domain accuracy, they find that it overfits

[2] `http://entm.ag/kpu`

[3] `https://www.wired.com/story/how-bots-ruined-clicktivism`

the generated text distribution rather than detecting outliers from human texts, resulting in increased "humanness" scores for random perturbations. This finding suggests that, given enough text from multiple sources, one can train an accurate provenance classifier.

Building on the above observation, Zellers et al. (2019) focus on fake news and create a model dubbed Grover. Grover can both generate news articles and detect auto-generated news based on provenance. We provide more details about their model below. In Section 3, we describe our adversarial setting in which fake news could be similar in distribution to real news, resulting in the failure of such classifiers, even when fine-tuned on relevant training data.

Grover's Architecture Grover's news *generator* is a Transformer-based LM (Vaswani et al., 2017). Though architecturally similar to GPT-2 (Radford et al., 2019), the generator was trained specifically on news article texts, conditioned on the article meta-data, and its output tokens are sampled using nucleus sampling (Holtzman et al., 2019). The generator is trained with a LM objective on a large news corpus from Common Crawl dumps.

The fake news *detector* is a simple linear classifier on top of the last hidden state of Grover's LM on the examined article. The LM parameters are initialized with the pre-trained generator values, from an earlier checkpoint, and are optimized jointly with the linear layer on the classification training set. To construct this dataset, fake news were generated by the pretrained generator and real news were sampled human-written articles.

3 Adversarial Setting

We adopt an adversarial setting similar to that of Zellers et al. (2019). Our **attacker** wishes to generate *fake* text, that contains unverified or false claims, en masse, using a language model to automate the process. The attacker's goal is to produce text that fools the verifier. Our **verifier** is **adaptive**: it receives a limited set of examples generated by the attacker, and trains a discriminator to detect the attacker's texts from legitimately-produced, *real* text, containing exclusively human-verified claims (news articles from relatively reputable sources, like the The New York Times are assumed to be real). We also experiment with a non-adaptive, **zero-shot** setting, where the verifier does not receive the attacker's examples.

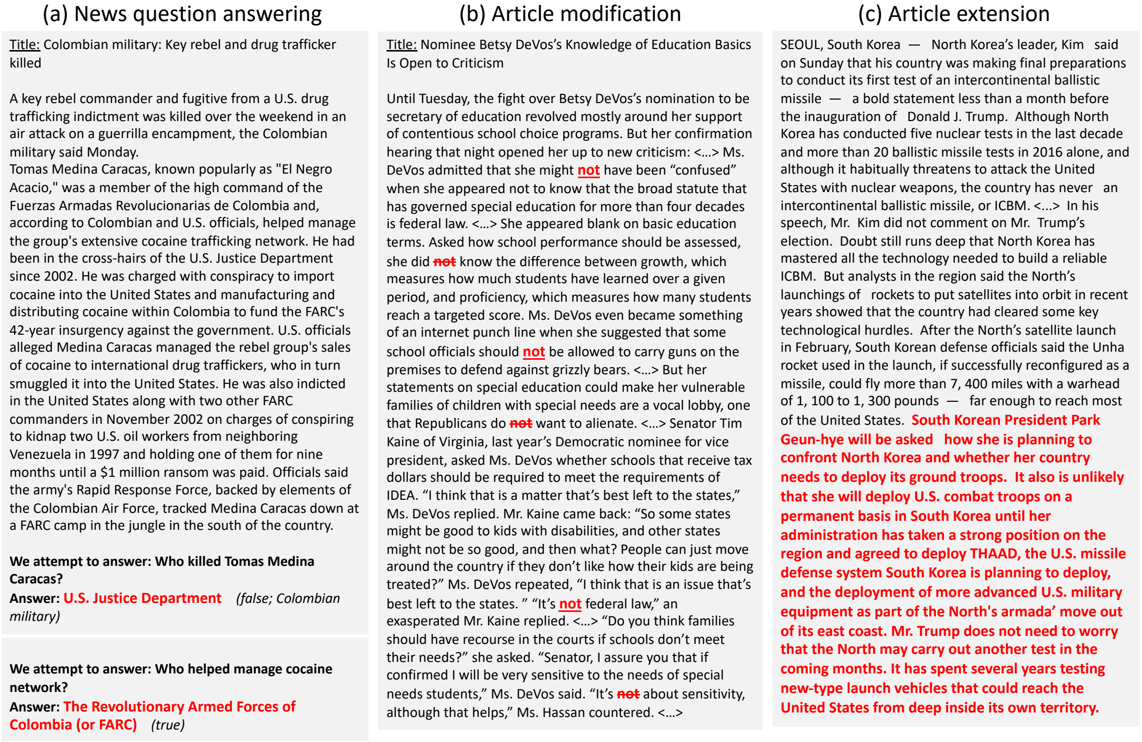
Notably, our attackers focus on automatically creating text with false information. We thus adopt the Council of Europe's view that fake news are purposefully or mistakenly fallacious news, as factual incorrectness is the one inherent difference between real and fake news (Wardle, Derakhshan, 2017). This veracity-focused approach was also used by Pérez-Rosas et al. (2018), though they perform the fake modifications manually. Conversely, the attackers in Zellers et al. (2019) focus on generating "viral and persuasive" content, considered as fake news irrespective of its correctness. The two views are complementary, but as our experiments indicate, truthfulness can be particularly subtle and challenging to detect using language-distributional features.

Our two veracity-based experiments are detailed in Section 4. These attackers are built to exhibit minimal distributional differences from a real news source, but intentionally include false or tampered statements. To assess the capacity of our detector, in Section 5 we show that the same defense performs well even on challenging provenance-based tasks.

Experimental Setup In each experiment, we collected a dataset with a "real" text class and a "fake" text class and used separate samples for testing and for fine-tuning. We used a Grover-Mega discriminator for all of the experiments. The weights of the Grover-Mega generator (used by the discriminator) were initialized from a checkpoint provided by Zellers et al. (2019) and fine-tuned for 10 epochs with our training samples. For evaluating the zero-shot setting defense, we applied a pretrained Grover-Mega discriminator by querying its Web interface.

4 Distributional Features Don't Distinguish Similar Sources

(1) Automatic False vs. True Question Answering
In our first experiment, we simulate a scenario where both the real and the fake texts are machine-generated. Specifically, an auto-completion text generator is used to extend a news article. A responsible user of this generator verifies the correctness of the output, whereas an attacker uses the same tool, but verifies *in*correctness.



(a) News question answering

Title: Colombian military: Key rebel and drug trafficker killed

A key rebel commander and fugitive from a U.S. drug trafficking indictment was killed over the weekend in an air attack on a guerrilla encampment, the Colombian military said Monday.
Tomas Medina Caracas, known popularly as "El Negro Acacio," was a member of the high command of the Fuerzas Armadas Revolucionarias de Colombia and, according to Colombian and U.S. officials, helped manage the group's extensive cocaine trafficking network. He had been in the cross-hairs of the U.S. Justice Department since 2002. He was charged with conspiracy to import cocaine into the United States and manufacturing and distributing cocaine within Colombia to fund the FARC's 42-year insurgency against the government. U.S. officials alleged Medina Caracas managed the rebel group's sales of cocaine to international drug traffickers, who in turn smuggled it into the United States. He was also indicted in the United States along with two other FARC commanders in November 2002 on charges of conspiring to kidnap two U.S. oil workers from neighboring Venezuela in 1997 and holding one of them for nine months until a $1 million ransom was paid. Officials said the army's Rapid Response Force, backed by elements of the Colombian Air Force, tracked Medina Caracas down at a FARC camp in the jungle in the south of the country.

We attempt to answer: Who killed Tomas Medina Caracas?
Answer: **U.S. Justice Department** *(false; Colombian military)*

We attempt to answer: Who helped manage cocaine network?
Answer: **The Revolutionary Armed Forces of Colombia (or FARC)** *(true)*

(b) Article modification

Title: Nominee Betsy DeVos's Knowledge of Education Basics Is Open to Criticism

Until Tuesday, the fight over Betsy DeVos's nomination to be secretary of education revolved mostly around her support of contentious school choice programs. But her confirmation hearing that night opened her up to new criticism: <…> Ms. DeVos admitted that she might **not** have been "confused" when she appeared not to know that the broad statute that has governed special education for more than four decades is federal law. <…> She appeared blank on basic education terms. Asked how school performance should be assessed, she did ~~not~~ know the difference between growth, which measures how much students have learned over a given period, and proficiency, which measures how many students reach a targeted score. Ms. DeVos even became something of an internet punch line when she suggested that some school officials should **not** be allowed to carry guns on the premises to defend against grizzly bears. <…> But her statements on special education could make her vulnerable families of children with special needs are a vocal lobby, one that Republicans do ~~not~~ want to alienate. <…> Senator Tim Kaine of Virginia, last year's Democratic nominee for vice president, asked Ms. DeVos whether schools that receive tax dollars should be required to meet the requirements of IDEA. "I think that is a matter that's best left to the states," Ms. DeVos replied. Mr. Kaine came back: "So some states might be good to kids with disabilities, and other states might not be so good, and then what? People can just move around the country if they don't like how their kids are being treated?" Ms. DeVos repeated, "I think that is an issue that's best left to the states. " "It's **not** federal law," an exasperated Mr. Kaine replied. <…> "Do you think families should have recourse in the courts if schools don't meet their needs?" she asked. "Senator, I assure you that if confirmed I will be very sensitive to the needs of special needs students," Ms. DeVos said. "It's ~~not~~ about sensitivity, although that helps," Ms. Hassan countered. <…>

(c) Article extension

SEOUL, South Korea — North Korea's leader, Kim said on Sunday that his country was making final preparations to conduct its first test of an intercontinental ballistic missile — a bold statement less than a month before the inauguration of Donald J. Trump. Although North Korea has conducted five nuclear tests in the last decade and more than 20 ballistic missile tests in 2016 alone, and although it habitually threatens to attack the United States with nuclear weapons, the country has never an intercontinental ballistic missile, or ICBM. <...> In his speech, Mr. Kim did not comment on Mr. Trump's election. Doubt still runs deep that North Korea has mastered all the technology needed to build a reliable ICBM. But analysts in the region said the North's launchings of rockets to put satellites into orbit in recent years showed that the country had cleared some key technological hurdles. After the North's satellite launch in February, South Korean defense officials said the Unha rocket used in the launch, if successfully reconfigured as a missile, could fly more than 7, 400 miles with a warhead of 1, 100 to 1, 300 pounds — far enough to reach most of the United States. **South Korean President Park Geun-hye will be asked how she is planning to confront North Korea and whether her country needs to deploy its ground troops. It also is unlikely that she will deploy U.S. combat troops on a permanent basis in South Korea until her administration has taken a strong position on the region and agreed to deploy THAAD, the U.S. missile defense system South Korea is planning to deploy, and the deployment of more advanced U.S. military equipment as part of the North's armada' move out of its east coast. Mr. Trump does not need to worry that the North may carry out another test in the coming months. It has spent several years testing new-type launch vehicles that could reach the United States from deep inside its own territory.**

Figure 2: Examples of the "fake news" in our experiments. (a) In the news question answering (Section 4), a CNN article is presented with two examples of questions (bold) from newsQA (Trischler et al., 2017) and Grover's generated answer (red). The first answer is verified by a human annotator to be false and the second as true. (b) In article modification ($m = 6$) (Section 4), the negations are marked with a cross-line for deletions and underline for addition. (c) In the article extension case (Section 5), the bold red text is the generation of GPT-2 medium to extend the human-written prefix.

In order to create such data, we use the newsQA dataset (Trischler et al., 2017) that contains CNN articles with questions about them. In this case, the extension (the answer to the question) could be easily validated. Lastly, to resemble an article extension task, we remove the sentence containing the answer from the article.[4] This question-answering application of language models was previously explored by Radford et al. (2019) and was recently used to automatically generate answers to medical questions.[5]

Following the setting of Radford et al. (2019), we insert the question into a template that is appended to the end of the article:

```
We attempt to answer: <question>
Answer:
```

Then, the first sentence produced by Grover's generator is used as a potential answer. We chose this template by examining common formats of questions in news articles and find it to produce mostly reasonable answers by Grover's generator.[6] For example, for an article about a man caught with a gun in Washington, and the question *"Where was the suspect stopped?"*, Grover generated the answer *"2 blocks from the U.S. Capitol."* See Figure 2 for more examples.

Finally, we manually assess the generated answers. We filter out nonsensical generations (29%) and label the rest as true or false by the article and external knowledge. Answers containing at least one false statement

[4] Post the generation, we remove the article's sentence with the highest word count tf-idf similarity with the question and answer (concatenated).

[5] `https://github.com/re-search/DocProduct`

[6] Evaluating against the dataset's ground-truth answers, Grover achieved token-level overlap F1 score of 0.19 on newsQA's test set and of 0.25 on our filtered evaluation set.

Telling fake from real (veracity-based)		precision	recall	F1	accuracy
adaptive	QA (false vs. true)	0.72	0.71	0.71	71%
	article modification ($m = 2$)	0.53	0.52	0.53	53%
	article modification ($m = 6$)	0.66	0.65	0.65	65%
	article modification ($m = 10$)	0.73	0.47	0.63	65%

Table 1: Detection results on the attacks of Section 4. Zero-shot defense performance (not included) is very low in all cases. We report (macro) F1 score and overall accuracy. Precision, and recall of the "fake" class are reported as well.

were labeled false. For this experiment, cases where the generated answer is true are used as real texts and the ones with false answers as fake texts.

(2) Guided Factual Modifications
In this experiment, the attacker uses a human-written news article and performs subtle modifications to semantically modify statements. This way, the fake, untruthful, article remains distributionally similar to human-written real ones. In this attack, we perform the modifications by removing and adding negations from statements. To remove negations, we randomly delete "not" or "no" occurrences from the text. Then, we look for other statements to add negations. When negations are added to syntactically-correct location, the new sentence is a negative inversion of the original (Rudanko, 1982). In order to find these locations automatically, we use the probabilities of a GPT-2 language model. Specifically, we randomly sample 100 locations in the article and choose the ones with the maximal score, defined as the probability for either "not" or "no", multiplied by the probability for the word following the negation.

To create the dataset for this experiment we use articles from The New York Times.[7] Original articles are labeled as real text and articles to which we invert m statements ($\frac{m}{2}$ by removing negations and $\frac{m}{2}$ by adding negations) are labeled as fake text.[8]

4.1 Results

For both of these attacks, we use a subset of the data for evaluation and the rest for fine-tuning the detector. The results are shown in Table 1. In question answering, the adapted classifier reached 71% in distinguishing true vs. false answers generated by Grover, exceeding a trivial majority baseline (51% on our evaluation set). This suggests that artificially produced "lies" manifest to some extent within the text's language distributional features. This was found true for human-written text as well (Rashkin et al., 2017). In our case, even the length of the answer is indicative: answers labeled as false contained 12.4 words on average compared to true answers, which had only 9.7. This might be due to our requirement for strictly-true text in the "real" class: longer sentences are more likely to include at least one false statement. When evaluated only on the short false answers (up to 10 words), Grover's accuracy is only 62%. Ultimately, however, despite its nontrivial ability to differentiate truth from lie, Grover's performance in this setting is much worse than in distinguishing human from machine text (see Section 5).

For the automatic article modification attack, Grover fails completely to detect two modifications ($m = 2$). Even when we invert up to 10 statements in an article, Grover's performance is still deficient. As we explain in Section 1, this is likely due to the similarity in Grover's feature space between texts of the same origin (where only slight — but possibly semantically significant — modifications were introduced). Even though the changes include negations in machine-originated locations and might introduce surprising factual claims, the distributional difference is subtle enough to cause the fine-tuned classifier to underperform.

5 Baselines: Distributional Features for Provenance-based Defenses

One might suspect that the low performance in Section 4 is due to limited capacity of Grover's detector. However, in this section we show that the detector succeeds in similar settings in which the classes are split by the generation source. Beyond distinguishing full texts sampled from a language model vs. human

[7] From *All the news* dataset: `https://www.kaggle.com/snapcrack/all-the-news`

[8] This attacker intentionally avoids changing the overall number of negations in corrupted articles, since the number of negations can be used as a fake-ness indicator.

Telling human from machine		precision	recall	F1	accuracy
zero-shot	full article	0.84	0.98	0.90	90%
	article extension ($g = 20\%$)	0.52	0.20	0.45	51%
	article extension ($g = 1\%$)	0.07	0.01	0.28	37%
adaptive	full article	0.93	0.94	0.94	94%
	article extension ($g = 20\%$)	0.90	0.97	0.95	95%
	article extension ($g = 1\%$)	0.91	0.95	0.94	95%
	QA (machine vs. human)	0.82	0.86	0.83	83%

Table 2: Detection results on the attacks of Section 5 in a zero-shot setting and the adaptive setting (where the discriminator is fine-tuned to a specific attacker). We report (macro) F1 score and overall accuracy. Precision, and recall of the "fake" class are reported as well.

ones (Zellers et al., 2019), we also experiment with more sophisticated attackers that automatically extend articles.

(1) Fully-generated Articles
First, unlike Zellers et al. (2019), we evaluate Grover against texts generated by a different model of similar size. For this out of domain attack, we use an unconditioned GPT-2 XL model (Radford et al., 2019). The dataset contains:
Real text: Examples from the WebText test set.[9]
Fake text: Generations from the released outputs of the GPT-2 XL model.

(2) Automatic Article Extension
In order to experiment with mixed human-machine articles, we build an attacker that automatically extends unfinished human-written articles. We simulate this attacker by querying a GPT-2 medium model, conditioned on the the first 500 words from an article.[10]

Real text: Articles from The New York Times.
Fake text: Articles that were automatically extended. We keep only the first few generated sentences in order to explore the defense against different percentages (g) of machine-generated text. The examples of the first attack were also included in order to preserve the ability to classify full articles.

(3) Automatic Question Answering (provenance detection)
In this experiment, we test whether the QA defense from Section 4 can perform better when it knows that true texts are always human-written. To this end, we repeat the same setting but use the gold answers from the newsQA dataset for the real text class.

5.1 Results

Zero-shot Setting As Table 2 depicts, the zero-shot classifier is effective in detecting the fully generated articles of a different model, with a 0.9 macro-F1 score.

The extended articles, which contain a substantial amount of human-written text, are mostly classified as human-written in this setting. This is unsurprising considering the dataset of full articles that it was trained on.

Adaptive Setting After fine-tuning on relevant examples, Grover improves the full article score and also performs well against article extension generations, reaching 0.94 macro-F1 for articles with a single generated sentence ($g = 1\%$). This demonstrates that when we can assume a text generator is only used by an attacker, having access to attacker-produced training examples enables a fine-tuned Grover to detect that source.

Although the question-answering setting is similar to article extension ($g = 1\%$) by containing a single machine-generated sentence, the adapted detector's score is 0.09 points lower. This drop in performance can be explained by using the (stronger) Grover-Mega model for the QA setting, compared to GPT-2 medium for article extension. Also, the appended template was chosen by empirically selecting the one that makes

[9] `https://github.com/openai/gpt-2-output-dataset`

[10] To use GPT2 for automatic generations online: `https://transformer.huggingface.co`

the generated answers look most reasonable for a human reader. Still, the score of the provenance-based QA defense is 0.12 points higher than the veracity-based QA. Therefore, in order to achieve high scores with this distributional-based defense, we need to assume that text generators are used only to create fake cases.

Overall, we conclude that this defense is highly effective when the sources of fake text and real text are different.

6 Conclusion

Previous studies on automatic fake news detection mainly fall into two categories. In the first, more common type, detection is based on the provenance of the article. In the second approach, detectors analyze the content and verify it against reliable sources.

Recently, improvements to automatic text generators were shown to enable the creation of realistic-looking news articles that are based on made-up facts. While a human reader might confuse these articles with real ones, their distributional features are sufficiently different for a provenance-based defense to detect them.

However, this approach does not account for the scenario where generators are also used for producing legitimate text, and nor for more sophisticated attackers that use the generator to create malicious content while keeping minimal distributional differences from a legitimate source.

We demonstrate the first setting by considering auto-completion of news articles with correct information. For demonstrating the second setting, we consider an attacker that uses the probabilities assigned by a language model to guide minimal edits that modify the correctness of an article's statements. Thus, defenses that perform detection of auto-generated text extremely well can still be fooled by generator-based attackers.

To inform the development of better detectors against all types of fake news, it is important to build diverse and challenging datasets. We recommend to extend our datasets and create a benchmark that represents content's veracity in a wide range of human-machine collaborating applications, from whole article generation to hybrid writing and editing. This reflects a definition of fake news that incorporates veracity rather than provenance.

7 Acknowledgments

We thank the members of the MIT NLP group for their helpful discussion and comments. Roei Schuster is a member of the Check Point Institute of Information Technology. This work is supported in part by DSO grant DSOCL18002; by the Blavatnik Interdisciplinary Cyber Research Center (ICRC); by the NSF award 1650589; and by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program.

References

Bakhtin Anton, Gross Sam, Ott Myle, Deng Yuntian, Ranzato Marc'Aurelio, Szlam Arthur. Real or Fake? Learning to Discriminate Machine from Human Generated Text // arXiv preprint arXiv:1906.03351. 2019.

Baly Ramy, Karadzhov Georgi, Alexandrov Dimitar, Glass James, Nakov Preslav. Predicting Factuality of Reporting and Bias of News Media Sources // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 3528–3539.

Gehrmann Sebastian, Strobelt Hendrik, Rush Alexander. GLTR: Statistical Detection and Visualization of Generated Text // Proceedings of the 57th Conference of the Association for Computational Linguistics: System Demonstrations. Florence, Italy: Association for Computational Linguistics, VII 2019. 111–116.

Hanselowski Andreas, Avinesh PVS, Schiller Benjamin, Caspelherr Felix, Chaudhuri Debanjan, Meyer Christian M, Gurevych Iryna. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task // Proceedings of the 27th International Conference on Computational Linguistics. 2018. 1859–1874.

Hashimoto Tatsunori, Zhang Hugh, Liang Percy. Unifying Human and Statistical Evaluation for Natural Language Generation // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 1689–1701.

Holtzman Ari, Buys Jan, Forbes Maxwell, Choi Yejin. The curious case of neural text degeneration // arXiv preprint arXiv:1904.09751. 2019.

Logan Robert, Liu Nelson F., Peters Matthew E., Gardner Matt, Singh Sameer. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 5962–5971.

Ma Jing, Gao Wei, Joty Shafiq, Wong Kam-Fai. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 2561–2571.

Mohtarami Mitra, Baly Ramy, Glass James, Nakov Preslav, Màrquez Lluís, Moschitti Alessandro. Automatic Stance Detection Using End-to-End Memory Networks // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 767–776.

Nyhan Brendan, Reifler Jason. Estimating fact-checking's effects // Arlington, VA: American Press Institute. 2015.

Pérez-Rosas Verónica, Kleinberg Bennett, Lefevre Alexandra, Mihalcea Rada. Automatic Detection of Fake News // Proceedings of the 27th International Conference on Computational Linguistics. 2018. 3391–3401.

Popat Kashyap, Mukherjee Subhabrata, Strötgen Jannik, Weikum Gerhard. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media // Proceedings of the 26th International Conference on World Wide Web Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017. 1003–1012. (WWW '17 Companion).

Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, Sutskever Ilya. Language models are unsupervised multitask learners // OpenAI Blog. 2019. 1, 8.

Rashkin Hannah, Choi Eunsol, Jang Jin Yea, Volkova Svitlana, Choi Yejin. Truth of varying shades: Analyzing language in fake news and political fact-checking // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. 2931–2937.

Rudanko Juhani. Towards a description of negatively conditioned subject operator inversion in English // English Studies. 1982. 63, 4. 348–359.

Schuster Tal, Shah Darsh J, Yeo Yun Jie Serene, Filizzola Daniel, Santus Enrico, Barzilay Regina. Towards Debiasing Fact Verification Models // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019.

Thorne James, Vlachos Andreas. Automated Fact Checking: Task Formulations, Methods and Future Directions // Proceedings of the 27th International Conference on Computational Linguistics. 2018. 3346–3359.

Thorne James, Vlachos Andreas, Christodoulopoulos Christos, Mittal Arpit. FEVER: a Large-scale Dataset for Fact Extraction and VERification // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018. 809–819.

Trischler Adam, Wang Tong, Yuan Xingdi, Harris Justin, Sordoni Alessandro, Bachman Philip, Suleman Kaheer. NewsQA: A Machine Comprehension Dataset // Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada: Association for Computational Linguistics, VIII 2017. 191–200.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, Polosukhin Illia. Attention is all you need // Advances in neural information processing systems. 2017. 5998–6008.

Vlachos Andreas, Riedel Sebastian. Fact Checking: Task definition and dataset construction // Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Baltimore, MD, USA: Association for Computational Linguistics, VI 2014. 18–22.

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

Wang William Yang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, VII 2017. 422–426.

Information Disorder: Toward an interdisciplinary framework for research and policy making. // . 2017.

Zellers Rowan, Holtzman Ari, Rashkin Hannah, Bisk Yonatan, Farhadi Ali, Roesner Franziska, Choi Yejin. Defending Against Neural Fake News // arXiv preprint arXiv:1905.12616. 2019.