Synthetic Disinformation Attacks on Automated Fact Verification Systems

Yibing Du†*, Antoine Bosselut‡*, Christopher D. Manning†
†Stanford University ‡EPFL
antoine.bosselut@epfl.ch, {yibingdu, manning}@stanford.edu

Abstract

Automated fact-checking is a needed technology to curtail the spread of online misinformation. One current framework for such solutions proposes to verify claims by retrieving supporting or refuting evidence from related textual sources. However, the realistic use cases for fact-checkers will require verifying claims against evidence sources that could be affected by the same misinformation. Furthermore, the development of modern NLP tools that can produce coherent, fabricated content would allow malicious actors to systematically generate adversarial disinformation for fact-checkers.

In this work, we explore the sensitivity of automated fact-checkers to synthetic adversarial evidence in two simulated settings: ADVERSARIAL ADDITION, where we fabricate documents and add them to the evidence repository available to the fact-checking system, and ADVERSARIAL MODIFICATION, where existing evidence source documents in the repository are automatically altered. Our study across multiple models on three benchmarks demonstrates that these systems suffer significant performance drops against these attacks. Finally, we discuss the growing threat of modern NLP systems as generators of disinformation in the context of the challenges they pose to automated fact-checkers.

1 Introduction

From QAnon’s deep state to anti-vaccination campaigns, misinformation and disinformation have flourished in online ecosystems. As misinformation continues to induce harmful societal effects, factchecking of online content has become a critical need to ensure trust in the information found online. However, systematic disinformation campaigns make manual efforts to filter misinformation unrealistic as claims must be verified promptly before they are allowed to spread and affect public opinion. Consequently, new research in automated fact-checking explores designing systems that can rapidly validate political, medical, and other domain-specific claims made and shared online (Thorne and Vlachos 2018; Guo, Zhang, and Lu 2019).

A popular emergent paradigm in automated fact-checking, Fact Extraction and Verification (FEVER; Thorne et al. 2018), frames the problem as claim verification against a large repository of evidence documents. As one of the first large-scale datasets designed in this framework, FEVER was released with 185k annotated claims that could be verified against Wikipedia articles. When checking a claim, systems designed in this framework search for related documents in the database, and retrieve relevant supporting or refuting evidence from these sources. Then, these systems evaluate whether the retrieved evidence sentences validate or contradict the claim, or whether there is not enough information to make a judgment. More recently, the SciFACT (Wadden et al. 2020) and COVIDFACT (Saakyan, Chakrabarty, and Muresan 2021) benchmarks re-purposed this framework for the sci-
entific domain by releasing datasets of medical claims to be verified against scientific content (Wang et al. 2020). While this framework has led to impressive advances in fact verification performance (Ye et al. 2020; Pradeep et al. 2021), current benchmarks assume that the available evidence database contains only valid, factual information.

However, check-worthy claims are often made about new events that may not be verifiable against extensive catalogues, and that must be checked rapidly to avoid strategic disinformation spread (Vosoughi, Roy, and Aral 2018). Consequently, deployed fact-checkers will need to operate in settings where their available evidence is sourced from contemporaneous reporting that may be sensitive to the same misinformation they are designed to verify. Currently, malicious actors remain limited by the cost of running disinformation campaigns (DiResta et al. 2018) and the risks of operational discovery,\(^3\) impeding the scale at which they can deploy these campaigns, and thus the balance of real and false content that fact-checkers must distinguish. However, the development of NLP tools capable of generating coherent disinformation (Zellers et al. 2019; Kreps, McCain, and Brundage 2020b; Buchanan et al. 2021) would allow malicious actors to over-load contemporaneous content with adversarial information and bias the evidence bases used by automated fact-checkers.

Furthermore, even in settings where claims may be verified against established and trusted knowledge, disinformation can still find its way into repositories of documents used by fact-checking systems (Kumar, West, and Leskovec 2016). Wikipedia, for example, which underlies FEVER (and other benchmarks; Petroni et al. 2021), acknowledges that much of the content on the platform may be incorrect, and remain so for long periods of time.\(^4\) For example, the Croatian Wikipedia was systematically contaminated by pro-nationalist bias over a period of at least 10 years.\(^5\) Moreover, studies have uncovered articles on Wikipedia that were edited to provide favorable accounts on specific topics (\textit{e.g.}, workers at a medical device company edited articles to present an optimistic view toward treatments that used their product\(^6\)). Modern NLP tools would allow malicious users to scale up production of disinformation on these platforms, and increase the perception of false consensus or debate on these topics.

In this paper, we evaluate whether automated disinformation generators can effectively contaminate the evidence sets of fact verification systems, and demonstrate that synthetic disinformation drastically lowers the performance of these systems. We define adversarial attacks in two settings: ADVERSARIAL ADDITION (ADVADD; §3), where synthetically-generated documents are added to the document base, and ADVERSARIAL MODIFICATION (ADVMod; §4), where additional automatically-generated information is inserted into existing documents. In both settings, we curate a large collection of adversarial disinformation documents that we inject into the pipelines of existing fact-checking systems developed for the FEVER, SCIFACT, and COVIDFACT shared tasks.\(^7\)

Our results demonstrate that these systems are significantly affected by the injection of poisoned content in their evidence bases, with large absolute performance drops on all models in both settings. Furthermore, our analysis demonstrates that these systems are sensitive to even small amounts of evidence contamination, and that synthetic disinformation is more influential at deceiving fact verification systems compared to human-produced false content. Finally, we provide a discussion of our most important findings, and their importance in the context of continued advances in NLP systems.

2 Background

In this section, we review the formulation of automated fact checking as fact extraction and verification, and recent advances in automated generation of textual disinformation.

Automated Fact-checking: Task

Current research into automated fact-checking generally follows the fact verification and extraction procedure of receiving a textual claim expressed in natural language \(\text{(e.g., “Hypertension is a common comorbidity seen in COVID-19 patients”)},\) collecting supporting evidence from a repository of available documents \(\text{(e.g., scientific manuscripts), and making a prediction as to the claim’s veracity based off the collected supporting evidence. Below, we define the two stages of this pipeline: evidence retrieval and claim verification.}\)

Evidence retrieval The evidence retrieval stage of fact verification systems is typically decomposed into two steps: \textit{document retrieval} and \textit{sentence retrieval}. During document retrieval, documents in the evidence repository that are relevant to the claim are selected. Existing methods typically use information retrieval methods to rank documents based on relevance \(\text{(Thorne et al. 2018; Wadden et al. 2020) or use public APIs of commercial document indices (Hanselowski et al. 2019; Saakyan, Chakrabarty, and Muresan 2021) to crawl related documents. In the sentence retrieval stage, individual sentences from these retrieved documents are selected with respect to their relevance to the claim, often using textual entailment (Hanselowski et al. 2019), or sentence similarity \(\text{(Thorne et al. 2018) methods. Typically, the number of retrieved sentences is capped for computational efficiency.}\)

Claim verification The claim verification stage of the pipeline evaluates the veracity of the claim with respect to the evidence sentences retrieved in the previous stage. Depending on the content found in the supporting sentences, each claim can typically be classified as \textit{supported} (SUP), \textit{refuted} (REF), or \textit{not enough information} (NEI, though some benchmarks omit this label). Systems must aggregate and weigh the evidence sentences to predict the most likely label.

\(^1\)\text{https://www.lawfareblog.com/outourcing-disinformation}\n\(^2\)\text{https://en.wikipedia.org/wiki/}\n\(^3\)\text{Wikipedia:Wikipedia_is_not_a_reliable_source}\n\(^4\)\text{https://meta.wikimedia.org/wiki/}\n\(^5\)\text{Croatian_Wikipedia_Disinformation_Assessment-2021}\n\(^6\)\text{https://www.thecut.com/business/archive/2015/08/wikipedia-editors-for-pay/393926/}\n\(^7\)\text{We will release these documents under a Terms of Use to promote research in fact-checking systems in adversarial settings}
Automated Fact-checking: Datasets

We briefly describe the provenance and structure of our studied datasets and refer the reader to the original works for in-depth descriptions of the construction of these resources.

**FEVER** The FEVER testbed (Thorne et al. 2018) is a dataset of 185,445 claims (145,449 train, 19,998 dev, 19,998 test) with corresponding evidence to validate them drawn from articles in Wikipedia. Because of its scale and originality, the FEVER dataset is one of the most popular benchmarks for evaluating fact verification systems (Yoneda et al. 2018; Nie, Chen, and Bansal 2019; Zhou et al. 2019; Zhong et al. 2020; Subramanian and Lee 2020).

**SIFIC** The SIFIC dataset (Wadden et al. 2020) contains 1,409 expert-annotated scientific claims and associated paper abstracts. SIFIC presents the challenge of understanding scientific writing as systems must retrieve relevant sentences from paper abstracts and identify if the sentences support or refute a presented scientific claim. It has emerged as a popular benchmark for evaluating scientific fact verification systems (Pradeep et al. 2021).

**COVID** The COVID dataset (Saakyan, Chakrabarty, and Muresan 2021) contains 1,296 crowdsourced claims crawled (and filtered) from the /r/COVID19 subreddit. The evidence is composed of documents provided with these claims when they were posted on the subreddit along with resources from Google Search queries for the claims. Refuted claims were automatically-generated by altering key words in the original claims.

Synthetic Disinformation Generation

Recent years have brought considerable improvements in the language generation capabilities of neural language models (Lewis et al. 2020; Ji et al. 2020; Brown et al. 2020; Holtzman et al. 2020), allowing users of these systems to pass off their generations as human-produced (Ippolito et al. 2020). These advances have raised dual-use concerns as to whether these tools could be used to generate text for malicious purposes (Radford et al. 2019; Bonnmasani et al. 2021), which humans would struggle to detect (Clark et al. 2021).

Specific studies have focused on whether neural language models could be used to generate disinformation that influences human readers (Kreps, McCain, and Brundage 2020a; Buchanan et al. 2021). One such study directly explored this possibility by training GROVER, a large-scale, billion-parameter language model on a large dataset of newswire text with the goal of generating text that resembles news (Zellers et al. 2019). In human evaluations of the model’s generated text, the study found that human readers considered the synthetically-generated news to be as trustworthy as human-generated content. While the authors found that neural language models could identify fake, generated content when finetuned to detect distributional patterns in the generated text, they hypothesized that future detection methods would need to rely on external knowledge (e.g., FEVER).

3 Adversarial Addition: Evidence Repository Poisoning

In this section, we simulate the potential vulnerability of fact-checking models to database pollution with misinformation documents by injecting synthetically-generated false documents into the evidence sets of fact verification models, and assess the impact on the performance of these systems.

**Approach**

Our method, Adversarial Addition (ADVADD), uses GROVER to produce synthetic documents for a proposed claim, and makes these fake documents available to the fact verification system when retrieving evidence. As GROVER requires a proposed article title and publication venue (i.e., website link) as input to generate a fake article, we use each claim as a title and set the article venue to wikipedia.com. We generate 10 articles for each claim and split them into paragraphs (n.b., FEVER DB contains first paragraphs of Wikipedia articles and SIFIC contains abstracts of scientific articles). Statistics for the number of documents generated for each benchmark are reported in Table 3. Additional implementation details for the experimental setting of each benchmark can be found in the Appendix.
**FEVER Study**

**Setup** For the FEVER benchmark, we select three high-ranking models from the leaderboard\(^8\) with open-source implementations: KGAT (Liu et al. 2020b), CorefBERT (Ye et al. 2020), and MLA (Kruengkrai, Yamagishi, and Wang 2021). For document retrieval, all models use the rule-based method developed by Hanselowski et al. (2019), which uses the MediaWiki API to retrieve relevant articles based on named entity mentions in the claim. For each claim and poisoned document, we extract all keywords and retrieve associated Wikipedia pages. If we find overlaps between the associated Wikipedia pages of a claim and a poisoned document, then the poisoned document is matched with the claim for document retrieval. Once the retrieved documents are available, the KGAT and CorefBERT models use a BERT-based (Devlin et al. 2019) sentence retriever to rank evidence sentences based on relevance to the claim (trained using pairwise loss). The MLA sentence retriever expands on this approach with hard negative sampling from the same retrieved documents to more effectively discriminate context-relevant information that is irrelevant to the claim. Claim verifiers vary between models, but are generally based off pretrained language models (e.g., CorefBERT, MLA) or graph neural networks (e.g., KGAT). We use the REF and NEI claims from the FEVER development set to study how the preceding systems are affected by the introduction of poisoned evidence.\(^9\)

**Impact of AdvAdd** We report the overall (and claim-stratified) performance change of the tested models in Table 2. For all models, we see a significant performance drop when GROVER-generated paragraphs are introduced into the available evidence set (AdvAdd-full), indicating that fact verification models are sensitive to synthetically-generated information. This drop approaches the performance of an oracle (AdvAdd-oracle), where only GROVER-generated documents are made available as evidence.

As confirmation that these attacks work as expected, we depict in Figure 2 how model predictions change once the synthetic disinformation is added to the evidence set. A significant number of claims that were originally predicted as REF or NEI are now predicted as SUP with the injected poisoned evidence. Consequently, we conclude that the poisoned evidence affects the model’s predictions in the intended way, and that cross-label changes for different pairings are rare. Furthermore, we also find that replacing the retrieved poisoned evidence with random retrieved evidence from FEVERDB does not cause the same performance drop (\(~\%\) vs. \(~\%\)), indicating that these effects are caused by the injection of poisoned evidence, and not merely the removal of potentially relevant evidence (see Appendix B for further details).

**Effect of disinformation scale** We also evaluate a setting where the attack is limited to retrieving only a single contaminated evidence sentence (AdvAdd-min). The performance drops in the min setting are still considerable, suggesting that even limited amounts of injected disinformation can significantly affect downstream claim verification performance.

Moreover, Figure 3 shows a histogram of the number of poisoned evidence sentences retrieved per claim and a strat-

---

\(^8\)https://competitions.codalab.org/competitions/18814#results

\(^9\)We discuss results related to SUP claims in the Appendix.

### Table 2: Effect of AdvAdd on FEVER claim verification. We **bold** the largest performance drop relative to the original evidence.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>CorefBERT Acc. (Ye et al. 2020)</th>
<th>KGAT Acc. (Liu et al. 2020b)</th>
<th>MLA Acc. (Kruengkrai et al. 2021)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>REF</td>
<td>NEI</td>
</tr>
<tr>
<td>Original</td>
<td>73.05</td>
<td>74.03</td>
<td>72.07</td>
</tr>
<tr>
<td>AdvAdd-min</td>
<td>34.80</td>
<td>47.22</td>
<td>22.38</td>
</tr>
<tr>
<td>AdvAdd-full</td>
<td><strong>28.59</strong></td>
<td><strong>39.63</strong></td>
<td><strong>17.54</strong></td>
</tr>
<tr>
<td>AdvAdd-oracle</td>
<td>21.18</td>
<td>27.09</td>
<td>15.26</td>
</tr>
</tbody>
</table>

---

### Table 3: Corpus statistics of evidence repositories

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Evidence Source</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEVER</td>
<td>FEVERDB</td>
<td>5,416,537</td>
</tr>
<tr>
<td></td>
<td>GROVER Docs</td>
<td>995,829</td>
</tr>
<tr>
<td></td>
<td>MediaCloud Docs</td>
<td>74,273,342</td>
</tr>
<tr>
<td>SciFact</td>
<td>Scientific Abstracts</td>
<td>5,183</td>
</tr>
<tr>
<td></td>
<td>GROVER Docs</td>
<td>21,963</td>
</tr>
<tr>
<td>COVIDFact</td>
<td>Google Search Results</td>
<td>1,003</td>
</tr>
<tr>
<td></td>
<td>GROVER Docs</td>
<td>2,709</td>
</tr>
</tbody>
</table>

---

Figure 2: CorefBERT’s predictions change from REF and NEI to SUP once AdvAdd poisons the evidence set.
We note that claims labeled NEI are far more sensitive to the (48.6% precision). While well-trained workers will improve (a,c) and NEI (b,d) claim verification accuracy for A the potential for such an attack to remain undetected. readers, pointing to the quality of the synthetic content, and at recognizing synthetic content, our results demonstrate the do not distinguish machine- from human-generated evidence mate the number of poisoned sentences (23.6% recall), and machine-generated. Our results show that humans underesti- e evidence examples (which could be from cal Turk where we presented three workers with five retrieved and 500 NEI claims from ebiewski 2018). is presented to a particular viewpoint (i.e., data voids; Boyd and Golebiewski 2018).

Quality of poisoned evidence We also evaluate whether poisoned evidence produced by ADVADD is of sufficient quality to bypass potential human detectors. For 500 REF and 500 NEI claims from FEVER, we ran a study on Mechani- cal Turk where we presented three workers with five retrieved evidence examples (which could be from ADVADD or from FEVERDB) and asked them to identify which examples were machine-generated. Our results show that humans underesti- mate the number of poisoned sentences (23.6% recall), and do not distinguish machine- from human-generated evidence (48.6% precision). While well-trained workers will improve at recognizing synthetic content, our results demonstrate the challenge of distinguishing these evidence sources for lay readers, pointing to the quality of the synthetic content, and the potential for such an attack to remain undetected.

<table>
<thead>
<tr>
<th>Source</th>
<th>Evidence Retrieval</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROVER</td>
<td>87%</td>
<td>65%</td>
</tr>
<tr>
<td>MediaCloud</td>
<td>99%</td>
<td>17%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Claim Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>KGAT CorefBERT</td>
</tr>
<tr>
<td>+ GROVER</td>
<td>70.76 73.05</td>
</tr>
<tr>
<td>+ MediaCloud</td>
<td>29.02 28.59</td>
</tr>
<tr>
<td></td>
<td>67.10 70.44</td>
</tr>
</tbody>
</table>

Table 4: Statistics and performance relative to the source of poisoned evidence: GROVER or MediaCloud

Comparison with human-compiled online evidence While we have shown that synthetic disinformation affects the performance of downstream claim verifiers when present in their evidence sets, the threat should be evaluated in comparison to the threat of already existing online misinformation on the same topic. Consequently, we use the MediaCloud\(^{10}\) content analysis tool to crawl web content related to FEVER claims. We crawl English-language news since January 2018 that contains the keywords of a claim anywhere in their text and extract articles with a title that contains at least one keyword from the claim. Finally, we process these articles to have the same format as the original Wikipedia database, yielding 74M total available documents for retrieval (Table 3).

In Table 4, we report the performance of an ADVADD setting where only MediaCloud-crawled documents are available to the retriever compared to our original setting where GROVER-generated documents were available. We observe that evidence crawled from online content has less of an influence on downstream fact verification performance (~3\% vs. ~40\% performance drop). While we are able to retrieve far more documents from MediaCloud due to the size of the database (99\% of claims retrieve a document from MediaCloud), the sentences from ADVADD-GROVER documents are selected more frequently in the sentence retrieval step. While this gap would likely be less pronounced with more contentious claims that yield competing viewpoints (Bush and Zaheer 2019), these results demonstrate that synthetic disinformation can be much more targeted to a particular claim of interest. Figure 3 supports this conclusion, where we observe smaller performance drops for ADVADD-MediaCloud (c,d) compared to ADVADD-GROVER (a,b) even when all retrieved sentences are sourced from the poisoned evidence.

SciFACT Study

Setup For SciFACT, we chose three systems for testing our attack: VeriSci (Wadden et al. 2020), ParagraphJoint (Li, Burns, and Peng 2021), and SciKGAT (Liu et al. 2020a). The VeriSci model was released by the creators of the SciFACT benchmark and retrieves relevant abstracts to a claim using TF-IDF. The ParagraphJoint model, one of the top systems

\(^{10}\) mediacloud.org
We generate a RoBERTa-based model predicts a veracity label. We generate individual evidence sentences (Reimers and Gurevych 2019). These paraphrases generally retain the meaning of the claim, but often remove contextualizing information that would be found in the context of the article in which the new sentence is inserted. Because the paraphrase method attempts to produce synthetic evidence that is semantically equivalent to the original claim, we test its efficacy relative to a method that merely introduces irrelevant content to the evidence document. Motivated by Jia and Liang (2017), we alter a claim by applying heuristics such as number alteration, antonym substitution, and entity replacement with close neighbors according to embedding similarity (Bojanowski et al. 2017). These modifications should not confuse humans, but would affect sensitive fact verification systems, providing a competitive baseline for assessing the strength of ADVMod-paraphrase.

Finally, our oracle reports the performance when the claim itself is appended to an evidence document.

### Results
Our results in Table 6 demonstrate that injecting poisoned evidence sentences into existing documents is an effective method for fooling fact verification systems. Our ADVMod-paraphrase method causes a significant drop on all tested models for both REF and NEI labeled claims. Furthermore, we also note that ADVMod-paraphrase achieves larger performance drops than the baseline method, ADVMod-KeyReplace, for most claim types (the KGAT model is slightly more sensitive to the baseline ADVMod-KeyReplace for the NEI claims), indicating that injections of disinformative content are more effective than non-targeted perturbances to the evidence (e.g., ADVMod-KeyReplace).

<table>
<thead>
<tr>
<th>Model</th>
<th>Evidence Set</th>
<th>Sentence selection</th>
<th>Sentence label</th>
<th>Abstract label</th>
<th>Abstract rationalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>VeriSci (Wadden et al. 2020)</td>
<td>Original</td>
<td>47.69</td>
<td>42.62</td>
<td>51.03</td>
<td>48.45</td>
</tr>
<tr>
<td></td>
<td>ADVAdd</td>
<td><strong>27.05</strong></td>
<td><strong>23.50</strong></td>
<td><strong>25.57</strong></td>
<td><strong>24.33</strong></td>
</tr>
<tr>
<td>SciKGAT (Liu et al. 2020a)</td>
<td>Original</td>
<td>55.61</td>
<td>51.69</td>
<td>58.04</td>
<td>57.41</td>
</tr>
<tr>
<td></td>
<td>ADVAdd</td>
<td><strong>39.44</strong></td>
<td><strong>36.97</strong></td>
<td><strong>37.46</strong></td>
<td><strong>36.98</strong></td>
</tr>
<tr>
<td>ParagraphJoint (Li, Burns, and Peng 2021)</td>
<td>Original</td>
<td>53.63</td>
<td>43.59</td>
<td>55.52</td>
<td>49.55</td>
</tr>
<tr>
<td></td>
<td>ADVAdd</td>
<td><strong>37.68</strong></td>
<td><strong>32.60</strong></td>
<td><strong>41.31</strong></td>
<td><strong>37.12</strong></td>
</tr>
</tbody>
</table>

Table 5: Effect of ADVAdd evidence on the SciFact benchmark. We bold performance drops relative to the original evidence.

on the SciFact leaderboard, uses a word embedding-based method to retrieve abstracts. Both use a RoBERTa-based module for rationale selection and label prediction. The SciKGAT model uses the same evidence retrieval as VeriSci, but the KGAT model (Liu et al. 2020b) to verify claims. We use the 300 claims from the development set to evaluate our method. We generate GroVER articles as with FEVER, but we set the venue URL to medicalnewstoday.com, which produces articles more likely to reflect scientific and medical content.

### Results
In Table 5, we observe large performance drops across all metrics for all models. Furthermore, we note that our disinformation generator, GroVER, is not trained on large quantities of scientific documents of the same format as the original evidence. Despite producing documents that are stylistically different, the disinformation is still retrieved as evidence, and affects the performance of the verifier.

### COVIDFact Study

**Setup** We run our analysis on the baseline system from Saakyan, Chakrabarty, and Muresan (2021). This model retrieves evidence documents for claims using Google Search and then selects evidence sentences based off high cosine similarity between sentence embeddings of the claims and individual evidence sentences (Reimers and Gurevych 2019). A RoBERTa-based model predicts a veracity label. We generate ADVAdd articles in the same manner as for SciFact, and run our analysis on the 271 REF claims from the test set.

**Results** In both the Top-1 and Top-5 settings from Saakyan, Chakrabarty, and Muresan (2021), we observe a ~14.4% performance drop on REF claims (83.8% → 69.4%). We note that COVIDFact random and majority accuracy is only 67.6% due to a label imbalance.

## 4 ADVISORIAL MODIFICATION: Evidence Document Poisoning

In Section 3, we investigated the effect of adding disinformation documents to the evidence repositories of fact verification systems, simulating the setting where the dynamic pace of news might lead to fake information being used to verify real-time claims. However, even in settings where information has more time to settle and facts to crystallize, misinformation can still find its way into repositories of documents used by fact-checking systems. Motivated by the possibility of malicious edits being made to crowdsourced information resources, we explore how NLP methods could be used to automatically edit existing articles with fake content at scale.

### Approach

Our method, ADVISORIAL MODIFICATION (ADVMod), simulates this setting in a two-stage process. First, we use off-the-shelf NLP tools to generate modified versions of the claim presented to the fact verifier. Then, we append our modified claims to articles in the evidence base that are relevant to the original claim. We modify the original claims in two ways.

In the paraphrase approach, we use a state-of-the-art paraphrasing model, PEGASUS (Zhang et al. 2019), to generate paraphrased versions of the original claim (see Table 7 for example). These paraphrases generally retain the meaning of the claim, but often remove contextualizing information that would be found in the context of the article in which the new sentence is inserted. Because the paraphrase method attempts to produce synthetic evidence that is semantically equivalent to the original claim, we test its efficacy relative to a method that merely introduces irrelevant content to the evidence document. Motivated by Jia and Liang (2017), we alter a claim by applying heuristics such as number alteration, antonym substitution, and entity replacement with close neighbors according to embedding similarity (Bojanowski et al. 2017). These modifications should not confuse humans, but would affect sensitive fact verification systems, providing a competitive baseline for assessing the strength of ADVMod-paraphrase.

Finally, our oracle reports the performance when the claim itself is appended to an evidence document.
were more robust against claims that should be worth opposing online (leading to less directly adversarial
Adding synthetic content to the evidence bases of fact verification, it will also be a liability when previously learned
Table 6: Effect of ADVMOD on FEVER claim verification. We bold the largest performance drop relative to the original evidence.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>CorefBERT Acc.</th>
<th>KGAT Acc.</th>
<th>MLA Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Ye et al. 2020)</td>
<td>(Liu et al. 2020b)</td>
<td>(Kruengkrai et al. 2021)</td>
</tr>
<tr>
<td></td>
<td>Total REF NEI</td>
<td>Total REF NEI</td>
<td>Total REF NEI</td>
</tr>
<tr>
<td>Original</td>
<td>73.05 74.03 72.07</td>
<td>70.76 72.50 69.01</td>
<td>75.92 78.71 73.13</td>
</tr>
<tr>
<td>ADVMOD-KeyReplace</td>
<td>53.82 66.50 41.15</td>
<td>42.82 68.90 16.74</td>
<td>60.93 81.83 40.02</td>
</tr>
<tr>
<td>ADVMOD-paraphrase</td>
<td>32.62 36.66 28.58</td>
<td>37.22 51.74 22.70</td>
<td>52.72 70.57 34.86</td>
</tr>
<tr>
<td>ADVMOD-oracle</td>
<td>4.78 7.94 1.61</td>
<td>11.90 23.78 0.02</td>
<td>25.17 45.96 4.37</td>
</tr>
</tbody>
</table>

Table 7: Sample ADVMOD sentences

| Original          | Damon Albarn’s debut album was released in 2011. |
| Paraphrase        | Albarn’s first album was released in 2011. |
| KeyReplace        | Matt Coldplay’s debut album was released in 202. |
|                   | Stefan Blur’s debut album was released in 1822. |

5 Discussion

Adding synthetic content to the evidence bases of fact verifiers significantly decreases their performance. Below, we discuss interesting findings and limitations of our study.

Synthetic vs. human disinformation As mentioned in Section 3, the performance of our test systems is more sensitive to poisoned evidence generated from GROVER than retrieved from MediaCloud, even as the number of documents retrieved from MediaCloud far exceeds the number generated from GROVER. While FEVER claims may not generally be worth opposing online (leading to less directly adversarial content being retrieved from MediaCloud), we note that language models have no such limitations, and can generate large quantities of disinformation about any topic. Consequently, while misinformation already makes its way into retrieval search results (Bush and Zaheer 2019), language models could cheaply skew the distribution of content more drastically (Bommasani et al. 2021), particularly on topics that receive less mainstream coverage, but may be of import to a malicious actor (Starbird et al. 2018).

Language models as a defense In the ADVADD and ADVMOD oracles settings, all tested systems performed better on claims labeled REF than for claims labeled NEI. This result implies that the GROVER-generated evidence was less adversarial for these claims, or that the pretrained models which these systems use to encode the claim and evidence sentences were more robust against claims that should be refuted. Consequently, we conclude that language models encode priors about the veracity of claims, likely from the knowledge they learn about entities during pretraining (Petroni et al. 2019), a conclusion also supported by contemporaneous work in using standalone language models as fact-checkers (Lee et al. 2020, 2021). While this property can be an advantage in some settings (i.e., language models pretrained on reliable text repositories will be natural defenses against textual misinformation), it will also be a liability when previously learned erroneous knowledge will counteract input evidence that contradicts it. Finally, we note that the presence of implicit knowledge in language models affecting the interpretation of input evidence implies that the training corpora of these LMs could be attacked to influence downstream fact verification. Prior work has explored poisoning task-specific training datasets (Wallace et al. 2021). As disinformation becomes more prevalent online, the pretraining corpora of LMs will require more careful curation to avoid learning adversarial content.

Limitations We identify three main limitations to our study. First, the FEVER document retrievers use the MediaWiki API to collect relevant Wikipedia articles based on entity mentions in the claim. We assume our synthetic content could be included in the retrieved documents if it were titled with a mention of the named entities in the claim. For SCIFACT, this limitation is not present because synthetic abstracts are retrieved using statistical IR methods. Second, our method ADVADD uses the actual claim to generate the synthetic article. In the absence of explicit coordination, synthetic poisoned evidence would be generated without knowledge of the exact claim formulation, reducing the realistic correspondence between the claim and the synthetic disinformation. If the GROVER model directly copied the claim during generation, performance drops might be overestimated based on unrealistically aligned evidence. For the ADVADD-full setting, we measure that this issue arises in ~20% of claims, which are predicted incorrectly more often, but does not affect the conclusions of our study. Finally, our FEVER and COVIDFACT studies are run using only claims labeled as REF and NEI, which we discuss in more detail in the Appendix.

6 Conclusion

In this work, we evaluate the robustness of fact verification models when we poison the evidence documents they use to verify claims. We develop two poisoning strategies motivated by real world capabilities: ADVADD, where synthetic documents are added to the evidence set, and ADVMOD, where synthetic sentences are added to existing documents in the evidence set. Our results show that these strategies significantly decrease claim verification accuracy. While these results are troubling, we are optimistic that improvements in automated synthetic content detection, particularly by online platforms with considerable resources, combined with human audits of fact-checker evidence (and their source), may potentially mitigate a considerable number of disinformation campaigns.


Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2020b. Fine-grained Fact Verification with Kernel Graph Attention Network. In Proceedings of ACL.


Thorne, J.; and Vlachos, A. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In COLING.


