The Limitations of Stylometry for Detecting Machine-Generated Fake News

Tal Schuster
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
tals@csail.mit.edu

Roei Schuster
Computer Science Department Tel Aviv University and Computer Science Department Cornell Tech
rs864@cornell.edu

Darsh J. Shah
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
darsh@csail.mit.edu

Regina Barzilay
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
regina@csail.mit.edu

Recent developments in neural language models (LMs) have raised concerns about their potential misuse for automatically spreading misinformation. In light of these concerns, several studies have proposed to detect machine-generated fake news by capturing their stylistic differences from human-written text. These approaches, broadly termed stylometry, have found success in source attribution and misinformation detection in human-written texts. However, in this work, we show that stylometry is limited against machine-generated misinformation. Whereas humans speak differently when trying to deceive, LMs generate stylistically consistent text, regardless of underlying motive. Thus, though stylometry can successfully prevent impersonation by identifying text provenance, it fails to distinguish legitimate LM applications from those that introduce false information. We create two benchmarks demonstrating the stylistic similarity between malicious and legitimate uses of LMs, utilized in auto-completion and editing-assistance settings.[1] *Our findings highlight the need for non-stylometry approaches in detecting machine-generated misinformation, and open up the discussion on the desired evaluation benchmarks.*

Submission received: 26 August 2019; revised version received: 5 January 2020; accepted for publication: 15 February 2020

1 Data: https://people.csail.mit.edu/tals/publication/are_we_safe/.

https://doi.org/10.1162/COLI_a_00380

1. Introduction

Many previous studies on **stylometry**—the extraction of stylistic features from written text—showed promising results on text classification. Two of stylometry's common applications are: (1) *Detecting the provenance* of text (i.e., identifying the author) in order to prevent impersonations (Tweedie, Singh, and Holmes 1996; Brennan, Afroz, and Greenstadt 2012; Afroz et al. 2014; Caliskan-Islam et al. 2015; Neal et al. 2017; Sari, Vlachos, and Stevenson 2017); and (2) *Detecting misinformation* in text due to deception (Enos et al. 2007; Mihalcea and Strapparava 2009; Ott et al. 2011; Feng, Banerjee, and Choi 2012; Afroz, Brennan, and Greenstadt 2012), fake news (Rashkin et al. 2017; Pérez-Rosas et al. 2018), or other false or illegal content (Choshen et al. 2019). In the former, the classifier identifies language features that correlate with a specific person or group. The latter, misinformation detection, relies on idiosyncrasies of lies, namely, style and language characteristics that are unique to text that is false or misleading.

Stylometry has recently gained attention as a potential answer to concerns that language models (LMs) could be used to mass-produce malicious text (Vosoughi, Roy, and Aral 2018; Radford et al. 2019), that (1) impersonates a human author's text and/or (2) is fallacious and misleading. Indeed, stylometry-based approaches have shown promising results for defending against human-impersonating language models (LMs) (Bakhtin et al. 2019; Zellers et al. 2019). However, as applications of text generation such as text auto-completion (Wolf et al. 2019; House 2019) and automatic question answering[2] become widely used, labeling text as generated by an LM might not indicate anything at all about its trustworthiness. This motivates our core inquiry subject: *Can stylometry be used to distinguish malicious uses of language models (LMs) from legitimate ones?*

We build the first benchmark for detection of LM-produced fake news that labels text as "real" or "fake" according to its veracity. Inspired by studies on deceitful behaviors in humans, showing that people try to diverge as little as possible from the truth when lying (Mazar, Amir, and Ariely 2008), we focus on automatic false modifications or additions to otherwise truthful news stories.

Our data sets contain articles produced by both malicious and responsible uses of language models, and the detector's task is to identify the malicious ones. In one data set, we produce text by prompting an LM to extend news articles with relevant claims. We simulate malicious user, who only accepts the LM's suggestion if the claim is factually false, and a responsible user, who only accepts correct claims. The produced sentences are short and concise statements, similarly to fake news and false claims as represented in human-generated data sets (Wang 2017; Augenstein et al. 2019). In another data set, we modify existing news articles to include false information by inverting article statements. In this case, the LM is used to automatically identify the most plausible edit locations. This is similar to (mis-)using an autocorrect tool that suggests local modifications.

We find that with the state-of-the-art stylometry-based classifier, even a single auto-generated sentence within a wall of human-written text is detectable with high accuracy, yet the *truthfulness* of a single sentence remains largely undecidable. Moreover, even a relatively weak LM can be used to produce statement inversions that the state-of-the-art stylometry-based model cannot detect. Thus, *stylometry fails to distinguish malicious from responsible behaviors*. This indicates that, unlike humans who expose stylistic cues when writing false content (Ott et al. 2011; Frank, Menasco, and O'Sullivan 2008; Matsumoto

2 https://github.com/re-search/DocProduct; https://openai.com/blog/openai-api/.

et al. 2011), LMs maintain consistency for both true and false content. Worse yet, while a provenance classifier can effectively detect a potentially malicious author or publication venue (given enough examples), it might not distinguish malicious from legitimate authors if they are both using the same LM to generate their text. In this regard, *malicious text generated by an LM might actually be harder to detect than hand-crafted malicious text*.

Our human evaluation tests show that humans are also fooled by machine-generated misinformation, but that access to external information sources can help. Therefore, we recommend future research on machine-generated misinformation to focus on non-stylometry strategies. Finally, we discuss what benchmarks are required for evaluating the performance of such detectors.

2. Background and Related Work

Stylometry for Human-Written Text. The use of statistical methods for analyzing human-written documents has been studied extensively since the early days of the field. One common application is provenance detection. For example, Mosteller and Wallace (1963) used word counts to predict the authorship of historical documents. Tweedie, Singh, and Holmes (1996) extracted other stylistic features and applied a neural network to the same task. Although these classifiers could be fooled by an aware writer that intentionally imitates other's style (Brennan, Afroz, and Greenstadt 2012), this approach was found useful for de-anonymizing cybercriminals in forums (Afroz et al. 2014), identifying programmers (Caliskan-Islam et al. 2015), and more (Neal et al. 2017). In a related line of study, stylometry was applied to, rather than detecting the specific person, identifying characteristics of the author, such as age and gender (Goswami, Sarkar, and Rustagi 2009), political views (Potthast et al. 2018), or IQ (Abramov and Yampolskiy 2019). Recently, as detailed later in this article, stylometry was used to distinguish machine- from human-writers.

Another common application of stylometry is detecting human-written misinformation. Mihalcea and Strapparava (2009) found specific words that are highly correlated with true and false statements. Ott et al. (2011) and Feng, Banerjee, and Choi (2012) used a richer set of features such as POS tag frequencies and constituency structure to identify deceptive writing. Following these observations and the increasing interest in fake news, recent studies applied stylometry on entire news articles (Pisarevskaya 2017), short news reports (Pérez-Rosas et al. 2018), fact and political statements (Nakashole and Mitchell 2014; Rashkin et al. 2017), and posts in social media (Volkova et al. 2017). The success of these studies is mostly attributed to stylistic changes in human language when lying or deceiving (Bond and Lee 2005; Frank, Menasco, and O'Sullivan 2008). In this work, we evaluate the viability of this approach on machine-generated text, where stylistic differences between truth and lie might be more subtle.

Machine-Generated Text Detection. Detecting text's provenance is similar to authorship attribution and, therefore, stylometry can be effective. Indeed, Gehrmann, Strobelt, and Rush (2019) show the existence of distributional differences between human-written texts and machine-generated ones by visualizing the per-token probability according to an LM. Bakhtin et al. (2019) learn a dedicated provenance neural classifier. Although their classifier achieves high in-domain accuracy, they find that it overfits the generated text distribution rather than detecting outliers from human texts, resulting in increased "human-ness" scores for random perturbations. Nevertheless, an advantage of such neural approaches over more traditional stylometry methods is that, given enough data,

the model learns hidden stylistic representations without the need to manually define any features.

Building on this observation, Zellers et al. (2019) focus on fake news and create a Transformer-based LM (Vaswani et al. 2017) dubbed Grover and train it on a large news corpus. Grover also includes a "neural fake news detector," a linear classifier on top of the hidden state of the last token of the examined article, fine-tuned to classify whether the news text was machine-generated or not. The experiments in this article are based on the Grover-Mega classifier, fine-tuned for the target task (see Section 3).

Fake News Detection Approaches Beyond Stylometry. The other most extensively studied NLP-based approach for fake news detection is based on fact-checking. This approach has recently gained increasing attention thanks to several synthetic (Thorne et al. 2018) and real-world data sets (Hanselowski et al. 2018; Wang 2017; Popat et al. 2018; Augenstein et al. 2019). The performance of current models is still far from that of humans (Thorne et al. 2019; Schuster et al. 2019), but with their advancements they can still play a positive role in detection.

Another line of work for fake news detection utilizes non-textual information such as how content is propagated, by which users, its originating URL, and other meta-data (Castillo, Mendoza, and Poblete 2011; Gupta et al. 2014; Zhao, Resnick, and Mei 2015; Kochkina, Liakata, and Zubiaga 2018; Liu and Wu 2018), as well as incorporation of users' explicit feedback, such as abuse reporting (Tschiatschek et al. 2018). Social network platforms, ISPs, and even individual users can use such methods to moderate content exposure. These approaches are beset by the challenges (Shu et al. 2017) of noisy and incomplete data, especially given the need to detect fake news early (Liu and Wu 2018) (before propagation and user engagement patterns are fully formed).

3. Adversarial Setting

"Fake News": Our Working Definition. Our attackers focus on automatically introducing false information into otherwise trustworthy content. We call the resultant manipulated articles **fake news**. This definition matches that of Zhou and Zafarani (2018) and is in line with the disinformation focused view of Wardle and Derakhshan (2017). Also, this is in line with how false claims are represented in many human-generated fake news data sets (Wang 2017; Augenstein et al. 2019). Conversely, Zellers et al. (2019) focus on entirely fabricated articles, a different type of fake news where the goal is mostly to create "viral and persuasive" content.

Our choice of creating articles with only a limited number of false statements is aligned with the way humans tend to deceive or lie. Psychological studies (Mazar, Amir, and Ariely 2008) support the age-old notion that, when lying, "the best policy for the criminal is to tell the truth as nearly as possible."[3] This helps preserve an honest self-image and, perhaps more importantly, reduces the chances of the lie being detected. For example, a study on the longest-surviving known fake Wikipedia article (Benjakob 2019)[4] revealed that many of the presented facts were only slightly altered from other, true facts.

3 Raskolnikov, "Crime and Punishment" (Dostoyevsky 1866).

4 https://en.wikipedia.org/wiki/Wikipedia:List_of_hoaxes_on_Wikipedia.

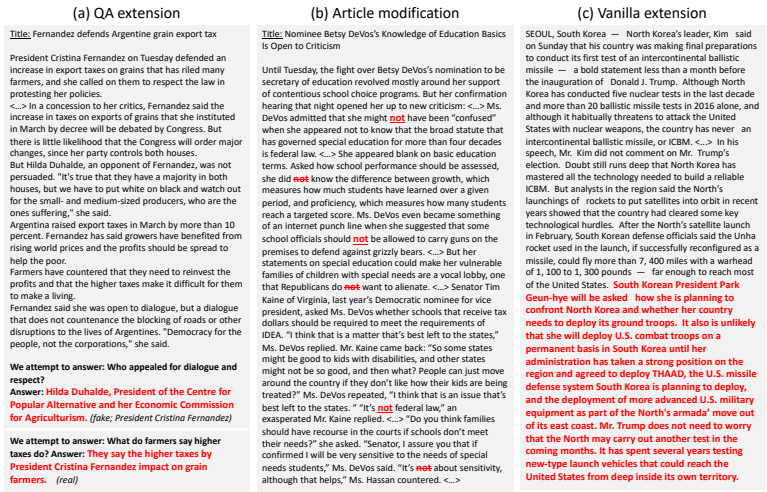

Figure 1
Examples of the fake class in our experiments. (a) In the news question answering (Section 4), a CNN article is presented with two examples of questions (bold) from newsQA (Trischler et al. 2017) and Grover's generated answer (red). The first answer is verified by a human annotator to be false 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 prefix.

Attack and Defense Capabilities. 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 fake text that the verifier classifies as real (see Figure 1 for examples). 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.

Training and Evaluating the Detector. In each experiment, we collected a data set 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 model's weights 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 defense, we applied a pretrained Grover-Mega discriminator by querying its Web interface. We report human performance on some of the attacks.

4. Stylometry Fails to Detect Machine-Generated Misinformation

We create two data sets, simulating two different uses of LMs to automatically produce fake news. In the first, the *extension scenario*, an auto-completion text generator extends a news article. A responsible user of this generator verifies the correctness of the output (producing real text), whereas an attacker verifies *in*correctness (producing fake text).

In the second, the *modification scenario*, the attacker uses a human-written news article and performs subtle modifications to semantically modify statements. Specifically, we add and remove negations. This follows the intuition that, if we take care to add negations in a syntactically correct fashion, the new sentence is a negative inversion of the original (Rudanko 1982). Yet, such changes are subtle enough to retain the original article's style and distributional features. See below for the full details on the creation of the data sets.

Additionally, we used about 100 examples from each data set to test human performance in detecting this form of misinformation. For the extension scenario, we assigned two subject volunteers with two different tasks. The first had the same task as Grover's detector, and the second was allowed to use external sources to verify facts (which Grover cannot do). For the modification scenario, the participants had the same task as Grover but we highlighted negations to help them focus on the relevant parts.

(1) Creating the Extension Data Set. We use the newsQA data set (Trischler et al. 2017) that contains CNN articles with corresponding questions and answers (segments copied verbatim from the text). Following Radford et al. (2019) and von Davier (2019), we embed the question in a template that is appended to the end of the input text:

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

Then, the first sentence produced by Grover's generator is used as a potential answer. We choose this template by examining common formats of questions in news articles and finding the one that empirically produces mostly reasonable answers by Grover's generator. For example, given text about a suspect caught in Washington, DC, and the question *"Where was the suspect stopped?"*, Grover answered *"2 blocks from the U.S. Capitol."*

We manually labeled the generated answers. We filtered out nonsensical ones (29%) and labeled the rest as real or fake texts by correctness. Answers containing at least one false statement (judging by the article and/or external knowledge) were labeled fake. We measured inter-annotator agreement on a subset of 100 examples labeled by two different annotators, and found it to be substantial (Cohen's kappa score of $k = 0.78$).

Finally, we remove the sentence containing the answer from each article by removing the sentence with the highest TF-IDF-weighted word-count similarity with the question and answer (concatenated).

(2) Creating the Modifications Data Set. We use articles from *The New York Times* (Thompson 2017). We perform the modifications by removing and adding negations from statements: first, we randomly delete $m/2$ "not" or "no" occurrences from the text. Then, we automatically find statements to add $m/2$ negations to, using the probabilities of a GPT-2 Medium language model. 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. By both adding and removing, we keep the total number of negations unchanged, avoiding anomalous negations counts. Original *New York Times* articles are labeled as real and modified ones as fake.

Table 1
Results: Section 4. We report (macro) F1 score and overall accuracy, as well as precision and recall of the "fake" class. Zero-shot performance (not included) was very low in all cases.

Machine-generated misinformation		precision	recall	F1	accuracy
	QA extension (false vs. true)	0.72	0.71	0.71	71%
adaptive	modification ($m = 2$)	0.53	0.52	0.53	53%
	modification ($m = 6$)	0.66	0.65	0.65	65%
	modification ($m = 10$)	0.73	0.47	0.63	65%

4.1 Results

The detector's performance figures are given in Table 1. Against QA extension, the detector reached 71% accuracy in distinguishing true vs. false extensions generated by Grover. Although this is above a trivial majority baseline (51%), some of the success could be a side effect of the way that the data set was constructed. Labeling instances with at least one false statement as fake makes these sentences generally longer, allowing a simple linear classifier that only uses the extension's length to reach 56%. When evaluated only on the short false answers (up to 10 words), the accuracy drops from 71% to only 62%. Ultimately, however, despite performing better than simple baselines, the detection in this setting is much worse than in distinguishing human from machine text (Section 5).

Comparing this with the results of the human evaluation is particularly instructive: The first subject, given the same task as the detector, attained an F1 score of 0.68, only marginally below the detector. The second participant, who was allowed to use external sources, attained an F1 of 0.84. We conclude that the stylometry-based classifier does not perform much better than humans in detecting potential misinformation, and that verifying against other resources can drastically improve results.

For the automatic article modification attack, the classifier fails completely to detect two modifications ($m = 2$), and is not much better even when $m = 10$ statements are inverted. Even though the change includes negations that might introduce surprising factual claims, the language style remains too similar to detect it. Humans, evaluated with $m = 2$, had an F1 of 0.74, with 0.91 and 0.59 recall for real and modified articles, respectively. Thus, although humans perform better than the model, many of the modified statements were interpreted as real. Manually examining these, we found that the added negation actually changed the meaning in 60% of the instances that were misclassified.

Overall, the results show that both attacks can produce articles containing misinformation that evade the stylometry detector and mislead humans.

5. Stylometry Detects Machine–Human Impersonations

One might suspect that the low performance in Section 4 is due to limited capacity of our detector or the small fraction of generated text. We now show that the detector performs well on the provenance task and is sensitive to differences in small text portions.

(1) Fully Generated Articles. We perform a first evaluation of Grover against texts generated by a different model of similar size, namely, GPT-2 XL (Radford et al. 2019). We include examples from the WebText test set (Radford et al. 2019), labeled real text, and the released unconditioned generations of the GPT-2 XL model, labeled fake text.

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.

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

(2) Automatic Article Extension (Vanilla). 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 first 500 words from an article. We include *New York Times* articles, truncated to match the length of the fake class, labeled as real text, and articles that were automatically extended, labeled fake text. We use different percentages (g) of machine-generated text, by stopping the generation after an appropriate number of sentences.

(3) Automatic QA-based Extension. In this experiment, we test whether the QA defense from Section 4 can perform better if we can assume that true texts are always human-written. To this end, we repeat the same setting but use the gold answers from the newsQA data set for the "real text" class.

5.1 Results

Zero-shot Setting. As Table 2 shows, the zero-shot classifier is effective in detecting the fully generated articles of a different model, with a 0.9 F-1 score.

The extended articles, containing a substantial amount of human-written text, are mostly classified as human-written in this setting. This is unsurprising considering the data set of full articles the detector was originally trained on.

Adaptive Setting. After fine-tuning, Grover improves on the full articles and performs well against article extension generations, reaching 0.94 F1 score for articles with a single generated sentence ($g = 1\%$). This stylometry detector is thus highly effective in distinguishing human from machine.

In the QA setting, also containing a single machine-generated sentence, the detector performs worse by 0.11 F1 points. This might be due to using the (stronger) Grover-Mega generator for the QA setting, and because the appended template for QA was selected to empirically maximize human "reasonableness" scores. Still, this setting allows the stylometry detector to reach much higher performance than in the veracity-based setting (Section 4). This indicates that a restricted benchmark, which assumes no legitimate use of LMs, might not reflect the misinformation detection performance if the model is deployed in a world where LMs are used for both legitimate and malicious purposes.

6. Discussion

Advancements in LM technology and their various applications have introduced a new challenge: distinguishing truthful text from misinformation, when the text is generated or edited by an LM. Our experiments indicate that LM-generated falsified texts are very similar in style to LM-generated texts containing true content. As a result, stylometry-based classifiers cannot identify auto-generated intentionally misleading content.

We conclude with the following recommendations:

(1) *Extending Veracity-based Benchmarks.* In order to better evaluate detectors against LM-generated misinformation, we recommend extending our benchmarks by creating other veracity-oriented data sets, that represent a wide range of LM applications, from whole-article generation to forms of hybrid writing and editing.

(2) *Improving Non-stylometry Methods.* Other detection approaches, as surveyed at the end of Section 2, are less affected by the use of LMs. Therefore, advancements in such methods can improve the detection of both human- and machine-generated misinformation. Notably, the fact-checking setting makes fewer assumptions on the available auxiliary information and can be applied even if the text was sent to the verifier through a private channel such as e-mail. However, because fact-checking requires advanced inference capabilities, incorporating non-textual information, when available, can yield better results.

7. Conclusion

The potential use of LMs in creating fake news calls for a re-evaluation of current defense strategies. We examine the state-of-the-art stylometry model, and find it effective in preventing impersonation, but limited in detecting LM-generated misinformation. This new kind of misinformation could be created by the same model that is used by legitimate writers as a writing-assistance tool, hiding stylistic differences between falsified and truthful content. This motivates (1) constructing more instructive benchmarks for NLP-based approaches and improving non-stylistic methods, and (2) addressing a set of challenges that spans many disciplines beyond NLP, including social networks, information security, human-computer interaction, and others.

Acknowledgments

We thank the anonymous reviewers and the members of the MIT NLP group for their helpful comments. R.S. is a member of the Check Point Institute of Information Technology. This work is supported in part by Google's TensorFlow Research Cloud program; 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

Abramov, Polina Shafran and Roman V. Yampolskiy. 2019. In Automatic IQ estimation using stylometric methods. *In Handbook of Research on Learning in the Age of Transhumanism*. IGI Global, pages 32–45, Hershey, PA.

Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of 2012 IEEE Symposium on Security and Privacy*. pages 461–475, San Francisco, CA, USA.

Afroz, Sadia, Aylin Caliskan, Rachel Greenstadt, and Dan Mcoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of 2014 IEEE Symposium on Security and Privacy*, pages 212–226, IEEE Computer Society, San Jose, CA, USA.

Augenstein, Isabelle, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong.

Bakhtin, Anton, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? Learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Benjakob, Omer. 2019. The fake Nazi death camp: Wikipedia's longest hoax, exposed. `https://tinyurl.com/y3m2wd9k`. Haaretz.

Bond, Gary D. and Adrienne Y. Lee. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19(3):313–329.

Brennan, Michael, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):12.

Caliskan-Islam, Aylin, Richard Harang, Andrew Liu, Arvind Narayanan, Clare Voss, Fabian Yamaguchi, and Rachel Greenstadt. 2015. De-anonymizing programmers via code stylometry. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 255–270, Washington, D.C.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 67–684.

Choshen, Leshem, Dan Eldad, Daniel Hershcovich, Elior Sulem, and Omri Abend. 2019. The language of legal and illegal activity on the Darknet. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4271–4279, Florence.

von Davier, Matthias. 2019. Training optimus prime, M.D.: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *arXiv preprint arXiv:1908.08594*.

Dostoyevsky, Fyodor. 1866. *Crime and Punishment*. The Russian Messenger.

Enos, Frank, Elizabeth Shriberg, Martin Graciarena, Julia Hirschberg, and Andreas Stolcke. 2007. Detecting deception using critical segments. In *International Speech Communication Association*, pages 2281–2284, Antwerp, Belgium.

Feng, Song, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–175.

Frank, Mark G., Melissa A. Menasco, and Maureen O'Sullivan. 2008. Human behavior and deception detection. In John G. Voeller, editor, *Wiley Handbook of Science and Technology for Homeland Security*. Wiley & Sons, Inc., pages 1–13.

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

Goswami, Sumit, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *International AAAI Conference on Web and Social Media*, San Jose, CA.

Gupta, Aditi, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. TweetCred: Real-time credibility assessment of content on Twitter. In *International Conference on Social Informatics*, pages 228–243, Cham.

Hanselowski, Andreas, P. V. S. Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, NM.

House, Patrick. 2019. I, language robot. `https://lareviewofbooks.org/article/i-language-robot/`. Los Angeles Review of Books.

Kochkina, Elena, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In

Proceedings of the 27th International Conference on Computational Linguistics, pages 3402–3413, Santa Fe, NM.

Liu, Yang and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *AAAI Conference on Artificial Intelligence*, pages 354–361, New Orleans, LA.

Matsumoto, David,, Hyi Sung Hwang, Lisa Skinner, and Mark Frank. 2011. Evaluating truthfulness and detecting deception. *FBI Law Enforcement Bulletin*, (80):1.

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

Mihalcea, Rada and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore.

Mosteller, Frederick and David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Nakashole, Ndapandula and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, MD.

Neal, Tempestt J., Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon L. Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50:86:1–86:36.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, OR.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, NM.

Pisarevskaya, Dina. 2017. Deception detection in news reports in the Russian language: Lexics and discourse. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 74–79, Copenhagen.

Popat, Kashyap, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels.

Potthast, Martin, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8). `https://openai.com/blog/better-language-models/`.

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

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

Sari, Yunita, Andreas Vlachos, and Mark Stevenson. 2017. Continuous n-gram representations for authorship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia.

Schuster, Tal, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong.

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

Thompson, Andrew. 2017. All the news dataset. `https://www.kaggle.com/snapcrack/all-the-news`. Kaggle.

Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong.

Thorne, James, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (fever) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels.

Trischler, Adam, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver.

Tschiatschek, Sebastian, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, pages 517–524.

Tweedie, Fiona J., Shreyan Singh, and David I. Holmes. 1996. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, 30:1–10.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Volkova, Svitlana, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653,Vancouver.

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

Wang, William Yang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver.

Wardle, Claire and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. `https://tinyurl.com/sy64l6s`.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*, pages 9054–9065, Curran Associates, Inc.

Zhao, Zhe, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405, Florence.

Zhou, Xinyi and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.