Can Fake News Detection be Accountable? The Adversarial Examples Challenge

(Extended Abstract)

Jérémie Bogaert, Quentin Carbonnelle,
Antonin Descampe, François-Xavier Standaert

UCLouvain, Louvain-la-Neuve, Belgium

Abstract

Automated fake news detection is an important challenge in view of the increasing ability of statistical language models to generate large amounts of (possibly fake) articles, so that recognizing them manually becomes unrealistic. Yet, the reliable deployment of such automated detection tools would require ensuring that they are accountable. Algorithmic accountability is known to be difficult to reach, especially when adversarial behaviors aim to make algorithms deviate from their expected mode of operation. In this paper, we illustrate with a case study that this challenge is further amplified in contexts where the labeling of the articles is prone to errors, which is the case of fake news detection.

1 Introduction

The proliferation of fake news on social media platforms has created an increasing need of solutions to detect them. While the generation of fake news and the necessary fact checking that it implies started as a mostly manual process (e.g., discussed in [17]), recent advances in natural language generation with machine learning algorithms have amplified the risk of so-called neural fake news [19]. The massive amount of articles that such tools can generate makes manual fact checking unrealistic and raises the question of their automated detection. As recently surveyed in [1, 20], solutions combining natural language processing and machine learning are attractive for this purpose, due to their ability to capture various features of newspaper articles. Besides, the sensitive nature of the fake news detection problem also requires that its deployment comes with guarantees of algorithmic accountability [3]. But as discussed in [2], such guarantees are hard to obtain in contexts where adversarial behaviors can target the robustness of machine learning classifiers, which is the case of fake news detection.

In this paper, we therefore study the robustness of fake news detection against adversarial examples [4, 8]. For this purpose, we first build a database of fake and reliable news. As already observed in the literature, this step is inherently challenging since there is no strict definition of what a fake news is [6]. We deal with this difficulty by combining a public dataset of fake news (`https://www.kaggle.com/mrisdal/fake-news`), which were scraped from blacklisted websites over a period of 30 days around the 2016 US election, and reliable news collected from the New York Times and the Guardian over approximately the same period. We additionally explore both data sets to show that they cover similar topics. While inevitably imperfect, this database is then used to show that standard machine learning classifiers can detect fake news with a good accuracy, but are also easy to fool with adversarial examples. Interestingly, it appears that the difficulty to define what a fake news is (possibly combined with label errors in the training sets) gives adversaries additional opportunities to generate fake news classified as reliable. So our results question again the possibility

to rely on accountable algorithms for sensitive tasks that can be targeted by adversarial behaviors and suggest different research avenues related to fake news in general.

We note that the topic of fake news is widely multidisciplinary [7] and even the more technical topic of automated fake news detection is already covered by a broad literature (e.g., the already mentioned [1, 19, 20] but also [5, 11, 13, 12] to name a few). Our contribution is not to improve automated fake news detection algorithms but to illustrate the difficult interplay between such algorithms and the need of algorithmic accountability when adversarial behaviors are considered. As a natural starting point in this direction, we show that there exist (for now simple) examples of fake news detectors that are weak against such adversarial behaviors, raising the question of how to prevent them, with technical and non-technical means. In other words, we put forward an under-discussed risk for the reliable deployment of such systems.

The rest of the paper is structured as follow. We start by describing the database we used for our investigations and discussing its unavoidable limitations in Section 2. We follow by showing simple examples of supervised fake news detection tools that can detect fake news with reasonable accuracy in Section 3. We finally exhibit how to craft adversarial examples against these classifiers in Section 4. We conclude by analyzing the impact of these findings and tracks for further research in Section 5.

2 Building a fake news database

Research on fake news detection has often been limited by the quality of existing datasets and their specific application contexts [12]. As already mentioned, this is mostly due to the difficulty of precisely defining what a fake news is, making their labeling for supervised learning challenging [6]. Besides, our goal to investigate adversarial examples is typically calling for "not too short" texts, excluding popular databases such as [18]. To the best of our knowledge, the database that comes the closest to our needs is the fake news corpus (`https://github.com/several27/FakeNewsCorpus`). However, preliminary investigations suggested quite significant dissimilarities between the topics of reliable and fake articles (namely, football for fake articles and politics for reliable ones). This similarity makes it unsuitable for our purposes, since it implies risks to classify the articles based on their topic more than their fake/reliable nature. We therefore prepared our investigations by attempting to build a better database.

Concretely, we started from a public dataset of 3500 fake news from Kaggle (`https://www.kaggle.com/mrisdal/fake-news`), which were scraped from 207 blacklisted websites over a period of 30 days around the 2016 US election. Some preliminary processing was applied to remove unwanted features of the articles (e.g., meta-data that is not in the original articles and was added by the blacklisted websites). No website contributed to more than 1% of the database to ensure that imperfect scraping, preprocessing or labeling for some websites cannot significantly impact the overall results. We then tried to build a complementary database of reliable news, covering the same period of time. For this purpose, we used the API developed by the New York Times and The Guardian, and scraped papers about World news and US news for the intended period. We had to slightly increase the window of time (Oct. 16 to Dec. 4 for the reliable news vs. Oct. 26 to Nov. 25 for the fake news), in order to collect 3500 articles (1500 from the New York Times, 2000 from The Guardian).

We performed a preliminary exploration of the topics covered by this database using the Term Frequency–Inverse Document Frequency (TF-IDF) statistic. It is a standard tool for information retrieval or summarization, which indicates the relevance of the words in some documents [10]. We first launched it on the full database to identify the most relevant words overall. Next, we launched it on the fake news corpus, leading to the results of Figure 1 (where the X axis lists the 40 most relevant words of the full

database). It confirms the coverage of the US elections (e.g., with Trump and Clinton in the first places), but also highlights the weight of some neutral words like 'said'.

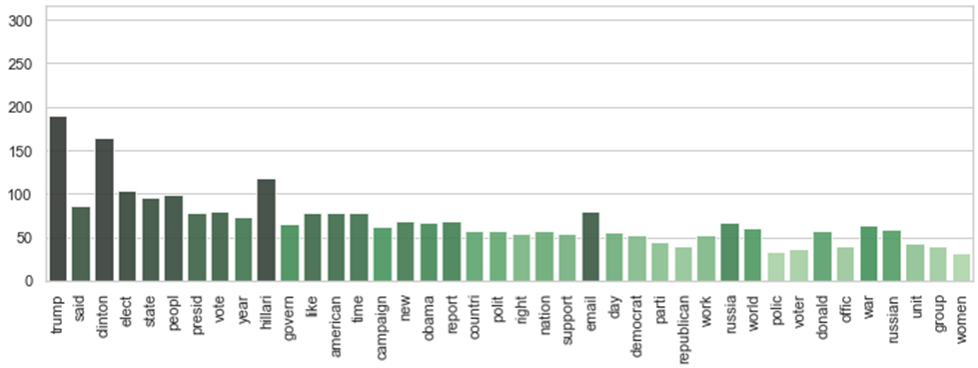

Figure 1: Evaluation of the fake news database's topics with TF-IDF score.

We finally computed the same TF-IDF scores for the reliable news, which are given in Figure 2 and confirm the coverage of similar topics. They also highlight some specific features of the fake news corpus already: for example the more frequent use of the (misspelled) first name Hilari or the importance of the word e-mail (relating to an ongoing affair of Mrs. Clinton's e-mail leaks during the 2016 elections).

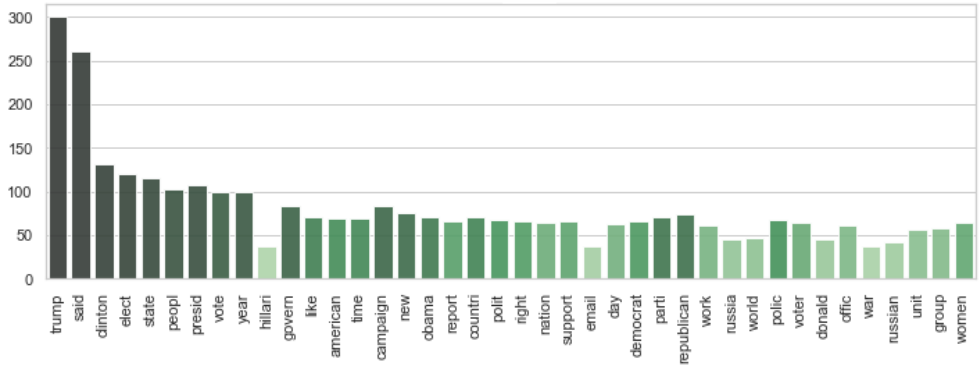

Figure 2: Evaluation of the reliable news database's topics with TF-IDF score.

In order to confirm that the resulting (7000-paper) database did not contain obvious parasitic patterns, we additionally visualized the data by feeding the t-distributed Stochastic Neighbor Embedding (t-SNE) tool of [16] with the vectors output by the TF-IDF transform.* t-SNE projects each high-dimensional object towards a two-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. As illustrated in Figure 3, the distribution of the topics is similar for fake and reliable articles while, for example, the separation between World and US (reliable) news can be distinguished in the right part of the figure. It suggests that the topics do not create obvious (parasitic) ways to discriminate fake and reliable news. So while the evaluations in this section do admittedly not provide any formal guarantee that no such patterns exist, we assume in the following that this database is good enough for our purposes.

* Reduced to 500 dimensions thanks to truncated Singular Value Decomposition (SVD). We also tried larger number of dimensions but it did not change our main observations.

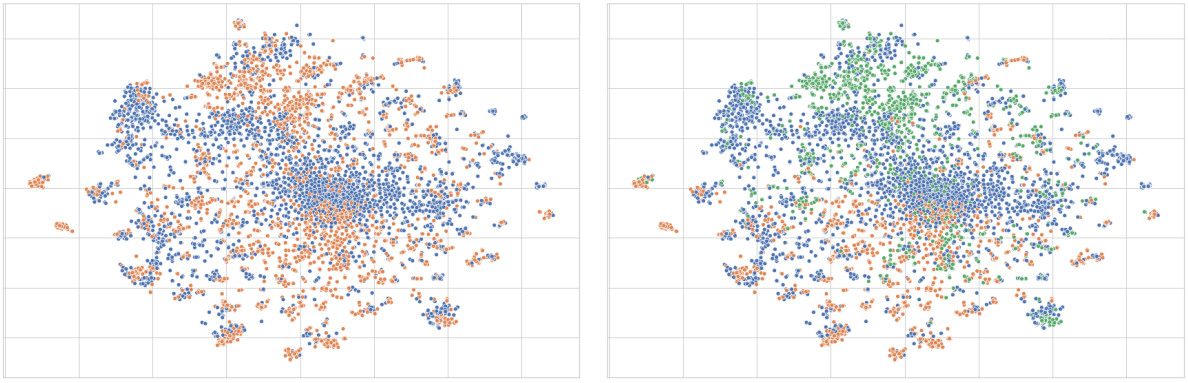

Figure 3: Visualization of the news' topics with t-SNE. Left: fake news (●) and reliable news (●). Right: fake news (●), reliable World news (●) and reliable US news (●).

3 Exemplary classifiers

The next step of our investigations was to build statistical classifiers for fake and reliable news, thanks to supervised machine learning. We considered various options for this purpose: logistic regression, naive Bayes, random forests and Long Short-Term Memory (LSTM). For place constraints, we focus our following descriptions on logistic regression and LSTM (the other classifiers did not significantly affect our main conclusions).

Concretely, all our classifiers were built starting with the same preprocessing: we removed punctuation, non-alphanumeric characters, multiple whites spaces, websites, stop words and short words). For the logistic regression, we then vectorized the articles as bag of words using TF-IDF while for the LSTM we used a Word2Vec model trained on our full dataset [9]. As Google's Word2Vec (`https://code.google.com/archive/p/word2vec/`), we fixed the embedding to 300-dimensional vectors. The rationale behind this choice is that the LSTM can take advantage of the words' order. The model hyperparameters were then tuned using a 5-fold cross-validation.

The accuracy of these classifiers is illustrated in Figure 4. It is significantly better than a random guess in both cases, and reaches $> 90\%$ for the logistic regression (we expect that the LSTM would reach this accuracy or even improve it with more training samples). Despite further optimizations are certainly possible, we assume these values to be sufficient for trying to decrease them with adversarial examples.

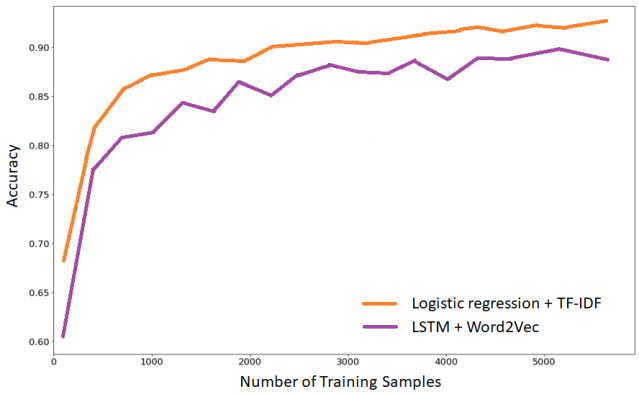

Figure 4: Learning curve of exemplary fake news detectors.

4 Crafting adversarial examples

An adversarial example is an input, unknown to the target machine learning algorithm, that makes it deviate from its public specifications and is purposely generated by an adversary having an interest in such a deviated behavior [4]. In the context of fake news detection, the goal of the adversary is to force the misclassification of a fake news as reliable, despite not changing its semantic content from the human perception viewpoint. The main question we tackle next is whether such adversarial examples can be crafted for the detectors of the previous section. In this context, the adversarial goal could vary from producing such adversarial examples at a low rate, possibly taking advantage of some manual processing, or to produce them at a high rate, automatically. As a first step to show the existence of a risk, we consider the easiest (low rate with manual intervention) context. We briefly discuss the relaxation of this context at the end of the section. The threat model could also vary from white box access to the models (i.e., knowing their parameters) to only black box access (i.e., only being able to observe input/output pairs). We discuss both options in the following.

4.1 Methodology

The high-level approach we used to craft adversarial examples at low rate is in 3 steps:

1. Identify Reliable Hot Words (RHW), which are the words having high probability to push the classification of any article towards the reliable class.

2. Identify the Target's Fake Hot Words (TFHW), which are the words having high probability to push the classification of a target article towards the fake class.

3. Human intervention to replace TFHW by RHW in a semantic-preserving manner.

The identification of the RHW and TFHW was performed using high-level ideas similar to [8, 2]. In the white box setting, we applied the Fast Gradient Sign Method (FGSM) which essentially boils down to maximizing the classifier's loss function, then using the gradient information to add or remove words independent of their position.[†] This gradient can be computed analytically for the logistic regression, and we used the numerical estimation provided by TensorFlow (`https://www.tensorflow.org/`) for the LSTM. In the black box setting, we used a more straightforward approach where we just removed words one by one and feed the classifier with the modified articles in order to identify words that impact the detection probability the most.

4.2 Experimental results

We will discuss the type of results we obtain with Example 1, which is initally detected as a fake news with 90% probability by the logistic regression, and 65% probability by the LSTM. As aforementioned, the first step in trying to fool the detection is to identify RHW and TFHW. RHW are identified on the reliable news corpus while TFHW are identified on the target article only. We illustrate this step with our black box approach applied to the beginning of the example (namely, the words *In what is being described as another 'bizarre'*), where the short words 'In', 'what', 'is', 'being' and 'as' do not impact the classification, while removing the words 'described' and 'bizarre' respectively decrease its probability of being detected as a fake news by 2.8% and 3.5%. By extending such an approach to the whole article (and the whole corpus of reliable news for RHW), we can then build lists of TFHW and RHW.

[†] This position could be exploited in the case of the LSTM, which we leave as a scope for further investigations (a direct exploitation would result in a quite computationally intensive strategy).

In what is being described as another 'bizarre' attempt to sabotage her own campaign, Hillary Clinton has desecrated a series of beloved US symbols, including punching a bison, setting fire to the Stars & Stripes and spitting at Jerry Seinfield. The Presidential hopeful seems determined to make a series of unprovoked errors, not least of which was agreeing to Bill hosting a sleepover for a group of Girl Guides. Short of dressing the Statue of Liberty in a Burka, Mrs Clinton has lurched from one PR blunder to another. Commented one journalist: 'The Presidential race is entering the final furlong and if Mrs Clinton was horse – and before you can say Benghazi – she's gone from bookie's favourite to an ingredient at the local glue factory'. Having already become the unwitting focus of various health scares and FBI investigations, Mrs Clinton's campaign is as orderly as a Marx Brothers movie. Her lead in the polls has been cut as video emerges of her lighting a cigar with a rolled up Bill of Rights, then proceeding to take a dump on the White House lawn. Hillary's erratic behaviour has seen her sing the Star-Spangled Banner in Korean, dress as Oprah Winfrey for Halloween and pebble-dash Mount Rushmore. Remarked a flummoxed advisor: 'She keeps doing the unthinkable – like making Donald Trump electable'. Share this story...

Example 1: Sample of the fake news database.

The main high-level observations that can be extracted from these lists are:

1. That they contain both semantically tainted words (e.g., 'Hilary', 'FBI', 'Donald') and semantically neutral words (e.g., 'said', 'commented', 'campaign').
2. That there is a higher proportion of semantically tainted words for TFHW.
3. That the lists of (ordered) words identified as RHW or TFHW for the logistic regression and the LSTM, significantly overlap but are not identical.

We can then build adversarial examples, e.g., for the (simpler) LSTM:

[...] 'bizarre' attempt to sabotage her own campaign, ~~Hillary~~ **Mrs** Clinton has desecrated
[...] Mrs Clinton's campaign is as ~~orderly~~ **neat** as a Marx Brothers movie [...]

By changing only two words (which do not affect the meaning of the article), it is now classified as a fake with only 45% probability (so as reliable with 55% probability). Interestingly, exactly those changes are not sufficient to misclassify the article with the logistic regression (which still classifies it as a fake, but with a probability reduced from 90% to 55%). Yet, another change of the second word does the job (suggesting that as usual with adversarial examples, they have a certain level of transferability):

[...] 'bizarre' attempt to sabotage her own campaign, ~~Hillary~~ **Mrs** Clinton has desecrated
[...] Remarked a flummoxed ~~advisor~~ **minister**: 'She keeps doing the unthinkable [...]

Since based on manual interventions, we did not craft such examples for large number of articles. We repeated the process for 5 fake news and succeeded to force a misclassification by changing a maximum of 15 words. The LSTM classifier was usually fooled with slightly less words which we assume is due to its initially lower accuracy.

Overall, and despite drawing general conclusions based on such small-scale examples is hard, one important observation is that in contrast with the case study of [2] where adversarial examples had to be mostly based on neutral words (to avoid being easy to spot by the human perception), fooling a fake news detection algorithm can be done with more tainted words (e.g., by changing 'Hillary' into 'Mrs' in our example). In other words, the fact that the fake or reliable nature of an article does not have a definition as clear as the topics used in the case study [2] makes the adversary's task easier.

4.3 Towards automation

Our handmade experiments naturally raise the question whether adversarial examples can be automated. As a first step in this direction (i.e., to evaluate the sensitivity of

our fake news detectors to semantically-neutral modifications), we evaluated a greedy strategy where we just substituted words by synonyms (sometimes causing syntax problems). One example obtained for the logistic regression is given below:

> [...] symbols, including punching a ~~bison~~ **buffalo** [...] [...] a group of ~~Girls~~ **Woman** guides
> [...] various health scares and FBI ~~investigations~~ **inquiry**, Mrs Clinton's campaign [...]

Out of 100 test articles, we could misclassify 22% (resp., 32%) with the LSTM (resp., logistic regression) classifier (presumably because our greedy strategy is better suited to models that do not exploit the words' order). sing advanced statistical language models should lead to more adversarial opportunities, which we leave as an open problem.

5 Conclusions and open problems

Advances in digital media are responsible for journalists to loose the monopoly on information production and dissemination, and raise new legitimacy concerns (e.g., regarding whether journalism can still offer quality and reliable news in the digital era) [15]. Algorithmic accountability is one of the emerging (and difficult to reach) goals aiming to mitigate such concerns [3]. In this work, we confirm that ensuring algorithmic accountability with robustness against adversarial behaviors is especially challenging in contexts such as fake news detection where the definition of optimization criteria is inherently fuzzy due to the lack of well defined classes. Our results are preliminary in many respects and therefore suggest various directions for further investigations. First, the difficult collection of a fake news / reliable news database would be improved by designing a tool able to automatically generate large amounts of fake news from reliable ones (e.g., by biasing them in a given direction). It would allow a better analysis of the syntactic or semantic patterns that fake news detection can exploit. Second, and as already mentioned, the generation of adversarial examples would benefit from automation in order to enable larger-scale experiments. Third, the evaluation of countermeasures (i.e., fake news detection tools able to cope with adversarial examples) is an important long-term goal as well. Early discussions in [4, 2] suggest that a purely technical solution may not be possible. The perspective of such negative conclusions therefore questions the need of complementary approaches, e.g., relying on the trust in the journalists who write stories more than on content, as suggested in [14].

References

[1] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4, 2015.

[2] Antonin Descampe, Clément Massart, Simon Poelman, François-Xavier Standaert, and Olivier Standaert. Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. AI & Society, 2021.

[3] Nicholas Diakopoulos. Algorithmic accountability. Digital Journalism, 3(3):398–415, 2015.

[4] Ian J. Goodfellow, Patrick D. McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. Commun. ACM, 61(7):56–66, 2018.

[5] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In KDD, pages 1803–1812. ACM, 2017.

[6] Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. Defining "fake news". Digital Journalism, 6(2):137–153, 2018.

[7] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. Science, 359(6380):1094–1096, 2018.

[8] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. In IJCAI, pages 4208–4215. ijcai.org, 2018.

[9] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In ICLR (Workshop Poster), 2013.

[10] Juan Ramos. Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, pages 29–48, 2003.

[11] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A hybrid deep model for fake news detection. In CIKM, pages 797–806. ACM, 2017.

[12] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol., 10(3):21:1–21:42, 2019.

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

[14] David Sterrett, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Tompson, Tom Rosenstiel, Jeff Sonderman, and Kevin Loker. Who shared it?: Deciding what news to trust on social media. Digital Journalism, 7(6):783–801, 2019.

[15] Jingrong Tong. Journalistic legitimacy revisited. Digital Journalism, 6(2):256–273, 2018.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(86):2579–2605, 2008.

[17] Chris J. Vargo, Lei Guo, and Michelle A. Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. New Media Soc., 20(5):2028–2049, 2018.

[18] William Yang Wang. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In ACL (2), pages 422–426. Association for Computational Linguistics, 2017.

[19] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In NeurIPS, pages 9051–9062, 2019.

[20] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv., 53(5):109:1–109:40, 2020.