The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors

Nicole O'Brien[1,3] **Sophia Latessa**[1,3] **Georgios Evangelopoulos**[1,3] **Xavier Boix**[1,2,3]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA
[2]Children's Hospital, Harvard Medical School, USA
[3]Center for Brains, Minds and Machines (CBMM)
{njobrien, sofiabl, gevang, xboix}@mit.edu

Abstract

The digital information age has generated new outlets for content creators to publish so-called "fake news", a new form of propaganda that is intentionally designed to mislead the reader. With the widespread effects of the fast dissemination of fake news, efforts have been made to automate the process of fake news detection. A promising solution that has come up recently is to use machine learning to detect patterns in the news sources and articles, specifically deep neural networks, which have been successful in natural language processing. However, deep networks come with lack of transparency in the decision-making process, i.e. the "black-box problem", which obscures its reliability. In this paper, we open this "black-box" and we show that the emergent representations from deep neural networks capture subtle but consistent differences in the language of fake and real news: signatures of exaggeration and other forms of rhetoric. Unlike previous work, we test the transferability of the learning process to novel news topics. Our results demonstrate the generalization capabilities of deep learning to detect fake news in novel subjects only from language patterns.[1]

1 Introduction

The information technology boom is an unprecedented opportunity for effective democracy: well-informed citizens are able to assess their governments more thoughtfully. However, the ability of individuals to become content creators and publishers, and the speed of online news source propagation, has led to the widespread of so-called fake news, a new turn in the cyclical history of propaganda [Vosoughi et al., 2018]. The speed of fake news propagation requires automated processes of detecting misleading sources and content [Lazer et al., 2018].

Many of the current approaches to automated detection are centered around "blacklisting" authors and sources that are known producers of fake news. As the fake news publishers have become more sophisticated in their propagation strategies, machine learning algorithms have been introduced to rapidly detect patterns in news sources and articles, cf. [Li et al., 2016, Shu et al., 2017, Thorne and Vlachos, 2018].

A research direction that has shown to be effective is to analyze the language used in fake news articles, using purely text-based approaches [Khurana, 2017, Rashkin et al., 2017, Pérez-Rosas et al., 2017, Horne and Adali, 2017, Potthast et al., 2017, Horne et al., 2018, Reimink, 2018, Baly et al., 2018]. Deep learning methods have achieved tremendous success in pattern recognition tasks and similarly demonstrated to outperform previous machine learning methods for detecting

[1]Demo, data and code can be found here: http://fakenews.mit.edu/

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

fake news [Singhania et al., 2017]. The success of deep learning stems from representations that emerge while the network learns to solve a task. Such representations could be key in detecting patterns that are difficult to capture with handcrafted representations, such as the patterns underlying fake news.

In previous work, fake news detectors are trained and tested on the same news topics and possibly simply capture a bias in the dataset at the topic level, e.g. there are more fake articles about "Trump" than "soccer". As a result, the detector may not generalize well to novel topics, which is when the fake news detector is more crucial in practice, ie. when there is less information available to assess the news. In this paper, we study a fake news detector that leverages deep neural networks, and by evaluating a topic that is not included in the training dataset, we demonstrate its generalization capabilities towards novel topics. We also address a well-known problem of deep learning, which is the lack of transparency in the decision-making process, i.e. the black-box problem. Not being able to know what triggers the decisions leaves the uncertainty as to whether the algorithm can be trusted or not. Our analysis reveals that the network learns to detect language patterns in fake news articles that can be generalized toward detecting fake news in novel topics. This is a first step towards understanding the reliability of deep learning based fake news detectors.

2 Methods

Dataset Collection for Fake and Real News. We follow the standard paradigm in the literature to classify articles into fake and real news. We used the fake news dataset from Kaggle comprised of approximately 12,000 articles, as samples of fake news [Getting Real about Fake News, 2016]. These articles and metadata were gathered from 244 different websites using the internet browser extension called BS detectors [2017], which maintains a blacklist of fake news source websites. We collected real news samples from The New York Times and The Guardian by using their APIs, which provide similar information to the Kaggle dataset (text and images within the document). Our real news dataset is comprised of 9,000 Guardian articles and over 2,000 New York Times articles. In total, our dataset has approximately 24,000 articles for training and evaluation. All of the articles were published within a 30-day interval between October, 26 (2016) to November, 25 (2016). These dates are specifically chosen because they span directly before, during and after the 2016 United States Presidential Election.

Note that there may be instances of fake news in The New York Times or The Guardian and instances of real news in the Kaggle dataset. However, these instances are an exception, and we expect that the detector will learn from the majority of articles that are consistent with the label treating the wrong instances as outliers and noise.

Cleaning the Data. The raw data contains traces that are not related to the content but can reveal the origin of the articles. Examples of such traces within the real news are: "Good morning. Here's what you need to know:' or "Share on Facebook." The fake news dataset also includes such traces as "Follow us on", "Read more", or "TRENDING". These are easily learnable by the neural network and lead to an artificially high accuracy due to the underlying correlations to fake or real data. We introduce a procedure to eliminate such source-related correlations from the detector. This includes removing the aforementioned traces, as well as author mentions, title, date specifications, digits, website links, and non-English words (by using the automatic spelling dictionary [Lachowicz, 2017]). Non-unique articles or incomplete articles are also removed, yielding 8,999 real and 7,401 fake news articles (16,400 articles in total). These lead to a dataset in which the differences between fake and real news are restricted in the language.

Evaluating Generalization to Novel News topics: Topic hold-out test set. There is usually a bias in the fake news topics, e.g. the amount of articles that involve "Hillary", "Wikileaks", and "republican" is higher in fake news than in real news, or words like "football" and "love" appear very frequently in the real news while rarely found in fake news. To assess the robustness of the detector to future fake news threads, we evaluate performance in a topic excluded from the training set. To do so, we establish our training set with all the articles that do not contain a topic that we select as hold-out, and then we evaluate the detector in the hold-out topic. We use the hold-out topic of "Trump", given that our dataset is composed of articles published during the 2016 United States Presidential Election. All the articles that contain the word "Trump" constitute the hold-out set (4,416 articles).

Deep Learning. To learn to detect fake news, we use the deep neural network introduced by Kim [2014], which was shown to be effective for text classification. Each word in the text is represented as a sparse vector in which only one entry is equal to 1 in order to indicate the word, i.e. a one-hot representation. The first layer in this network is a pre-trained *word2vec* embedding, which maps each word to a 1,000-dimensional representation in which words with similar meaning are represented with a small distance [Mikolov et al., 2013]., The *word2vec* embedding is pre-trained as part of a neural network that predicts the subsequent word in a sentence, and the learned representation in this task is used in our network. The word embeddings are used as input to a convolutional layer with 128 filters of size equal to 3 words (we found this to work better than 2 and 4 words). The entire layer is max-pooled into a 128-dimensional vector, which captures the maximum value of each feature in all the article. A fully connected output layer with softmax produces probabilities for two outputs: fake and real. The network is trained to minimize the cross-entropy loss with the Adam optimizer. The hyper-parameters (learning rate, batch size, momentum, weight decay, dropout) are found by cross-validation on the training set.

Opening the Black-Box: Attributing Article Detection to Individual Words. We introduce a procedure to visualize what patterns in an article are more useful to classify it as fake or real news. This consists of finding the words that are "most fake" and "most real" by back-propagating the output of the network to the article, similarly as in [Zeiler and Fergus, 2014]. First, we select the units of the pooling layer that contributed most to the output. Let $\{w_i\}$ be the weights of a unit in the output layer (note that there is one set of weights for each of real and fake output), and let $\{a_i\}$ be the activations of the pooling layer. Since the output is $\sum_i w_i a_i$, we select the units of the pooling layer by ranking the units that contributed the most to the output, i.e. we select the units of the pooling layer with highest and lowest $w_i a_i$ values for both real and fake outputs. We select the top-20 units for each article. Then, we trace the units that caused the activation of the max-pooling, a_i, back to the convolutional layer. From the convolutional layer, we can directly find the words that caused the activation, which are the 3 words used as the input of the convolutional units (recall that the convolutional layer uses a filter size equal to 3 words).

3 Experiments

Quantitative Evaluation of Generalization to Novel News Topics. We evaluate the accuracy as the average number of articles correctly classified as fake or real news. For the hold-out topic, "Trump", the detector accurately recognizes $87.7\% \pm .9$ of the articles (average is taken on 5 different network initializations). This shows that a purely text-based convolutional network detector can provide effective baselines with good sensitivity and specificity. Given that it only takes language patterns into account, we can use it jointly with other techniques, such as automated fact-checkers, to further boost the detection accuracy.

We also evaluated the detector in novel articles of the news topics in the training set, i.e. the evaluation paradigm used in previous works. We use as testing set 4,000 articles randomly chosen. The accuracy of the detector is $93.5\% \pm .2$. The gap between this accuracy and the accuracy on hold-out topics demonstrates that there is overfitting to the bias in the topics and highlights the importance of evaluating generalization and transferability to hold-out topics. This is also an important practical consideration because the detector should be operational when the topic is novel and there is less information available to assess the article.

Qualitative Analysis of Fake and Real News Language Patterns. We now analyze the language patterns (triplets of words) responsible for classifying each article. We analyze the test hold-out set corresponding to the topic "Trump". In Figure 1, we show two extracts of articles with the emergent, most relevant patterns for the classification highlighted (the triplet of words plus the words of the original article that are discarded during the data cleaning). We can observe qualitatively a language bias in the fake news, a subtle tendency to exaggeration and use strong words to catch the attention of the reader. In order to have a general overview of this bias, in Figure 2, we visualize the 1,000 most frequent words in the extracted triplets that appear exclusively either in real or fake news from the "Trump" articles, classified by part of speech (the figure displays a randomly selected subset of words alphabetically sorted). Note that these are words that have been useful for the detector at least in one article, but it does not

Fake News: The person who received the most votes free from interference or tampering needs to be in the White House. It may well be Donald J. Trump, but further due diligence is required to ensure that American democracy is not threatened. Although the election was called on Nov 8th, the Democratic Party's ongoing campaign to delegitimize the new incoming President is still ongoing. For those of us with long enough memories, the Democratic Party, their media operatives and the Clinton Campaign were claiming that Trump and the GOP would be engaged in this very same behavior after Hillary Clinton won the Presidency (as expected). Notice how the shoe is now on the other foot. Its interesting reading past stories by news sources linked to the Clinton Campaign and John Podesta as Politico has, who ran a feature on Aug 16th entitled, Why the GOP Will Never Accept President Hillary Clinton which lays out the case of Hillary's lock on the White House and how the evil Republicans will not accept her eventual election victory. Will Donald Trump respect the peaceful transition of power?
Real News: Civil-rights campaigner and congressman John Lewis was in tears as he accepted America's National Book award for young people's literature in Manhattan on Wednesday night, speaking of how as a child he had been turned away from the public library for being black. Lewis won the prestigious US honour for the third volume of his graphic memoir March, which tells of his vital part in the civil rights movement in the 1960s. "This is unreal. This is unbelievable," said Lewis as he took to the stage with his visibly moved co-authors Andrew Aydin and Nate Powell. Recounting how he grew up "very, very poor" in rural Alabama, Lewis said there were "very few books in our home", recalling a trip in 1956 to try and borrow some books from the library. "I had a wonderful teacher in elementary school who told me: 'Read, my child, read', and I tried to read everything. I love books," said Lewis. "When I was 16 years old, some of my brothers and sisters and cousins [were] going down to the public library trying to get public library cards, and we were told the library was for whites only, not for coloureds. To come here and receive this award this honour is too much. Thank you."

Figure 1: *Qualitative examples.* The triplets of word that were most useful for the detector are hihglighted. For testing more articles visit `http://fakenews.mit.edu/`

Real Verbs: adapting, aiming, appeared, backing, campaigning, challenges, compared, debating, delivering, disappointed, drew, emerged, ensuring, fails, hit, improve, insisting, kept, leaving, offering, play, praised, ran, reducing, resisted, running, scrambling, staring, takes, urged
Fake Verbs: breaking, carrying, continue elect, fed, follow, getting, happening, help, indicate, let, lying, need, occupying, please, provided, registered, seems, spending, stated, sworn, tell, translated, want, went

Real Nouns: adaptation, aim, amazon, apprentice, artist, bridge, buyers, card, chair, charity, coal, comment, concerns, contestants, criticism, dawn, discrimination, emphasis, events, families, feeling, headline, inquest, job, journalists, legislation, manager, matches, memories, music, novelist, passion, period, potter, records, requests, scores, servants, smith, speech, streak, tackle, telephone, title, travel, vacancy, winners
Fake Nouns: ambassador, axis, bias, cause, combat, congressman, convention, corruption, courtesy, depression, dominion, enemy, eyes, fight, fund, gods, imperialist, interests, judge, lawyers, lieutenant, mailbox, missiles, oil, organization, paper, persons, police, poll, prophecy, pussy, reality, registration, republics, rumors, sense, space, systems, theories, unity, wars, wiles

Real Adjectives: aboriginal, artistic, chaotic, disappointing, eighth, fierce, grateful, guilty, lame, minimum, narrow, optimistic, popular, radical, rural, sharp, sombre, theatrical, welcome, worst
Fake Adjectives: able, bipartisan, covert, deep, divine, false, general, ill, inflammatory, largest, libertarian, moderate, numerous, patriotic, powerful, secular, standard, violent, worth

Figure 2: *Words frequently found useful to classify real and fake news.* Words are randomly selected, classified into parts of speech and sorted alphabetically. The complete list can be found here: `https://tinyurl.com/yao37vrl`

mean that they will be useful in every single article they appear. As before, we can see that the detector captures subtle differences in the language of fake and real news in our dataset.

4 Conclusions

We have shown that convolutional neural networks can be a powerful tool to detect fake news in novel topics, solely from the language (no syntax, semantics or source analysis). The difference in accuracy in topics in the training dataset and hold-out topics, shows the importance of measuring the generalization capability of the detector in hold-out topics. We opened the black-box of neural network detectors by looking into the words of an input article that are most relevant for classification, and we found a bias in the language of fake news in our dataset. This methodology is generic and can be applied to analyze any neural network or differentiable detector. Additional research is needed to understand if, for example, these language patterns can assist humans to detect fake news.

Acknowledgements. We thank Sabbi Lall, Sooyoung Kim, and especially Tomaso Poggio and Jack Hidary for very useful discussions and comments. This work was supported by the National Science Foundation Science and Technology Center Award CCF-123121.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

BS detectors. `http://bsdetector.tech`, 2017.

Getting Real about Fake News. `https://www.kaggle.com/mrisdal/fake-news`, 2016.

Benjamin D Horne and Sibel Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.

Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 235–238. International World Wide Web Conferences Steering Committee, 2018.

Urja Khurana. The linguistic features of fake news headlines and statements. Master's thesis, University of Amsterdam, June 2017.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Dom Lachowicz. Enchant - `https://github.com/AbiWord/enchant`, 2017.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1–16, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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, 2017.

Edie Reimink. Is this thesis fake news? linguistic methods for categorizing news. Master's thesis, Yale Universits, May 2018.

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

Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 3han: A deep neural network for fake news detection. In Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, editors, *Neural Information Processing*, pages 572–581, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70096-0.

James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*, 2018.

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

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.