Peering under the hood of fake-news detectors

February 4 2019

Credit: CC0 Public Domain

New work from MIT researchers peers under the hood of an automated fake-news detection system, revealing how machine-learning models catch subtle but consistent differences in the language of factual and

false stories. The research also underscores how fake-news detectors should undergo more rigorous testing to be effective for real-world applications.

Popularized as a concept in the United States during the 2016 presidential election, fake news is a form of propaganda created to mislead readers, in order to generate views on websites or steer public opinion.

Almost as quickly as the issue became mainstream, researchers began developing automated fake news detectors—so-called neural networks that "learn" from scores of data to recognize linguistic cues indicative of false articles. Given new articles to assess, these networks can, with fairly high accuracy, separate fact from fiction, in controlled settings.

One issue, however, is the "black box" problem—meaning there's no telling what linguistic patterns the networks analyze during training. They're also trained and tested on the same topics, which may limit their potential to generalize to new topics, a necessity for analyzing news across the internet.

In a paper presented at the Conference and Workshop on Neural Information Processing Systems, the researchers tackle both of those issues. They developed a deep-learning model that learns to detect language patterns of fake and real news. Part of their work "cracks open" the black box to find the words and phrases the model captures to make its predictions.

Additionally, they tested their model on a novel topic it didn't see in training. This approach classifies individual articles based solely on language patterns, which more closely represents a real-world application for news readers. Traditional fake news detectors classify articles based on text combined with source information, such as a Wikipedia page or

website.

"In our case, we wanted to understand what was the decision-process of the classifier based only on language, as this can provide insights on what is the language of fake news," says co-author Xavier Boix, a postdoc in the lab of Eugene McDermott Professor Tomaso Poggio at the Center for Brains, Minds, and Machines (CBMM) in the Department of Brain and Cognitive Sciences (BCS).

"A key issue with machine learning and artificial intelligence is that you get an answer and don't know why you got that answer," says graduate student and first author Nicole O'Brien '17. "Showing these inner workings takes a first step toward understanding the reliability of deep-learning fake-news detectors."

The model identifies sets of words that tend to appear more frequently in either real or fake news—some perhaps obvious, others much less so. The findings, the researchers say, points to subtle yet consistent differences in fake news—which favors exaggerations and superlatives—and real news, which leans more toward conservative word choices.

"Fake news is a threat for democracy," Boix says. "In our lab, our objective isn't just to push science forward, but also to use technologies to help society. ... It would be powerful to have tools for users or companies that could provide an assessment of whether news is fake or not."

The paper's other co-authors are Sophia Latessa, an undergraduate student in CBMM; and Georgios Evangelopoulos, a researcher in CBMM, the McGovern Institute of Brain Research, and the Laboratory for Computational and Statistical Learning.

Limiting bias

The researchers' model is a convolutional neural network that trains on a dataset of fake news and real news. For training and testing, the researchers used a popular fake news research dataset, called Kaggle, which contains around 12,000 fake news sample articles from 244 different websites. They also compiled a dataset of real news samples, using more than 2,000 from the New York Times and more than 9,000 from The Guardian.

In training, the model captures the language of an article as "word embeddings," where words are represented as vectors—basically, arrays of numbers—with words of similar semantic meanings clustered closer together. In doing so, it captures triplets of words as patterns that provide some context—such as, say, a negative comment about a political party. Given a new article, the model scans the text for similar patterns and sends them over a series of layers. A final output layer determines the probability of each pattern: real or fake.

The researchers first trained and tested the model in the traditional way, using the same topics. But they thought this might create an inherent bias in the model, since certain topics are more often the subject of fake or real news. For example, fake news stories are generally more likely to include the words "Trump" and "Clinton."

"But that's not what we wanted," O'Brien says. "That just shows topics that are strongly weighting in fake and real news. ... We wanted to find the actual patterns in language that are indicative of those."

Next, the researchers trained the model on all topics without any mention of the word "Trump," and tested the model only on samples that had been set aside from the training data and that did contain the word "Trump." While the traditional approach reached 93-percent accuracy,

the second approach reached 87-percent accuracy. This accuracy gap, the researchers say, highlights the importance of using topics held out from the training process, to ensure the model can generalize what it has learned to new topics.

More research needed

To open the black box, the researchers then retraced their steps. Each time the model makes a prediction about a word triplet, a certain part of the model activates, depending on if the triplet is more likely from a real or fake news story. The researchers designed a method to retrace each prediction back to its designated part and then find the exact words that made it activate.

More research is needed to determine how useful this information is to readers, Boix says. In the future, the model could potentially be combined with, say, automated fact-checkers and other tools to give readers an edge in combating misinformation. After some refining, the model could also be the basis of a browser extension or app that alerts readers to potential fake news language.

"If I just give you an article, and highlight those patterns in the article as you're reading, you could assess if the article is more or less fake," he says. "It would be kind of like a warning to say, 'Hey, maybe there is something strange here.'"

More information: "The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors." cbmm.mit.edu/sites/default/fil… -news-paper-NIPS.pdf

Provided by Massachusetts Institute of Technology