3-2020

Detecting fake news in social media: An Asia-Pacific perspective

Meeyoung CHA

Wei GAO
Singapore Management University, weigao@smu.edu.sg

Cheng-Te Li

Detecting Fake News in Social Media: An Asia-Pacific Perspective

By Meeyoung Cha, Wei Gao, Cheng-Te Li

Published in Communications of the ACM, April 2020, Vol. 63 No. 4, Pages 68-71

https://doi.org/10.1145/3378422

In March 2011, the catastrophic accident known as "The Fukushima Daiichi nuclear disaster" took place, initiated by the Tohoku earthquake and tsunami in Japan. The only nuclear accident to receive a Level-7 classification on the International Nuclear Event Scale since the Chernobyl nuclear power plant disaster in 1986, the Fukushima event triggered global concerns and rumors regarding radiation leaks. Among the false rumors was an image, which had been described as a map of radioactive discharge emanating into the Pacific Ocean, as illustrated in the accompanying figure. In fact, this figure, depicting the wave height of the tsunami that followed, still to this date circulates on social media with the inaccurate description.

Social media is ideal for spreading rumors, because it lacks censorship. Confirmation bias and filter-bubble effects further amplify the spread of unconfirmed information. Upon public outcry, independent fact-checking organizations have emerged globally, and many platforms are making efforts to fight against fake news. For example, the state-run Factually website in Singapore has been known to clarify falsehoods since its inception in May 2012, which was followed recently by the implementation of the Protection from Online Falsehoods and Manipulation Act (POFMA) in October 2019. In Taiwan, the government officially created a feature on the website of the Executive Yuan (the executive branch of Taiwan's government) to identify erroneous reporting and combat the spread of fake news. Taiwan's Open Culture Foundation has also developed and introduced the well-known anti-fake fact-checking chatbot Cofacts in May 2018. The Indonesia government since 2018 has held weekly briefings on hoax news; that same year, the country revised its Criminal Code to permit the imprisonment for up to six years of anyone spreading fake news. Governments in the Asia and Oceania region, including South Korea, Singapore, Japan, Taiwan, Philippines, Cambodia, Malaysia, have enacted relevant laws to prevent fake news from spreading.

Nonetheless, fact-checking of fake news remains daunting, and requires tremendous time and effort in terms of human investigation. Moreover, it is prone to low efficiency and inadequate coverage due to the complexity of the topics being checked, and is incapable of keeping up with the fast production and diffusion of falsehoods online. This article will review some of the latest techniques to automatically debunk fake news, many of which were initiated in the Asia and Oceania region.

Figure. The recent presidential election in Taiwan was fraught with fake news and disinformation, inflaming supporters on both sides.

Research on understanding and debunking false information spans multiple disciplines, including social psychology, information management, and computer science. Computational approaches to automate fact-checking have attracted interest, especially in data mining, natural language processing, and artificial intelligence. Existing approaches primarily rely on training classifiers, for which past events or claims are gathered and labelled as real or fake, and significant features are extracted to generate appropriate data representation. Recent techniques on deep learning further improve performance and enable the interpretability of fact-checking by representation learning and attention-based models. We categorize these methods into four streams.

- Feature-engineering method: One of the first data-driven research projects on fake news was initiated in East Asia,[3] which tried to build a classifier to debunk fake news using features crafted from temporal (for example, frequency of spikes over time), structural (density, clustering), and linguistic (for example, usage of negation and persuasion words) factors. Among them, linguistic features were found to be consistently effective over short (that is, several days) and long (that is, two months) time.[2] Modeling time-evolving evidence as features based on social interactions can further boost performance.[5] However, this stream requires ad-hoc processing of raw data for constructing features, which are painstakingly detailed, biased, and labor-intensive.
- Matching-based method: Assuming that false claims will spark responses from skeptical users who question their veracity, false rumors can be detected by searching and clustering tweets.[11] Text patterns of skepticism about factual claims are particularly useful, such as expressions like "Really?" While rule-based matching requires manual specification, semantic matching can be a viable alternative. False claims may be debunked by retrieving evidence from relatively reliable sources, such as checked news and Wikipedia articles, by an evidence-aware neural attention model that learns to highlight words related to the claim.[7]
- Representation-learning method: Deep neural networks, such as Recurrent Neural Networks (RNN), are exploited to learn latent representations of text content based on low-level features such as term frequency or word embeddings.[4] Some modeled rumors as information campaigns through Generative Adversarial Networks (GAN).[6] Recently, a co-attention network-based model was developed not only to find the correlation between posts and their comments for more accurate detection, but also to automatically identify which users, sentences, comments, and words contain fake signals.[8]
- Multimodal method: News spread in social media exhibits multiple modalities, such as text, image, and social context. Automatic fact-checking is trending to evidence collection and consolidation from multimodal information. Example methods include the RNN-based multimodal feature extraction and fusion model for rumor detection.[1] Adversarial neural networks have been developed for multimodal fake news detection by learning an event's invariant representation,[10] which removes tight dependencies of features on the specific events in the training data for a better generalization. Tri-relationship embedding models publisher-news relations and user-news interactions simultaneously to recognize fake news in the early stage of news dissemination.[9]

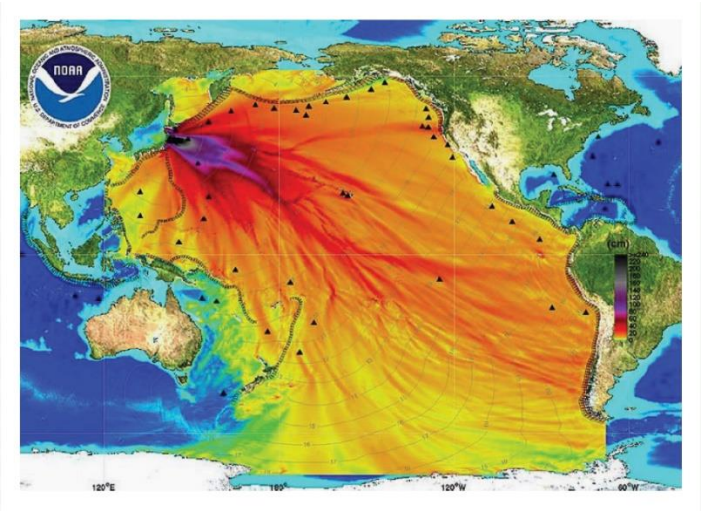

Figure. A map falsely introduced as showing the spread of radioactive seepage from the Fukushima region (https://www.snopes.com/fact-check/fukushima-emergency/). The original image was produced by the National Oceanic and Atmospheric Administration, @NOAA.

Technically, the approaches were developed from non-learning-based methods, traditional supervised learning methods, neural representation learning methods toward semi-supervised, and more recently, unsupervised methods. Content-wise, information was adopted from a single modality, such as pure text or images, toward the combination of multiple modalities. Semantically, prior methods began with shallow patterns, and hand-crafted features have been advanced by utilizing automatic feature learning, which is now tending to cross fact-checking using more sophisticated learning techniques across heterogeneous content and structures. For practical applications, the learning algorithm is shifting from employing a massive collection of user responses to relying on a limited set of observations, pursuing higher detection accuracy in the early stage of news propagation. Moreover, to reduce fake news, stakeholders require models to provide explainable outcomes that highlight which users and publishers are creating fake news, on which topics, through what types of textual and social manners.

Research on fact-checking is still in its infancy. Existing approaches bear several noticeable limitations because social media content comprises different modalities that reflect the dynamics and diversity of current events. Supervised models trained entirely on historical events that consider different types of features as a single representation can cause overfitting when the training set is multifaceted. Furthermore, while most algorithms learn to assess and discriminate inquiries, viewpoints, and stances of users on newsworthy claims, the outcome can suffer from low recall, since not all falsehoods may spark responses on social media. Also, many fake responses are intentionally created by certain groups of users to fool or attack people from other groups. Since the responses could come from everyday users or just be pertinent, the information they convey is generally too noisy and subjective to deduce a reliable veracity assessment.

To encourage advanced research in fighting against fake news, Taiwan's Ministry of Science and Technology (MOST) has created a special call for relevant research projects for which it will provide funding support. South Korea's National Research Foundation (NRF) and the Japan Science and Technology Agency (JST) also financially support research projects on detecting and tracking disinformation. We believe more and more countries in Asia and Oceania region will devote resources to the war on fake news.

The fake news phenomenon is taking new turns. YouTube and instant messaging (IM) services (for example, Whatsapp, Kakaotalk) are emerging as hotbeds of fake news. According to a survey conducted by the Korea Press Foundation, 34% of Korean YouTube viewers report having watched or received videos containing fake news. Taiwan's Open Government Foundation g0v reported that in 2017 only 46% of chatbot responses on

that nation's most popular IM app LINE is correct. Fake news on streaming platforms and IM services is particularly concerning because it contains visual content, which is more persuasive than mere text posts. Also, IM may reinforce the credibility of fake claims because people are more likely to follow trusted social contacts blindly. Data synthesis techniques like GANs can produce high-quality fabricated videos of celebrities and politicians that may appear in fake news. Intelligent fact-checking chatbots that incorporate a GAN to retrieve and generate natural-language evidence and explanations, such as the Cofacts program in Taiwan, are being progressively developed and implemented on IM services to catch false information and prevent it from spreading instantly. Consequently, the task of detecting fake news in the era of big data, social media, and artificial intelligence calls for greater attention from the research community.

References

1. Jin, Z., Cao, J., Guo, H. and Luo, J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In Proceedings of ACM Multimedia 2017, 795–816.

2. Kwon, S., Cha, M., Jung, K. Rumor detection over varying time windows. PLOS ONE 12, 1 (2017), e0168344.

3. Kwon, S., Cha, M., Jung, K., Chen, W. and Wang, Y. Prominent features of rumor propagation in online social media. In Proceedings of ICDM 2013, 1103–1108.

4. Ma, J. et al. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of IJCAI, 2016, 3818–3824.

5. Ma, J., Gao, W., Wei, Z., Lu, Y. and Wong, K.-F. Detect rumors using time series of social context information on microblogging websites. In Proceedings of CIKM, 2015, 1751–1754.

6. Ma, J., Gao, W. and Wong, K.-F. Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In Proceedings of WWW, 2019, 3049–3055.

7. Popat, K., Mukherjee, S., Yates, A. and Weikum, G. DeclarE: Debunking fake news and false claims using evidence-aware deep learning. In Proceedings of EMNLP, 2018, 22–32.

8. Shu, K., Cui, L., Wang, S., Lee, D. and Liu, H. dEFEND: Explainable Fake News Detection. In Proceedings of KDD, 2019, 395–405.

9. Shu, K., Wang, S. and Liu, H. Beyond News Contents: The Role of Social Context for Fake News Detection. In Proceedings of WSDM, 2019, 312–320.

10. Wang, Y. et al. EANN: Event Adversarial Neural Networks for multi-modal fake news detection. In Proceedings of KDD, 2018, 849–857.

11. Zhao, Z., Resnick, P. and Mei, Q. Enquiring minds: Early detection of rumors in social media from enquiry posts. In Proceedings of WWW, 2015, 1395–1405.

Authors

Meeyoung Cha is an associate professor in the School of Computing, Korea Advanced Institute of Science and Technology and the Institute for Basic Science, South Korea.

Wei Gao is an assistant professor in the School of Information Systems at Singapore Management University.

Cheng-Te Li is an associate professor in the Institute of Data Science and Department of Statistics at National Cheng Kung University, Tainan, Taiwan.