Human-Aware Interdisciplinary Models to Identify and Understand Disinformation

Kai Shu, Illinois Institute of Technology, kshu@iit.edu
Huan Liu, Arizona State University, huanliu@asu.edu

The use of social media has accelerated information sharing and instantaneous communications. Disinformation, such as fake news, hoaxes and conspiracy theories, has increasingly become a hindrance to the functioning of online social media as an effective channel for trustworthy information. There are cases in which deliberately fabricated disinformation is weaponized to divide people and create chaos in societies. The low barrier to enter social media enables more users to participate and makes them stay engaged longer, incentivizing people with hidden agenda to use disinformation to manipulate information and sway opinions. Since users have limited expertise or means to ensure protection against disinformation, the increased connectivity makes users increasingly susceptible to disinformation and becoming unwilling spreaders while being victims of disinformation. Therefore, it is imperative to understand the dissemination of disinformation and investigate how we can improve resistance against it, taking into account the tension between the need for information and the need for security and protection from disinformation. The goal of this research is to *study the scientific underpinnings of digital disinformation and develop a computational approach to human-aware protection and mitigate its rampant spread.*

However, it is a non-trivial task to combat disinformation in a digital age. To enable a trustworthy social media cyberspace, we envision two major challenges including data and users. First, the *data challenge* has been a major roadblock for researchers in their attempts to develop effective defensive means against disinformation. For example, relying on professional fact-checkers to annotate disinformation is labor-intensive and time-consuming. While the effectiveness of machine learning models usually need a large amount of training data, which is not readily available in the disinformation domain. In addition, social media data is large-scale, multimedia, mostly user-generated, sometimes anonymous, often partial and noisy. Second, The *user challenge* arises from the fact that users have very different backgrounds and knowledge, and have disparate preferences or needs. It is clearly impractical to provide a one-size-fits-all solution in building a robust and fact-based information cyberspace. For example, users on social media can have different credibility, biases, and their opinions and stances can be rather diverse. In addition, professional fact-checkers may provide more reliable judgement of identifying disinformation, while falling short in providing timely decisions in a large-scale. Incorporating heterogeneous user feedback to computational models needs proper encoding methods to reflect the source reliability and inter-agreements. In an attempt to solve these challenges, using fake news as an example, we propose to investigate human-aware machine learning models to identify and understand disinformation, with the following two tasks.

Task 1: *Detecting disinformation with human-aware supervision.* Prior works on detecting fake news rely on large amounts of labeled instances to train supervised models. Such large labeled training data is difficult to obtain in the early phase of fake news detection. To overcome this challenge, learning with weak supervision from human feedback presents a viable solution. We exploit multiple weak signals from different sources from user engagements and fact-checking information, along with contents, and their complementary utilities to detect fake news. We jointly leverage limited

amounts of clean data with human-aware supervision signals to train a fake news detector in a meta-learning framework which estimates the quality of different weak instances. We propose to derive supervision from historical social media engagements and fact-checking information. First, user engagements such as comments contain indicative signals such as sentiments, stance, and credibility, to help detect and fake news. Second, professional fact-checkers provide detailed explanations to further justify the annotations of fake news pieces. To this end, we develop a Label Weighting Network to model the weight of these weak labels that regulate the learning process of the fake news classifier.

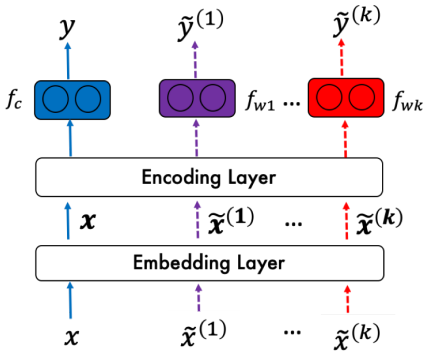

Figure 1: Modeling human-aware supervision to detect disinformation

The LWN serves as a meta-model to produce weights for the weak labels and can be trained by back-propagating the validation loss of a trained classifier on a separate set of clean data. It is suitable for early detection as only news content is needed in the testing phase. Our empirical results show that weak supervision from social media engagements can contain complementary information in addition to news content to improve fake news detection. In addition, fact-checking information provides explainable cues to make predictions more understandable.

Task 2: *Understanding disinformation dissemination from human behaviors.* Recent years have witnessed remarkable progress made towards the computational detection of disinformation. To mitigate its negative impact, however, we argue that a critical element is to understand *why* people spread fake news. Central to the question of why is the need to study the fake news sharing behavior. Deeply related to user characteristics and online activities, fake news sharing behavior is important to uncover the causal relationships between user attributes and the probability of this user to spread fake news. One obstacle in learning such user behavior is that most data is subject to *selection bias*, rendering partially observed fake news dissemination among users. To discover causal user attributes, we confront another obstacle of finding the *confounders* in fake news dissemination . Drawing on theories in causal inference, we first propose a principled approach to unbiased modelings of fake news dissemination fake news dissemination under selection bias. We then consider the learned fake news sharing behavior as the measured confounder and further identify the user attributes that potentially cause users to spread fake news. Our empirical results show that our proposed estimators achieve higher accuracy of predicting fake news that users are more likely to spread than standard estimators. For the task of identifying the causal attributes of suspicious users, we first show that the predictive accuracy can be improved by incorporating the unbiased embeddings of fake news sharing behavior. We then find that multiple user attributes are potential causes of users being suspicious.

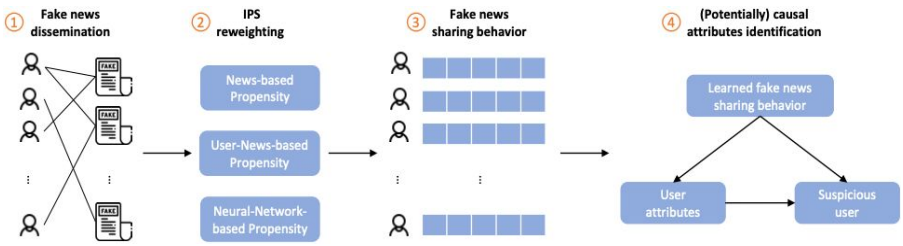

Figure 2: Modeling user behaviors to understand disinformation spreading.