

Human-Aware Interdisciplinary Models to Identify and Understand Disinformation

Kai Shu, Illinois Institute of Technology, kshu@iit.edu

Huan Liu, Arizona State University,

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

Figure 2: Modeling user behaviors to understand disinformation spreading.

modelings of 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 tas