I am developing the largest dataset of multi-lingual similar news article pairs with both either supervised learning and un-supervised learning methods with an active learning framework, and will integrate the pipeline into an open source platform (https://mediacloud.org/). Also I reasoned the causality of social feedback series with generalized linear mixed model and design a method to mediate the influence of confounder on predictor under certain circumstances.