ABSTRACT

LEARNING ALGORITHMS FOR DETECTING DISINFORMATION ON SOCIAL MEDIA

By
Courtland VanDam

Advisor: Pang-Ning Tan

Social media has become a widely accessible medium for the free exchange of ideas and knowledge. Users share their opinions and details of their personal lives, including first hand accounts of emerging/disaster events, to a wide audience. However malicious entities may abuse users’ trust to disseminate disinformation, i.e. false and misleading information. Disinformation disseminated on social media can have a significant impact offline. For example, fake news is suspected to have influenced the 2016 U.S. political election. Rumors on social media can mislead criminal investigations, e.g. the investigation of the 2013 Boston Bombing. A news agency’s compromised account lead to public panic and the stock market declined. Information phished from social media could provide hackers with access to other types of accounts, e.g. bank account. To mitigate the impacts of disinformation, detection is important.

This dissertation proposes algorithms to detect two approaches hackers use to disseminate disinformation | hashtag hijacking and compromising accounts. Hashtag hijacking occurs when one group of users takes control of a hashtag to use it in a different context than what was intended upon creation. Anyone can participate in hashtag hijacking, but to be successful, a coordinated effort among several accounts posting that hashtag is needed. Therefore, hashtag hijacking is an example of organized disinformation, i.e. multiple accounts disseminate similar misleading information in a coordinated pattern. Hashtags are used to provide context to social media posts, so those seeking to learn more about a given situation search for posts containing related hashtags. However, opponents of an idea or attention-seeking users can mislead the public via hashtag hijacking. This dissertation proposes an unsupervised multi-modal matrix factorization algorithm for detecting hijacked hashtag. This algorithm can detect hijacked hashtags where their underlying contexts were unknown
Another approach for disseminating disinformation is compromising account, where the hacker is posing as the genuine user. A social media account is compromised when it is accessed by a third party, i.e. hacker, without the genuine user’s knowledge. Compromised accounts are damaging to the account holder as well as the accounts audience, e.g. followers. Hackers who post from a compromised account can damage the user’s reputation, e.g. by posting hateful rhetoric. They also utilize the inherent trust among users to disseminate misleading information, e.g. phishing website. In this dissertation, I propose two compromised account detection algorithms, CADET and CAUTE. CADET is an unsupervised multi-view learning framework that employs nonlinear autoencoders to learn the feature embedding from multiple views. Anomalous behavior in one view, e.g. abnormal time of day, may not indicate a compromised account. However, by aggregating the anomalous behavior from multiple views, CADET projects the embedded features into a common lower-rank feature representation and detects compromised accounts in the shared space. CAUTE is a deep learning framework which simultaneously learns the encodings of users as well as the posts they publish to detect compromised accounts. CAUTE identifies anomalous posts based on approximation error of the encoders. CAUTE can determine the likelihood that a post comes from the designated user by observing a sample of their social media posts.