Detecting Toxic Content using Open Source Social Media:

A Content Centric Approach

Research Note, by The SecDev Group, 2014

Notice:

What is Kanishka?

The Kanishka Project is a multi-year initiative funded by the Government of Canada to support terrorism-focused research. Unveiled on June 23, 2011, the project is named after the Air India Flight 182 plane that was bombed on June 23, 1985, killing 329 people, most of them Canadians.

The initiative invests in research to increase understanding of the recruitment methods and tactics of terrorists, to help produce more effective policies, tools and resources for law enforcement and people on the front lines. Although the project's primary focus is on research, it also supports other activities necessary to build knowledge and create a network of researchers and students that spans multiple disciplines and research organizations.

The overarching goal of the Kanishka Project is to improve Canada's ability to counter terrorism and violent extremism at home and abroad. This report provides an account of one of the case studies funded by a grant provided to the The SecDev Group under the Kanishka Project.

About SecDev Kanishka Research

Over the past year The SecDev Group engaged in a set of practical experiments exploring techniques and methods for detecting violent extremist content and communities at risk of radicalization online.

Our approach was inspired by the public health approach to violence reduction developed by the World Health Organization. We started with four basic assumptions:

- Violent extremist groups are active and savvy users of social media spaces;
- While pathways to radicalization and violence are highly idiosyncratic, socialization plays an important role; therefore tracking and analyzing on-line ties and toxic content has potential utility.[1]
- Open-source social media (OSSM) analytics has the potential to generate information that could prove useful to improving public safety through the prevention of violent extremism;
- Methods and techniques are in their infancy. Our work is exploratory. A main purpose is to raise questions and identify areas for further research.

Our open source research explored different techniques for identifying online networks that encourage violence, as well as toxic content and its audiences. We also did some initial exploration of audience geo-location, as we thought this could provide potentially useful information for local preventative strategies.

[1] Ragheb, Abdo. 2014**.** *Review of Social Science Literature on Radicalization to Assess Operational Utility for Open Source Social Media Research in the Interests of Prevention of Violent Extremism.* The SecDev Group

This note provides a summary account of a series of social media experiments using a "content first" approach to surfacing online violent extremist networks.

Research Objectives and Methodology

In this part of the research, we set out to explore the approach of using keywords and in-group phrases as a starting point for assessing open source social media analytical techniques to surface online violent extremist networks.

Two separate experiments were conducted to identify online activity associated with the White Supremacist (WS) movement. The first experiment tested the premise that individuals who hold radical views also produce and consume radical content. The second experiment tested the assumption that radicalized individuals share common interests, which can serve as a proxy for group affiliation.

Data collection for both experiments was performed against publicly accessible content posted on the Facebook platform (FB).

Analysis

To proceed with the first experiment, working operational indicators – core incendiary keywords that are used by this group – were developed. FB Pages and Profiles returned by indicator-based searches were assessed for presence of:

1. ideological extremity;
2. incitement to violence; and
3. seriousness of intent.

The results of these searches generally failed to yield extremist or violent extremist content. Moreover, drawing a valid distinction between offensive and inciting language was often difficult.[2]

The experiment was repeated again, this time using a core, highly specific phrase (the so-called "14 words")[3] used by this in-group – which we considered to be a potential "hallmark." The network returned was indeed highly specific, with a high degree of radical content. The "hallmark" approach cut out the mainstream "noise," and was effective in surfacing a VE online network.

Network analysis of the data collected from the FB Pages and Profiles thus identified, showed that VE pages were deeply embedded in otherwise non-violent and non-extreme network of conservative political pages.

The second experiment began with a hypothesis that Facebook "interest" pages could be another way to surface VE online networks. **Facebook Interest Pages** are automatically generated from Wikipedia entries for specific topics, and, because they do not contain content, are not subject to removal by

[2] After this research was completed, we became aware of a more fruitful operational concept – that is "dangerous speech." We recommend that this concept be applied going forward.

[3] As popularized by David Lane's profile on Southern Poverty Law: David Lane's profile on Southern Poverty Law: http://www.splcenter.org/get-informed/intelligence-files/profiles/david-lane.

Facebook Community Standards (which removes Facebook Pages and Groups that post extreme content.

The experiment began with the development of an "interest profile" for members of the WS moment.[4] This interest profile was then used to perform FB Graph Search queries to locate matching FB Interest Pages. In turn, the FB Interest Pages were used as proxies to identify specific FB Pages and Groups whose members represented interest in the profile.

The data collected from the FB Pages and Profiles was examined to identify the presence and structure of a WS community. Analysis positively identified a WS FB group – the New British Union, a more extreme splinter group from the English Defence League (EDL) – that was actively recruiting far-right members of the EDL.[5]

Findings

Simple keyword searches based on behavioural or sociological indicators, such as calls to action or "us vs. them" rhetoric, are not reliable operational tools. However, qualitative open source research can often identify specific "in-group" phrases, which do provide a more efficient detection mechanism.

Using FB Interest Pages proved to be an effective means for identifying individuals and groups espousing radical ideology, seemingly in spite of ongoing removal of FB Pages and Groups featuring toxic content.

Based on these experiments, we can conclude that:

1. It is possible to identify and operationalize content-based hallmarks of violent extremist groups/online networks;
2. These groups and online networks are likely to be deeply embedded in otherwise legitimate communities; and
3. Detection methods that exploit the nuances of the FB Graph Search can be effective in detecting members of communities at risk of radicalization toward violent extremism; however, the analysis and monitoring of FB Pages – even the public pages of members of extremist groups – could contravene the public's evolving normative expectations around online privacy, even if framed as an 'early warning and intervention' activity. This issue requires informed discussion and debate. Perceptions could change, depending on who was doing the monitoring, with what authorities, and for what purpose, exactly. A clear and transparent monitoring policy could help

[4] An interest profile is a means by which Facebook automatically classifies groups, based on what users enter into their Facebook profile. For example, under interest they may put "heavy metal rock" (or in the case of WS, "lynching"). If that content category exists in Wikipedia, then Facebook will generate a list of all users of Facebook that have indicated that their interest is in "heavy metal rock." Unlike Facebook groups, interest page groups are not created by users. There is no function or option to join an interest group. Rather it is an automated means of grouping together users indicating they share a similar interest. In this respect, interest groups are a useful way to discover otherwise invisible communities. However, this method is not without its faults. Since the list is automatically generated, without any user intervention, it can yield many false positives, as it cannot recognize irony or sarcasm.

[5] A member of the NBU was convicted of firebombing a mosque, six months after he joined the more radical online grouping.