CCINT: Cyber-Crime INTelligence Framework for Detecting Online Radical Content

Name
Company
Contact info

Name
Company
Contact info

The presenter describes the background of the problem and explains what the study will focus on

1. INTRODUCTION

Several criminal and terrorist organizations have benefited tremendously from the worldwide reach, growth, and speedof the Internet. By utilizing the Internet and multiple social media platforms, they are now able to spread their views, widen their reach, and have opportunities to recruit people from all over the world. This has also given them a media platform to broadcast their messages and different propaganda material aiming to spread fear, radicalize and recruit potential members. Previous study has demonstrated that the use of Internet by terrorist groups has significantly increased in the recent years [1]. Several social media platforms such as Twitter and Facebook are working towards keeping these platforms clean by suspending those who are promoting violent content or extremist behavior. However, due to the volume and speed of the generated data it is still challenging to detect those misbehaving users accurately and in a timely manner.

Recent research has focused on studying the online behavior of pro-extremists users mainly by performing content-based analysis in order to identify distinguishing textual features that can aid in automatic detection of these users [2]. However, using this approach alone has several shortcomings including producing a large number of false positives, having a high dependency on the data, and it can be easily evaded by adapting the users writing styles through automated tools.

Building on the findings of previous research efforts, in this paper we propose a novel CyberCrime INTelligence Framework (CCINT) to detect online extremist content that is based on multi-modal approach including textual (syntactic and semantic) features, behavioral features based on social network analysis, as well as psychological-based features. We study the effects of adding these psychological, and personality features to the accuracy of our model using Linguistic Inquiry and Word Count (LIWC). We perform an experiment on the Twitter platform using our approach with the aim of detecting radical content and pro-extremist tweets. We adopt machine learning methodology to classify tweets and use our proposed approach for features identification. We envisage that this approach can be utilized by law enforcement investigators and security analysts to aid in detecting and limiting online radical propaganda.

2. RELATED WORK

In this section, we study existing literature focusing on the detection of online accounts and/or content advocating to terrorists or extremists. This covers issues related to racism, hate speech, cyberbullying and religious radicalism. In the recent years, there has been an increase in online accounts advocating and supporting terrorist groups such as ISIS. This phenomenon attracts researchers to study the online existence and research ways to automatically detect these accounts and limit their spread. Ashcroft et al. make an attempt to automatically detect Jihadist messages on Twitter [2]. They adopt a machine-learning method to classify tweets as ISIS supporters or not. They focus on English tweets that contain a reference to a set of predefined English hashtags related to ISIS. One of the limitations of their approach is that it is highly dependent on the data. Choudhary et al., [3] surveyed existing literature on counter terrorism and social network analysis. Some of the studied problems in this domain are related to identifying key-players, finding behavior patterns, community discovery, and disrupting terrorist networks. They found that the use of Social Network Analysis (SNA) is one of the most successful methods for counter terrorism in social networks.

3. APPROACH

Automatic detection ofradicalized content is a challenging task as it requires identification of the intent behind the message. Using single type of features such as textual syntactical features is prone to generating large setoffalse positives, which results in the analyst losing trust in the system. Thus, we propose a novel approach to detect radical and pro-extremist online content, which is different from existing work in that it uses multi-modal approach. We combine heterogeneous features including textual, social, and psychological into a representation that can be

used to detect radicalism. Using an unsupervised machine learning method, we cluster online text messages into groups of radical/non-radical messages. Moreover, our detection approach is designed to be semi-automatic to integrate the security analyst into the detection process. By making use of the analysts' intuition and experience to guide the detection process, we can minimize false positives generated by the system.

The presenter explains how the approach is new

3.1 Features Identification

Three classes of information are used to identify relevant features to detect radical content. Some of these features are user-based and others are message-based, as discussed in detail below:

1) Textual-based features: This class consists of text-based features calculated using text mining and natural language processing methods, such as bag of words, n-grams, most frequent words, ratio of bad words (violent words), uppercase letters, and number of emoticons. We use term frequency inverse document frequency (tf-idf) to calculate a composite weight for each term occurring in any given message. Features like n-grams, and words frequency were adopted from the literature since they have been successfully used to classify radical content. Additionally, we use upper-case letters as indicative features since they are used to convey emphasis for a word or an anger (yelling) behavior.

2) Psychological-based features: This class consists of the analysis of psychological properties of the authors of the messages. Inspired by previous research within the fields of terrorism and psychology that suggests that terrorists differ from non-terrorists in their psychological profile [4], we measure five psychological features: Personality traits (OCEAN) is a model with five domains of personality (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism). Based on lexical features a score for each personality domain is calculated per user, which together represent the person's personality. Thinking Style which focus on measuring the degree to which the person is an analytical thinker who relies on facts or feelings when making decisions. Interests focus on measuring the person interests such as work, friendship, and religion. Tone measures whether the person have a positive emotion or a negative one (joy, sadness, anger, etc.). Clout measures if the person is speaking from experience and is confident or more tentative and anxious style. For example, a radical message that promotes for violence will show high levels of anger, anxiousness, and neuroticism.

3) Behavioral-based features: This class consists of measuring behavioral features by constructing the social graph of a particular user to capture their relationships and influences within their respective community. Additionally, this class of features capture user's interactions with others through Like actions, and engagement in discussions using Mention, and Reply actions. For example, by constructing the social network for a particular user, we can measure the degree of influence this user has over his social network (e.g., using centrality measures), or identify how a particular user is linked to other known radical extremist users.

3.2 Clustering and Integrating Users-in-the-Loop

After extracting the features from the three categories previously defined, we combine them in a final feature vector. Due to the sparsity and large number of features we perform dimensionality reduction to reduce the feature space using a statistical approach called Principle Component Analysis (PCA). The new features vectors with reduced dimensionality are input to unsupervised machine learning clustering algorithms, such as K-Means and hierarchical clustering. These clustering will group the messages into groups based on the author's behavior, psychological traits, and textual violence words associated with each message. The results of the clustering algorithm will be presented to the security analysts and prompt them to tag a subset of the results as being radical or not. These tags are labels that are fed back to the system through a supervised machine learning algorithm to further learn and enhance the results accuracy. Additionally, the security analyst will be able to configure several parameters of the machine learning algorithms for example by setting the value of *K* for the K-Means algorithm.

4. RESULTS AND CONTRIBUTIONS

Initial Results are presented and a figure is added

4.1 Results

To evaluate our approach, we performed an experiment on Twitter platform. We collected data using the Twitter Streaming API following snowball sampling approach. We started the collection by filtering tweets based on an initial seed of keywords that represent the top terrorist organizations as reported by the US National Counter Terrorism Center [5]. We then extracted hashtags from the

collected data (10K unique hashtags), and added the most frequent ones (top 2%) to our seed in order to collect additional tweets. By the end of the collection process our dataset consists of around 700K tweets. Figure 1 describes the flow of our approach starting from data gathering, feature extraction, to machine learning, and the analyst's interactions through configuration and tagging of the results. Our initial results suggest that the proposed approach of combining multi-modal features is promising in detecting radical content. Moreover, the use of psychological features mainly social tone, interests, and personality traits are among the most discriminatory features.

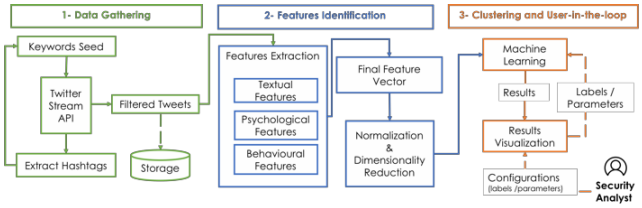

Fig. 1. Flow diagram describing our approach from data gathering, to features identification and creating the final feature vector containing textual, psychological, and behavioral features, and finally the clustering and analyst interaction with the results.

4.2 Contributions

The main contributions of this work are:

- Presenting a novel approach for detecting radicalized online messages based on multi-modal approach.
- Utilization of heterogeneous features to extract discriminatory characteristics that can identify radical content. Unlike previous efforts, these features do not only focus on lexical analysis of the messages but add additional dimensions such as psychological and behavioral aspects to improve the detection accuracy.
- Designing and developing an interactive system that allows analysts to integrate their intuition and experience into the detection process. This is achieved through the configuration and labeling process shown in Fig. 1.
- Validating our approach through an experiment performed on the Twitter platform.

5. REFERENCES

[1] R. Torok, "Developing an explanatory model for the process of online radicalisation and terrorism," Security Informatics, vol. 2, no. 1, p. 6, 2013.

[2] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, "Detecting jihadist messages on twitter," in Proceedings of the Intelligence and Security Informatics Conference (EISIC), 2015 European, Sept 2015, pp. 161–164.

[3] P. Choudhary and U. Singh, "A survey on social network analysis for counter-terrorism," International Journal of Computer Applications, vol. 112, no. 9, 2015.

[4] [4] A. W. Kruglanski and S. Fishman, "The psychology of terrorism: syndrome versus tool perspectives," Terrorism and Political Violence, vol. 18, no. 2, pp. 193–215, 2006.

[5] US National Counter Terrorism Center, "Counter Terrorism Guide," https://www.nctc.gov/site/index.html, [Online; accessed 14-Feb-2017].