Using Social Media to Understand Cyber Attack Behavior

Amy Sliva[1(✉)], Kai Shu[2], and Huan Liu[2]

[1] Charles River Analytics, 625 Mount Auburn Street, Cambridge MA 02138, USA
asliva@cra.com
[2] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287-8809, USA
{kai.shu,huan.liu}@asu.edu

Abstract. As networked and computer technologies continue to pervade all aspects of our lives, the threat from cyber attacks has also increased. However, detecting attacks, much less predicting them in advance, is a non-trivial task due to the anonymity of cyber attackers and the ambiguity of network data collected within an organization; often, by the time an attack pattern is recognized, the damage has already been done. Evidence suggests that the public discourse in external sources, such as news and social media, is often correlated with the occurrence of larger phenomena, such as election results or violent attacks. In this paper, we propose an approach that uses sentiment polarity as a sensor to analyze the social behavior of groups on social media as an indicator of cyber attack behavior. We developed an unsupervised sentiment prediction method that uses emotional signals to enhance the sentiment signal from sparse textual indicators. To explore the efficacy of sentiment polarity as an indicator of cyber-attacks, we performed experiments using real-world data from Twitter that corresponds to attacks by a well-known hacktivist group.

Keywords: Cybersecurity · Social media analytics · Sentiment analysis

1 Introduction

As networked and computer technologies continue to pervade all aspects of our lives, the threat from cyber attacks has also increased. The broad range of cyber-attacks, such as DDoS attacks, data breaches, and account hijacking, can have a strong negative impact on individuals, businesses, and broader society. Therefore, understanding these attacks and predicting them before they occur is an emerging research area with widespread applications. However, detecting attacks, much less predicting them in advance, is a non-trivial task due to the anonymity of cyber attackers and the ambiguity of network data collected within an organization; often, by the time an attack pattern is recognized, the damage has already been done. Evidence suggests that the public discourse in external sources, such as news and social media, is often correlated with the occurrence of larger phenomena, such as election results or violent attacks. Social media, in particular, turns users into "social sensors," empowering them to participate

J. I. Kantola et al. (Eds.): AHFE 2018, AISC 783, pp. 636–645, 2019.
https://doi.org/10.1007/978-3-319-94709-9_62

in an online ecosystem that interacts with behavior in the physical world. We believe the same principle can apply to cyber attacks, where open source data may provide indicators to help understand the social and behavioral phenomena leading up to an attack.

In this paper, we propose an approach that uses sentiment polarity as a sensor to analyze the social behavior of groups on social media as an indicator of cyber attack behavior. For example, extreme negative sentiment towards an organization may indicate a higher probability of it being the target of a cyber attack. However, measuring sentiment itself in social media is a challenging task due to the lack of ground truth datasets with sentiment labels and the need to extract effective and robust features from short and noisy social media posts. Both challenges make standard supervised sentiment analysis methods inapplicable. Instead, we developed an unsupervised sentiment inference method that uses emotional signals to enhance the sentiment signal from sparse textual indicators. In this method, we incorporate both emotion words and emoticons and model the correlations among them in an unsupervised manner.

To explore the efficacy of sentiment polarity as an indicator of cyber attacks, we performed experiments using real-world data from Twitter that corresponds to attacks by a well-known hacktivist group. The experimental results show that the proposed sentiment prediction framework can recognize distinct behavioral patterns associated with these attacks. We also performed a temporal analysis on the sentiment for these attacks, which provides deeper understanding of the progression of ongoing cyber attack behaviors over time.

Our contributions are summarized as follows:

(1) We propose to utilize sentiment polarity in social media as a sensor to understand and predict social behaviors related to cyber attacks;
(2) We apply an unsupervised sentiment analysis using emotional signals, which models emotion indications without requiring labeled sentiment data beforehand;
(3) We conduct experiments on real-world Tweet data related to several cyber-attacks by a well-known hacktivist group to demonstrate the effectiveness of the proposed sentiment prediction framework.

The remainder of this paper is organized as follows. In Sect. 2, we describe the use of sentiment in social media for understanding real-world events and its potential application to cyber attacks. In Sect. 3, we describe a sentiment model for social media that we applied to this problem. In Sect. 4, we present the results of experiments using Twitter data related to hacktivist attacks to illustrate the role of sentiment in this discourse. Finally, in Sect. 5 we present conclusions and plans for future work.

2 Sentiment Analysis for Behavioral Understanding

Sentiment analysis, the automated identification and quantification of opinions in a piece of text, has been an important task for natural language processing and computational linguistics. Because of the nature of social media platforms, such as Twitter and Facebook, as forums for explicitly sharing opinions and experiences, this rich social discourse can be exploited for understanding sentiment around a variety of topics

related to real-world observed behaviors. For example, sentiment analysis has been used in prediction of public opinion and political polls [10], stock market prediction [1], and analysis of large social movements or protests [2, 12].

While cyber attacks are often regarded as a technical problem for network security experts, the individuals and groups that perpetrate these attacks are still acting in accordance with the same types of social and behavioral factors that characterize these political events, stock market shifts, or social movements. Because cyber attacks are grounded within this social discourse, social media has already demonstrated value as a means for analyzing and understanding attacks, for example in threat intelligence fusion for systematic detection of cyber attacks [9] or detection of malicious cyber attack discussions [8]. Researchers have also used social media, such as Twitter and blog posts, to extract details about cyber attacks, such as attack identifiers (e.g., source IP address or MD5 hash of malware) [7], or the trending popularity of common vulnerabilities and exposures (CVEs) and their propensity to turn into real attacks [11].

We propose to extend this existing body of research on using social media to understand cyber attacks by analyzing the sentiment used to communicate cyber attack motivations, plans, or outcomes. Just as sentiment in social media can provide predictive indicators for other types of observed behaviors (e.g., political behaviors, social movements), we propose that it can also provide a way to understand and potentially predict cyber attacks. People on social media engage in massive discussion of upcoming and ongoing attacks, using it as a platform for describing possible motivations and emerging techniques, as well as for recruitment of participants for some large-scale hacktivist attacks. Social media provides us with abundant data, such as user profiles or network structure, to provide context for sentiment analyses of a post's textual content. This context enables analysis of sentiment towards particular targets of interest or motivating events, or sentiment patterns of social networks known to be part of cyber attack organizations. Preventing cyber attack damage may remain the purview of network security experts, but leveraging sentiment in the social discourse around cyber attack behaviors may enable prediction or early detection of attack preparations.

3 Modeling Sentiment in Social Media

Social media posts, such as tweets or Facebook updates, are distinct from other types of text communication. While they are often more explicit in terms of conveying sentiment than news media or other forms of communication, the posts themselves are often short and use quite informal language, making them harder to analyze using standard natural language processing approaches. Further, despite the massive quantity of social media data, there is a dearth of data sets that have been pre-annotated with sentiment information needed to train automated sentiment analysis tools. In this section, we discuss an approach to sentiment analysis in social media that addresses both of these challenges.

Many approaches to sentiment analysis in social media use supervised learning approaches [3], which require large sets of labeled training data and a large number of features for analysis. Acquiring such data is often very time consuming and labor intensive, especially when trying to develop a representative sample over large-scale

social media data. Unsupervised methods, on the other hand, do not require extensive annotation of training data, but tend to be based on predefined dictionaries of positive and negative words [1, 10]. Although manually labeling data for supervised learning is very costly, amassing vast quantities of unlabeled data for unsupervised analysis is relatively easy in social media.

To leverage the unique nature of social media data as a sentiment sensor to understand cyber attacks, we apply a novel unsupervised approach we developed in prior work [6]. Rather than relying exclusively on traditional linguistic features, this approach exploits the presence of emotional signals contained in social media posts. Examples of these emotional signals are given in Table 1. In this approach, we investigated the following problems: Are the emotional signals available in social media potentially useful for sentiment analysis? How can the emotional signals be explicitly represented and incorporated into an unsupervised sentiment analysis framework? Is the integration of emotional signals helpful for real-world sentiment analysis applications?

Table 1. Emotional signals used to indicate positive and negative sentiment in social media

Positive	:-), (-:, =), (= , (:, :), :D, :d, d:, :), (:, 8), (8, 8), ;), ;), ;), (;,;-), (-;, (;, ^ _ ^
Negative	:-(,)-:, = (,) = , :(,):, 8(,)8, :-(

Abundant emotional signals are observed in social media. Emotional signals are any information that could be correlated with sentiment polarity of a document or the words in the document. For example, when communicating in the physical world, it is common for people to supplement vocal interaction with gestures and facial expressions. Similarly, in social media, users develop visual cues that are strongly associated with their emotional states. These cues, known as emoticons (or facial expressions), are widely used to show the emotion that a user's post represents. When the authors use emoticons, they are effectively marking up the text with an emotional state. In this case, an emoticon is considered as an emotional signal. To link these emotional signals with the other content of a social media post, we look to emotional consistency theory [4, 5], which is well-established in the social sciences and models the fact that simultaneously occurring mental processes—emotions, speech, etc.—are compatible with one another. This theory suggests that words and emotional signals that often co-occur will be consistent with the same sentiment orientation, even when the posts are short.

Using this insight, our approach to sentiment analysis in social media models emotional indication in several ways:

- **Post-level Emotion Indication.** Post-level emotion indication strongly reflects the sentiment polarity of a post. The key idea of modeling post-level emotion indication is to make the sentiment polarity of a post as close as possible to the emotion indication of the post.
- **Word-level Emotion Indication.** The overall sentiment of a post is positively correlated with the sentiment of the words in that post. By modeling word-level

emotional signals, we can use this relationship to infer the sentiment polarity of a post.

Using these models of emotion indication, we then model the correlation between emotional signals and the text of a social media post at two separate levels:

- **Post-level Emotion Correlation.** To model post-level emotion correlation, we construct a post-post graph. The key idea is that if two nodes are close in the graph, their sentiment labels are also close to each other. This intuition is consistent with traditional supervised sentiment analysis, in which it is assumed that sentiment labels of two posts are more likely to be consistent when their textual similarity is high.
- **Word-level Emotion Correlation.** Similar to the interpretation of the post-level emotion correlation, we construct a word-word graph. The basic idea here is to build a latent connection to make sentiment labels of two words as close as possible if they are close in the graph.

Using this unsupervised framework, we developed a sentiment model that quantifies the sentiment of a social media post from 0 (negative) to 1 (positive). To train this model, we collected historical Twitter data from January through December 2016 related to cyber security topics, querying for specific known sources of attacks (e.g., names of hacking groups), discussions about cyber attack tactics (e.g., DDOS, phishing, etc.), specific known attack names (e.g., botnet, low orbit ion cannon, etc.), and announcements of new vulnerabilities and discussion of attacks by security experts. We collected a total of 498,019 total tweets using this method. These tweets were then automatically labeled as either positive (1) or negative (0) based on the words and emotional signal correlations indicated within the post; tweets with a mix of signals were discarded. The final training set consisted of 3,414 automatically labeled tweets. Using this dataset, we trained a logistic regression model to classify tweets according to their sentiment polarity. We performed 10-fold cross-validation on the training data, with the results shown in Table 2. The results show that, using this unsupervised method on our cyber-related training data, we can produce a high-quality model that can infer the sentiment of a tweet. In the next section, we explore how we can use this sentiment model to help understand and predict cyber attacks in the wild.

Table 2. Performance results of the sentiment model trained on cyber-related tweets

Precision	Recall	F1	Accuracy
0.858 ± 0.092	0.844 ± 0.116	0.849 ± 0.049	0.852 ± 0.092

4 Experiments: Relating Sentiment and Cyber Attacks

In the previous section, we presented a model for assigning sentiment labels to social media posts. Here, we investigate the utility of this model for understanding and predicting cyber attacks. To explore the efficacy of this model, we created a dataset consisting of several real-world cyber attacks perpetrated by a well-known hacktivist

group. This group seeks social change through cyber attacks on government, corporate, and religious websites and networked systems and makes extensive use of social media for recruitment of attack participants and publicity of attack motivations and outcomes. We selected three attacks (referenced below as Attack1, Attack2, and Attack3) that occurred between 2015 and 2017 for in-depth analysis, collecting historical Twitter data related to these attacks (using keywords associated with the motivation, target, and perpetrating organization of the attack) for a period of three weeks before and one week after the attack itself. Using this data, we conducted two types of experiments: (1) analysis of sentiment classification results; and (2) analysis of temporal sentiment trends.

For our first experiment, we wanted to analyze the ability of our sentiment classifier to produce useful results for these real-world cyber attacks. To do this, we used a standard approach in unsupervised machine learning where clustering is used to assess the discriminatory power of a trained classifier. A successful classifier will be more discriminatory, that is, will be able to identify distinct, non-overlapping classes of behavior such that the clusters are maximally different from one another, but items within each cluster are maximally similar. In our case, we wanted to show that our trained sentiment model could discriminate between positive and negative sentiment tweets related to real-world cyber attacks. For this experiment, we assumed that the sentiment space can be divided into two or three distinct categories (i.e., either just positive or negative, or positive, negative, and neutral). To measure the quality of our sentiment model, we used two standard clustering metrics: separation (i.e., the inter-cluster distinctiveness) and cohesion (i.e., intra-cluster similarity). These can be combined into a single composite metric known as the silhouette score; the silhouette score ranges from 0 to 1, with a score closer to 1 indicating more discriminatory clusters. The results of this cluster experiment are summarized in Table 3, showing very high silhouette scores for all attacks for both two and three clusters. We see slightly better results when we assume three clusters, indicating that our trained model is good at using emotional signal indicators in social media posts to differentiate between positive, negative, and neutral sentiment.

Table 3. Silhouette scores for each cyber attack

Cyber attack	Clusters	
	2	3
Attack1	0.914	0.976
Attack2	0.921	0.981
Attack3	0.908	0.986

We used principal component analysis (PCA), a feature reduction method, to project the clustering results into 2D space for easier visualization. Figure 1 shows the cluster analysis results for Attack1, Attack2, and Attack3. In all three attacks, we observe three very distinct clusters for positive, negative, and neutral sentiment, which is consistent with the high silhouette scores.

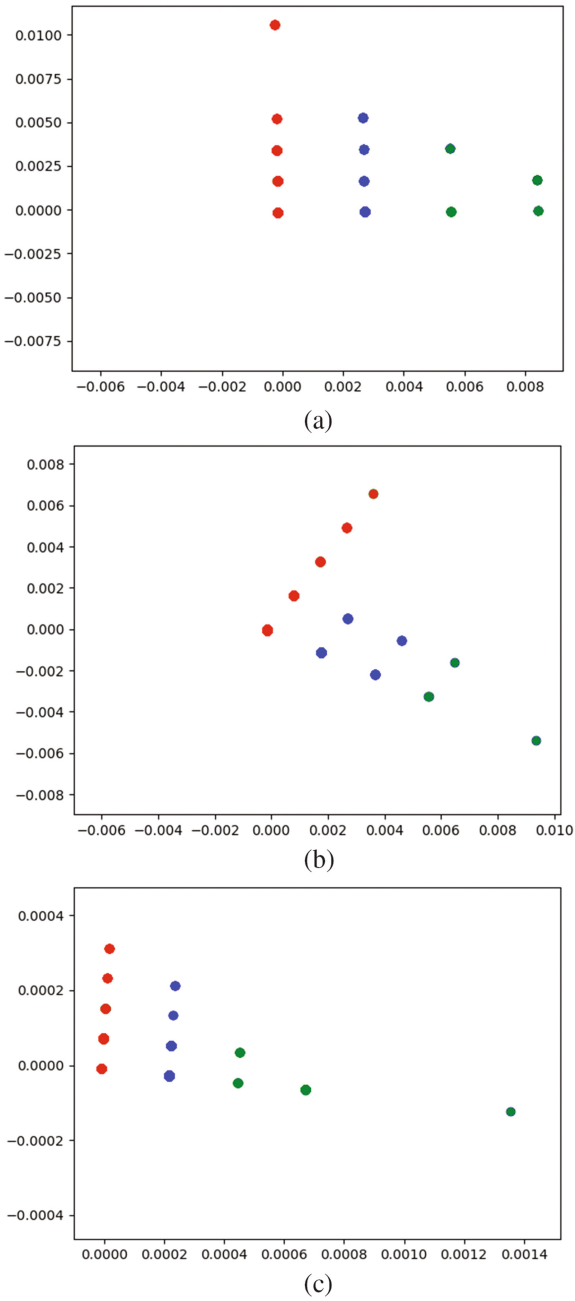

Fig. 1. Clustering results for hacktivist cyber attacks (a) Attack1, (b) Attack2, (c) Attack3

In our second experiment, we looked at temporal trends in the sentiment around each of the hacktivist attacks to better understand how sentiment relates to cyber

attacks and how it can potentially be used as a predictive indicator or early detection mechanism. For each attack, we tracked both the temporal changes in the quantity of tweets related to the attack as well as changes in sentiment over time.

Figures 2 and 3 illustrate the results of this analysis for Attack3.

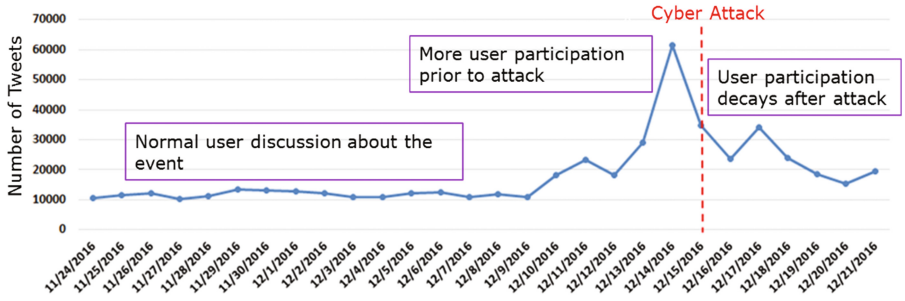

Fig. 2. Number of tweets over time before and after Attack3

Looking at these trends, we observe changes in sentiment and increased participation in the days leading up to an attack. The growing negative sentiment leading up to an attack may indicate growing anger toward the target of the attack. The subsequent spike in positive sentiment immediately before an attack parallels results in social psychology studying other types of violent attacks (i.e., terrorism), where there is a growth of in-group favoritism or enhancement immediately before an attack as the participants encourage each other and become increasingly motivated for the event. We observe similar sentiment and frequency patterns for the other two attacks as well. These results indicate that sentiment in social media may be able to provide predictive and explanatory information that can help us better understand certain types of cyber attack behavior.

5 Conclusions and Future Work

In this paper, we explore the use of social media data as a sensor for understanding cyber attack behaviors, leveraging the sentiment of posts to identify patterns that may help predict and explain cyber attacks. We apply a model of sentiment analysis that was designed to explicitly leverage the emotional signals present in social media, using unsupervised learning to exploit the massive quantities of available data without the prohibitive costs of manually labeling a training set for supervised learning. We then conducted several experiments using real-world Twitter data related to three hacktivist attacks. First, we demonstrated that our model of sentiment can successfully discriminate between positive, negative, and neutral sentiment related to cyber attacks. Second, we looked at temporal trends over tweet frequency and sentiment before and after these attacks, showing promising results for using explaining cyber attack behaviors and illustrating the potential for using sentiment for prediction.

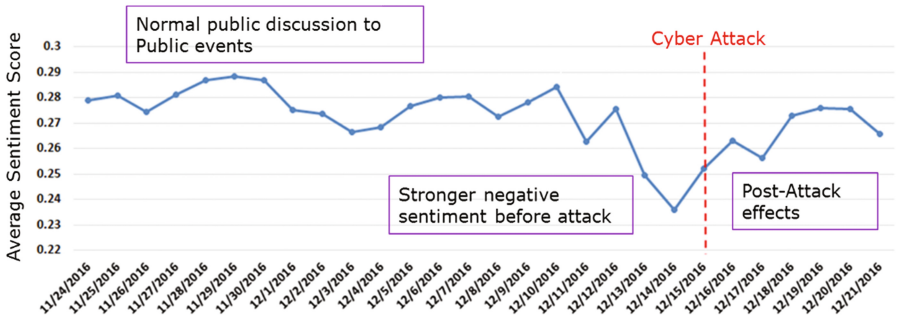

Fig. 3. Average sentiment over time before and after Attack3

The results of our early analyses are very encouraging for using sentiment in social media to understand cyber attacks. However, there are several directions for future work. First, we would like to expand our analysis beyond our initial three case studies, identifying trends in sentiment and behavioral patterns that may be common across a variety of cyber attacks. Second, we plan to analyze these sentiment results at a more granular level, separating out the positive and negative signals related to the attack target and motivation rather than looking at the average, which may obscure some more subtle variation. Third, we plan to look at other social features, such as credibility or veracity, to better understand the underlying social and behavioral patterns.

Acknowledgments. This material is based upon work supported by ONR grant N00014-17-1-2605, and the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0108. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, ONR, or the U.S. Government.

References

1. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
2. Gerbaudo, P.: Tweets and the Streets: Social Media and Contemporary Activism. Pluto Press
3. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1, 12 (2009)
4. Grawe, K.: Psychological Therapy. Hogrefe Publishing, Kirkland (2004)
5. Grawe, K.: Counseling and Psychotherapy Investigating Practice from Scientific, Historical, and Cultural Perspectives. Neuropsychotherapy: How the Neurosciences Inform Effective Psychotherapy. Lawrence Erlbaum Associates, Mahwah (2007)
6. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 537–546. ACM (2013)

7. Liao, X., Yuan, K., Wang, X., Li, Z., Xing, L., Beyah, R.: Acing the IOC game: toward automatic discovery and analysis of open-source cyber threat intelligence. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 755–766. ACM (2016)
8. Lippmann, R.P., Campbell, J.P., Weller-Fahy, D.J., Mensch, A.C., Campbell, W.M.: Finding Malicious Cyber Discussions in Social Media. MIT Lincoln Laboratory Lexington, United States (2016)
9. Modi, A., Sun, Z., Panwar, A., Khairnar, T., Zhao, Z., Doupé, A., Black, P.: Towards automated threat intelligence fusion. In: 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), pp. 408–416. IEEE (2016)
10. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. ICWSM **11**(122–129), 1–2 (2010)
11. Sabottke, C., Suciu, O., Dumitras, T.: Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits. In: USENIX Security Symposium, pp. 1041–1056 (2015)
12. Wu, C., Gerber, M.S.: Forecasting civil unrest using social media and protest participation theory. IEEE Trans. Comput. Soc. Syst. (2017)