Assessing cyber-incidents using machine learning

Article *in* International Journal of Information and Computer Security · January 2018

DOI: 10.1504/IJICS.2018.095298

4 authors:

Ross Gore
Old Dominion University
84 PUBLICATIONS **1,185** CITATIONS
SEE PROFILE

Saikou Y. Diallo
Old Dominion University
162 PUBLICATIONS **2,252** CITATIONS
SEE PROFILE

José julian Padilla
University of Guadalajara
108 PUBLICATIONS **1,282** CITATIONS
SEE PROFILE

Barry Ezell
Old Dominion University
59 PUBLICATIONS **1,041** CITATIONS
SEE PROFILE

Assessing Cyber-Incidents Using Machine Learning

ROSS GORE, Old Dominion University
SAIKOU DIALLO, Old Dominion University
JOSE PADILLA, Old Dominion University
BARRY EZELL, Old Dominion University

One of the difficulties in effectively analyzing and combating cyber attacks is an inability to identify when, why and how they occur. Victim organizations do not reveal this data for fear of disclosing vulnerabilities and attackers do not reveal themselves for fear of being prosecuted. In this paper, we employ two machine learning algorithms (text classification and topic modeling) to identify: (1) if a text-based report is related to a cyber-incident and (2) the topic within the field of cyber-security the incident report addresses. First, we evaluate the effectiveness of our approach using a benchmark set of cyber-incident reports from 2006. Then, we assess the current state of cyber-security by applying our approach to a 2014 set of cyber-incident reports we gathered. Ultimately, our results show that the combination of automatically gathering and organizing cyber-security reports in close to real-time yields an assessment technology with actionable results for intelligence and security analysts.

Categories and Subject Descriptors: C.2.2 [**Computer Security**]: Cyber-Incident Analysis

General Terms: Machine Learning, Algorithms, Analysis

Additional Key Words and Phrases: classification, cyber-security, threat assessment

ACM Reference Format:
Ross Gore, Saikou Diallo, Jose Padilla and Barry Ezell 2014. Assessing cyber-incidents using machine learning. *ACM Trans. Info. Syst. Sec.* 9, 9, Article 99 (September 9999), 16 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Information has become the critical asset in the operation and management of virtually all modern organizations. Organizations embed information, communication, and networking technologies into their core mission processes as a means to increase their operational efficiency, exploit automation, reduce response times, improve decision quality, minimize costs, and/or maximize profit. However, this increasing dependence has resulted in an environment where equipment failure, malicious insiders or external attacks on an organization's information technology infrastructure can be crippling.

One of the difficulties in constructing effective cyber-security is the inability to recognize how, when and where cyber-incidents occur. Victims do not report when they have suffered an attack for fear of revealing vulnerabilities that can be further exploited. Furthermore, attackers do not reveal the details of their intrusion for fear of prosecution. The lack of information sharing makes it difficult to identify new attack

This work is supported by these grants:.
Author's addresses: R. Gore S. Diallo Jose Padilla and Barry Ezell, Virginia Modeling, Analysis and Simulation Center, Old Dominion University; Suffolk, Virginia 23435.

DOI: http://dx.doi.org/10.1145/0000000.0000000

mechanisms (e.g. malware) and targets (e.g. vulnerabilities) in a timely fashion. Given these unknowns it is difficult to assess the evolution of cyber-incidents in real time [Richardson and Director 2008].

Furthermore, there have been limited analyses performed in the context of information from actual threats and attacks. Studies employing questionnaires administered over the Internet do not reflect valid data from actual attacks because there is no way of restricting individuals to a single response set [Rogers 1999]. Internet questionnaires have become acceptable benchmarks in the cyber-security domain because of the lack of formal routines for reporting cyber-attacks. The majority of organizations do not have a formal infrastructure in place to collect information on intrusions and those organizations that do only use it to respond effectively, not to perform scientific research.

The lack of empirical data and ensuing analysis creates opportunity. Data from cyber-incident reports contain both quantitative and qualitative data embedded in unstructured text. As a result, machine learning can be employed to automate the process of gaining insight about attacks from cyber-incident reports. Our two-tiered, machine learning-based, approach shows promise in its ability to differentiate reports related to cyber-incidents from other news. Furthermore, for the cyber-incident reports it automatically separates the reports into topics represented by a set of keywords. This analysis helps in the identification and assessment of emergent: (1) cyber-related current events, (2) attack mechanisms (e.g. malware), and (3) attack targets (e.g. vulnerabilities). While our approach requires a large quantity of incident reports retrieved regularly, we introduce an automated means to gather cyber-incident report data in close to real-time. Addressing this issue allows our approach to be immediately actionable for security and intelligence agencies. Finally, we apply our approach to the recently gathered report data to assess the current state of cyber-security. Ultimately, our work has furthered the state of the art in effectively identifying and assessing cyber-incidents.

2. BACKGROUND

Here we introduce the need for automated cyber intelligence analysis. Next, we give an overview of two areas of machine learning needed to understand our approach: text classification and topic modeling.

2.1. Cyber Intelligence Analysis

Cyber intelligence analysis is the process of producing formal descriptions of cyber threat situations and entities with appropriate statements of probability about the future actions of situations and by those entities [Clark 2012]. The process requires triangulating ground truth about a specific topic from various information sources of questionable credulity. Researchers have explored automated tools to assist analysts in correlating information from these sources about cyber-security threats to identify ground truth past and future attacks. One effort employs entity recognizers from the field of natural language processing to extract the names of people, organizations and locations from news articles. Once extracted the tool applies probabilistic models to learn the structure (in terms of people and location) of the identified organizations [Newman et al. 2006]. Other researchers use automated analysis to explore the complexity of terror attack incidents from online news reports from 2001 to 2006. Based on evidence from temporal and event data mining they found that terror attacks are increasing in complexity and incidence [Yang et al. 2006]. Another study used text classification for improving the identification of threats from malicious insiders related to weapons of mass destruction by training their tool on a document corpus from the Center for Nonproliferation Studies (CNS) [Oberhauser 2010].

These studies illustrate: (1) the growing need for intelligence analysis in the field of cyber-security and (2) the promise of using automated tools rooted in machine learning and statistics to identify and assess the large amount of unstructured data available via the World Wide Web (WWW). As means to this end, our approach applies text classification and topic modeling. We provide an overview of each next.

2.2. Text classification

Text classification is the study of classifying textual documents into predefined categories. A variety of approaches to text classification exist including naive bayesian, k-nearest neighbor, neural networks and support vector machines. The naive bayesian approach is the most widely used. It uses joint probabilities of words and categories to estimate the probability that a given document (composed of words) belongs to a given category. Documents within a certain probability are considered relevant to a category [Zhang and Li 2007; Sebastiani 2002; Baharudin et al. 2010].

The k-nearest neighbor approach to text classification identifies the k-neighbors that are most similar to a given document. The categories of these neighbors are then used to decide the category of the given document. A similarity threshold is also used for each category [Tan 2005; Sebastiani 2002; Baharudin et al. 2010].

Neural networks are a well established solution to learn patterns by constructing complicated networks of weighted edges according to training data. However, recently feed forward, back propagation (FF/BP) neural networks have also been employed in the text classification domain. The network is trained using term frequencies or other similar metrics as inputs. Based on the training data, the network predicts the category of the document [Lam and Lee 1999; Sebastiani 2002; Baharudin et al. 2010].

The final text classification approach we review is Support Vector Machine (SVM). SVM requires both a positive and negative training set for classification. This is not required by the previous described techniques. The positive and negative training set is needed for the SVM to identify the decision surface that optimally separates the positive training set from the negative training set in n-dimensional space. In many recent applications SVM has outperformed the other text classification approaches. However, it is important to recognize that SVM does require the additional negative training set and additional computational time and memory [Joachims 1998; Tong and Koller 2002; Sebastiani 2002; Baharudin et al. 2010].

2.3. Topic Modeling

A topic model is a statistical model for discovering the abstract topics that occur in a collection of documents. It assumes that the distribution of keywords within a document reflect the document's topic. For example, topic models assume that the terms *dog* and *bone* appear more often in documents about dogs, the terms *cat* and *meow* appear more often in documents about cats, and the terms *the* and *is* appear equally in both. A document typically concerns multiple topics in different proportions. As a result, in a document that is 10% about cats and 90% about dogs, topic models assume 9 times more dog words than cat words. A topic model captures this idea in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the document keywords: (1) the number of different topics in a set of documents, (2) the keywords that define those topics and (3) the ratio of topics in each document [Blei and Lafferty 2009; Steyvers and Griffiths 2007; Blei and Lafferty 2007]. In our work we employ the most common topic modeling framework, Latent Dirichlet Allocation (LDA) [Blei et al. 2003].

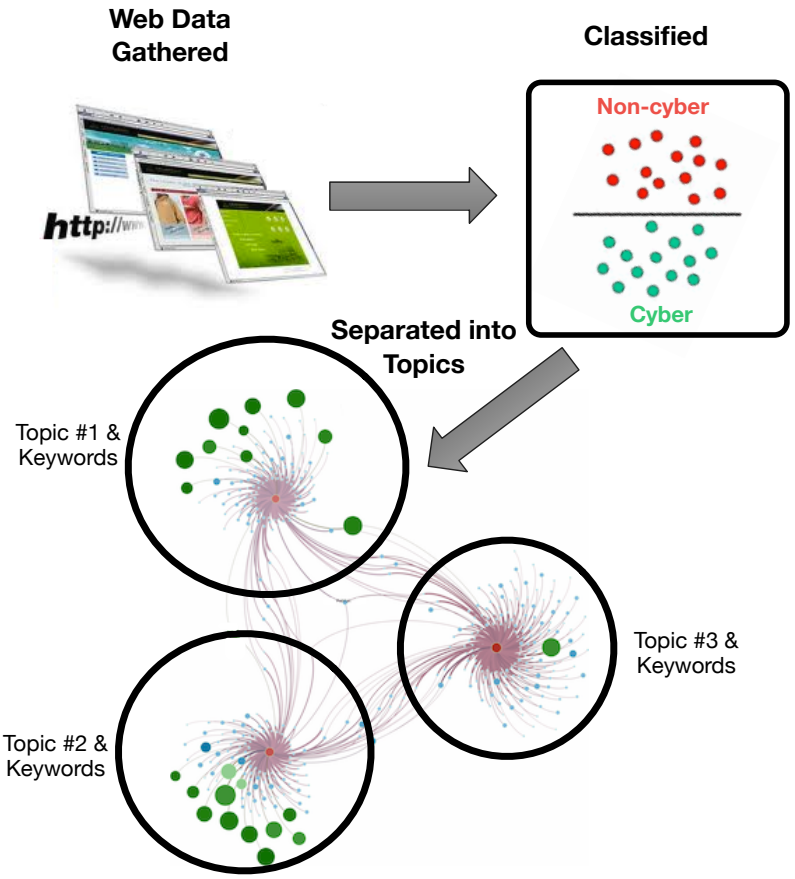

Fig. 1. Overview of our approach to classifying and analyzing cyber-reports.

3. CLASSIFICATION ALGORITHM

We employ text classification and topic modeling to create a two-tiered approach for the identification and analysis of cyber-incident reports. The structure of this approach is shown in Figure 1. It is important to note that despite the use of machine learning techniques within our work, our goal is *assessment* not prediction. First, news reports are gathered from the web. Next, we apply text classification to classify any text-based document as either reporting on a cyber-incident or not. Then, for those reports which are related to cyber-incidents we organize them into topics using topic models. This separation enables insight into: (1) cyber-related events, (2) attack mechanisms (e.g. malware), and (3) attack targets (e.g. vulnerabilities). In the "Separated into Topics" portion of Figure 1, the pink core at the center of each outlined circle represents the identified topic and the solid green circles surrounding the core represent the keywords that define the topic.

3.1. Gathering Web Data

For our initial experiments, we use the English language subset of the Nielsen BuzzMetrics dataset created in 2006. The dataset consists of about 14 million weblog documents in XML format collected by Nielsen BuzzMetrics for May 2006. The marked-up

fields of each document include: title, body and category. The English language subset of the dataset makes up 51% ($\simeq$ 7 million posts) of the entries [Nielson 2006]. Of the $\simeq$ 7 million documents there are 5,493 entires related to cyber-security attacks, threats or events. For example, the following is an excerpt from a weblog post entitled *Cyber Blackmail Is On The Rise*, "Hackers have moved away from unauthorized use of infected computers via trojan horse scripts to directly blackmailing victims. Cyber blackmailing is done by encrypting data or corrupting system information and then demanding a monetary ransom for its return to the victim" [Nielson 2006].

The Nielsen BuzzMetrics dataset and specifically the 5,493 post cyber subset of the dataset have been used in prior machine learning studies [Tsai and Chan 2007; Ng et al. 2007; Chen et al. 2008]. To form our testbed dataset we used the 5,493 documents in the cyber subset (5,493) and added to it 14,507 documents chosen at random from the remainder of the English language subset. The result is a 20,000 weblog dataset where $\simeq$1/4 of the weblog documents are cyber-incident reports.

Using an older, categorized dataset provides a platform for rigorous evaluation. The Nielsen BuzzMetrics dataset is established in the machine learning community and reflects web content generated in the wild. Furthermore, because the set of documents is categorized and includes a category for cyber-security it allows us to objectively evaluate the text classification portion of our approach. Finally, triangulating the objectives and implementations of cyber-incidents that were pervasive in May 2006 with those uncovered by our topic modeling requires the context afforded by an older dataset. The applicability of our approach to a current dataset is addressed in Section 4.

3.2. Text Classification: Is it a Cyber-Incident Report?

Text classification is a general inductive process that trains models by learning, from a set of pre-classified documents, the characteristics of a particular subject category. From these characteristics, the inductive process gleans the characteristics that a new unseen document should have in order to be classified into a particular category [Sebastiani 2002]. In our approach we ignore all marked up fields in each entry of the data set and only analyze the body of the document as unstructured text. This decision enables flexibility. By treating the weblog posts as unstructured text, our approach can be applied to research publications, lecture notes, tweets, funding solicitations, etc, without modification. Furthermore, we remove all the stop words from the documents. Stop word removal eliminates common words like *a*, *the*, *is*, *and* from the documents. This practice is common in text classification [Silva and Ribeiro 2003].

In order to maximize the effectiveness of the implementation of our classification approach we explore three different algorithms: (1) naive bayes (NB), (2) k-nearest neighbor (k-NN) and (3) support vector machine (SVM). The implementation for each algorithm is provided by the Apache Mahout library [Owen et al. 2011]. A neural network approach to classification is not explored for this stage because the unstructured nature of the data does not lend itself to developing the needed metrics for training [Lam and Lee 1999].

Each algorithm is evaluated on its ability to effectively assign a Boolean value of true or false to each document, d_i, in our dataset. A value of $true$ assigned to a document, d_i, by an algorithm indicates that the document represents a cyber-incident report, while a value of $false$ indicates that it does not. Membership within the set of cyber-indicident report documents is mutually exclusive. This means that each document, d_i, is either a cyber-incident report or it is not. Besides documents used for training, no additional knowledge of what a cyber-incident report is given to the algorithms prior to classification.

Each algorithm was evaluated using cross-validation, a widely-used evaluation methodology for text classification systems [Stone 1974; Kohavi et al. 1995]. A 50-

Table I. Classification Results.

	Accuracy (%)	Precision (macro/micro) (%)	Recall (macro/micro) (%)	F-measure (macro/micro)
NB	80.80	63.40/63.9513	60.52/62.4987	0.6005/0.6322
k-NN	87.80	87.97/94.9437	55.08/56.8221	0.6646/0.7109
SVM	89.40	81.38/82.3882	76.19/76.1396	0.7614/0.7913

fold cross validation was adopted in which the 20,000 weblog documents were divided into 50 equal portions, with 400 documents each. Testing was performed for 50 iterations, in each of which 49 portions of the data (19,600 weblog documents) were used for training and the remaining portion (400 documents) was used for testing. The data were rotated during the process such that each portion was used for testing for exactly one iteration.

We measured the effectiveness of each system using precision, recall, F-measure and accuracy. Precision measures the fraction of documents that the algorithm classified as a cyber-incident report that are cyber-incident reports, while recall measures the fraction of cyber-incident report documents within the dataset that the algorithm classified as cyber-incident reports. F-measure is a single measure that equally weights precision and recall. Accuracy measures the prediction correctness of the algorithm. These measures are commonly used in text classification and are quantified in Equations 1 - 4:

$$Precision = \frac{\text{\# of docs correctly classified by algorithm}}{\text{\# of all docs classified by algorithm}} \tag{1}$$

$$Recall = \frac{\text{\# of docs correctly classified by algorithm}}{\text{\# of all docs classified by algorithm}} \tag{2}$$

$$F - measure = \frac{\text{\# of docs correctly classified by algorithm}}{\text{\# of all docs classified by algorithm}} \tag{3}$$

$$Accuracy = \frac{\text{\# of docs correctly classified by algorithm}}{\text{\# of all docs classified by algorithm}} \tag{4}$$

There are two popular ways to calculate averages across the data for these metrics: macro-averaging and micro-averaging [Chai et al. 2002; Cohen and Singer 1999; Lam et al. 1999; Joachims 1998; Yang and Liu 1999]. In macro-averaging, the performance metrics are calculated for each iteration, and the average of all the iterations is obtained. In micro-averaging, the average is calculated across al the individual classification decisions made by each algorithm. As a result, accuracy and macro-averaging statistics are significant up to two decimal places, while micro-averaging and F-measure statistics are significant up to four decimal places.

The results of accuracy, precision, recall, and F-measure are summarized in Table I. Because F-measure represents a balance between precision and recall, we focus our discussion on accuracy and F-measure. The results demonstrated that our NB classification approach did not perform as well as our k-NN or SVM approach. It achieved the lowest accuracy and F-measure. The SVM classification approach achieved the highest accuracy and F-measure.

In order to study whether the differences among the different algorithms were statistically significant, two statistical tests were performed. The first is a micro sign-test that looks at all the classification decisions individually and uses a binomial distribution to determine whether the decisions made by any two algorithms are significantly different [Cohen 1995]. The number of observations n is defined to be the number of

Table II. Micro sign-test results.

vs.	k-NN	SVM
NB	0.00001 [a]	0.00001 [a]
k-NN		0.09725 [b]

Table III. Macro t-test results.

(Accuracy) vs.	k-NN	SVM		(F-measure) vs.	k-NN	SVM
NB	0.00001 [a]	0.00001 [a]		NB	0.08273 [b]	0.00001 [a]
k-NN		0.16272		k-NN		0.00246 [a]

times that the two systems made *different* classification decisions. The second test was a macro t-test that takes the performance of each iteration as an individual observation in order to determine whether the performances of two approaches are significantly different [Yang and Liu 1999]. Our number of observations was 50, since there were 50 iterations of testing for each approach. The macro t-test was applied to both accuracy and F-measure.

The p-values of the micro sign tests are show in Table II while the p-values of the macro t-tests on accuracy and F-measure are shown in Table III. Within the tables the superscript a denotes statistical significance at the 1% level and the superscript b denotes statistical significance at the 10% level. The results show that the superior performance of the k-NN and SVM classification approaches compared to the NB approach are statistically significant for each test. Furthermore, the superior performance of the SVM approach is statically significant compared to k-NN approach in the micro-sign test and the macro t-test for F-measure. However, there is not a statistically significant difference in the accuracy of the k-NN approach and the SVM approach. One possible reason that the SVM approach does not drastically outperform the k-NN approach is that in order to minimize the use of resources and time we employed a linear model in the SVM approach. It is possible that the performance of the SVM approach might improve if a non-linear model could be adopted, however, this would also consume more time and resources.

In order to analyze the effect of the number of training examples on the F-measure achieved by each approach, we ran the experiments on the systems while varying the size of the training data. We started with 400 documents in the first run, and increased the number of training documents by 400 in each subsequent run. There were thus 49 runs in total from (400 to 19,600 training documents). In each run, a 50-fold cross validation similar to the one described above was used, and 400 documents were used for testing with rotation. The macro-averaged F-measure for each iteration was recorded and the results are shown in Figure 2.

From the graph shown in Figure 2, we can see that the performances of all three approaches become relatively stable after 9,800 (NB), 3,600 (NN), and 5,200 (SVM) training documents respectively. It is important to note that all these approaches achieve relative stability with half or less of the training documents. We hypothesize this behavior is due to the lack of structure we impose on documents in the dataset. Recall, the body of each document at this stage each document is simply treated as unstructured text, all marked up fields are ignored. This simplicity enables the algorithms to achieve stable performance with less training data.

3.3. Topic Modeling: Organizing the Separated Cyber-Incident Reports

Next, we use topic modeling to organize the cyber-incident reports identified by our classification approach. Given the superior performance of our SVM classification approach we use the set of cyber-incident documents it classifies for this analysis. Recall, topic models allow the probabilistic modeling of keyword frequency in documents. The

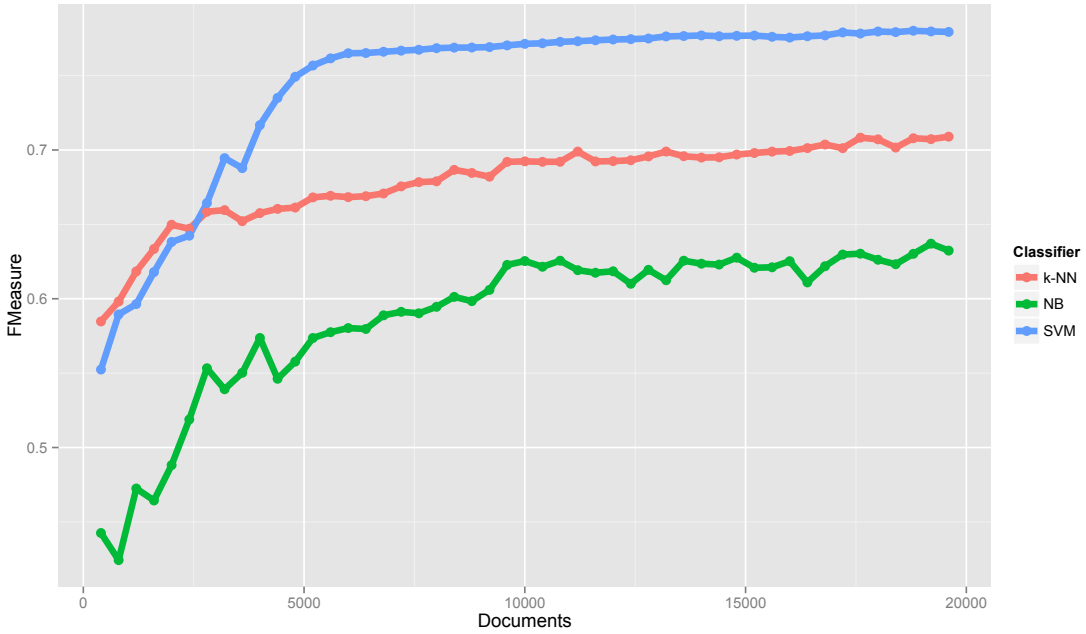

Fig. 2. Macro F-measure vs. Number of Training Documents.

models organize portions of the documents into topics based on the frequency of keywords they contain.

This analysis requires requires preprocessing the documents. Recall, we already removed marked up fields of the documents besides the body and we removed the stop words. Next, we stem and prune the documents. Stemming and pruning removes common prefixes and suffixes from the remaining words in the documents. The total number of keywords after stemming and pruning in the SVM cyber-incident report corpus is 804. Finally we use LDA topic modeling to organize the remaining keywords in the documents into topics. We employ the `topicmodels` package provided for the R programming language in this process [Dalgaard 2008; Hornik and Grün 2011].

The analysis found the maximum separation of groups of keywords by dividing them into three topics. Using the top keyword for each topic we have labeled these topics: (1) Malware Mechanisms, (2) Macintosh Vulnerabilities and (3) NSA Phone Call Database. The top ten keywords that makeup each topic are shown in Table IV.

The NSA Phone Call Database topic relates to a USA Today report on May 11, 2006 about the Bush administration and National Security Administration (NSA) using a large database of all phone calls made in the United States to counter terrorism [Cauley 2006]. The two other topics (Malware Mechanisms and Macintosh Vulnerabilities) initially appear less dated. The topic of Malware Mechanisms provides examples of types of cyber-attacks (spyware, trojan, worm) and their implementations (scan, adware, spybot, remove, http). Meanwhile the topic of Macintosh Vulnerabilities reflects increasing number of reports of cyber-attacks targeted at Macintosh computers [Frei et al. 2006; Amorosi 2011]. Of particular interest is the term *tiger* which reflects the name of the newest and most popular operating system on Macintosh computers in May 2006.

Table IV. Most Influential Terms in Identified Topic Areas Identified from Nielsen Buzzmetrics Dataset.

Malware Mechanisms	Macintosh Vulnerabilities	NSA Phone Call Database
malwar	mac	nsa
spywar	secur	bush
spybot	appl	phone
viru	tig	program
remov	attack	presid
scan	system	cia
adwar	safari	american
trojan	pref	record
web	anti	call
user	secur	administr

Our approach reveals that cyber criminals were attacking vulnerabilities on Macintosh computers through at least two issues related to the internet browser, Safari. First, the autofill option was enabled allowing spyware to easily mine personal information once installed. The second issue enabled easier access for software based cyber-attacks. The security preferences were changed in Tiger in an attempt to give users the ability to select different security options based on usage. The increase in control resulted in some users lowering the preferences below their previous levels offering bots, adware, trojan horses and other viruses easy access [Robinson 2007; Frei et al. 2008]. The application of our approach on the Nielsen BuzzMetrics dataset reveals how actionable these results would have been for intelligence and security agencies in 2006. Next, we demonstrate how search engine APIs can be employed to grow our dataset of cyber-incident reports in close to real-time. This capability allows us to construct a current (2014) cyber-incident report dataset and apply our approach to explore its applicability.

4. GROWING OUR DATASET & APPLYING OUR APPROACH IN 2014

Since 2006 cyber-security threats have evolved. Attackers have become more creative in response to the continuing efforts of the antivirus community and the growth of the counter-cyber industry. As a result the topics identified from the Nielsen Buzzmetrics dataset are no longer an accurate reflection of the current state of cyber-incidents. Assessing the current state of cyber-incidents with our two-tiered approach requires collecting a current dataset. To meet this need we gathered a random sample of the articles indexed by Bing News from July 11, 2014 to August 6, 2014 resulting in 368,121 news articles.

We refer to this set of articles as the *Bing News dataset*. Bing News marks up fields containing meta-data about each article it indexes, just as Nielsen did with the 2006 weblog data we analyzed. Once again, we ignore all marked up fields expect the body. Recall, this enables each gathered article to be represented as unstructured text. In turn, this decision enables flexibility allowing our approach to be applied to any text-based corpus.

Next, we apply our SVM classifier to separate the Bing News dataset into two categories: (1) articles related to cyber-incidents and (2) articles unrelated to cyber-incidents. Prior to application, we train our SVM classifier using the he 20,000 document subset of the Nielsen Buzzmetrics dataset described in Section 3.2. From the 368,121 document dataset the classifier identifies 4,041 as related to cyber-incidents. Since Bing does not have a cyber-security categorization for its news articles we cannot evaluate the performance of our classifier for the Bing dataset. However, we can gain insight into the effectiveness of the classifier by applying topic modeling to organize the 4,041 document cyber-corpus it identified.

Table V. Most Influential Terms in Identified Topic Areas Identified From Bing News Dataset.

Malware Mechanisms	Mobile Vulnerabilities	Degree Programs
malwar	mobil	degre
viru	app	digit
botnet	smart	forens
encrypt	sms	master
ransomwar	wireless	grant
darknet	jailbroken	gchq
sinkhol	android	nsa
spywar	password	student
zombi	instal	academ
key	bluetooth	govern

Recall, in order to apply topic modeling we need to perform additional preprocessing by stemming and pruning the 4,041 documents for words with common prefixes and suffixes. The total number of keywords after stemming and pruning in the Bing News cyber-incident report corpus is 742. Applying topic modeling separates the 742 keywords from the 4,041 documents into three topics. Using a top ten keyword for each topic we have labeled these: (1) Malware Mechanisms, (2) Mobile Vulnerabilities and (3) Degree Programs. The top ten keywords that makeup each topic are shown in Table V and the distribution of documents among those topics over the 25 days we collected data is shown in Figure 3.

4.1. The Malware Mechanism & Mobile Vulnerability Topics

Once again a *malware* and a *vulnerabilities* topic are identified. However, the keywords uncovered by the topic modeling in our approach shows the change in 2014 cyber-incidents compared to those in 2006. First we review the evolution of attack mechanisms (malware) from 2006 to 2014, then we discuss new vulnerabilities. For each topic, we demonstrate how our approach which was not trained on 2014 data was able to identify and isolate keywords related to 2014 cyber-incidents. Finally, we discuss how the assessment enabled by our approach is actionable for intelligence analysts.

4.1.1. Malware Mechanisms. Many of the malware mechanisms used in 2006 are still prevalent in 2014 [Jiang et al. 2010; Gostev 2012]. For example, in both 2006 and 2014 reports describe attack mechanisms as general malware and viruses. However, new attack mechanisms are also evident. The adware programs that were prevalent in 2006 have been replaced by rasomware. Ransomware uses encryption to make user's files inaccessible as means to extort money from them. It is important to note that while the term ransonware is new, extortion attacks are not [Luo and Liao 2009]. Recall, when describing the Nielsen Buzzmetrics dataset in Section 3.2 we provided an example document which discussed *cyber blackmailing*. According to our approach in 2014, the term *cyber blackmailing* is not used frequently, instead *ransomware* is discussed in cyber-incident reports. Our approach, trained on 2006 data was able to identify that ransomware was a prevalent mechanism for cyber-attacks despite being unfamiliar with the term.

The inclusion of other terms related to recent cyber-incidents also demonstrates the evolution of malware mechanisms and the utility of our approach. Two examples are the terms *darknets* and *sinkholing*. *Darknets* refer to closed, anonymous areas of the Internet that attackers have recently began to use to resist surveillance and avoid identification. *Sinkholing* is a mechanism anti-virus companies started using in 2010 to redirect traffic from infected clients to company servers [Crenshaw 2011]. Recently, attackers have began *encrypting* their malware with *keys* so that anti-virus companies

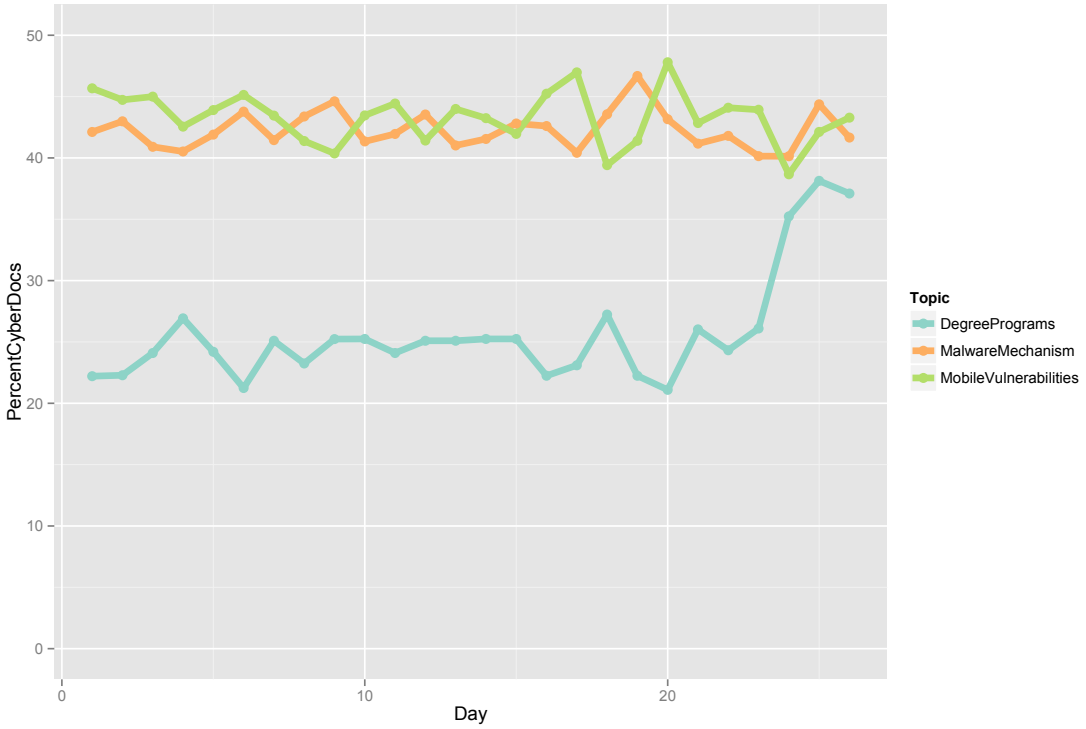

Fig. 3. The distribution of documents classified as cyber-incident reports in 2014. Recall, one document can belong to multiple topics.

will not have the ability to modify and redirect its target address [Massacci et al. 2011; Chen et al. 2010]. Each of these cases reflect new mechanisms for cyber-attacks not present in our training data. However, they were able to be uncovered as keywords by applying topic modeling to those documents that our approach classified as cyber-incident related. Next, we explore the keywords present in the Mobile Vulnerabilities topic that is present in the 2014 Bing News dataset.

4.1.2. Mobile Vulnerabilities. The inclusion of the term *mobile* as the top keyword in the vulnerabilities topic reveals a great deal about the evolution of cyber-attacks since 2006 [Gostev 2012]. In 2006 Apple Macintosh had not yet released their iPhone and most people still considered smartphones as personal device assistants (PDAs), not cellular phones. Since then, things have changed drastically. In 2012 over 400 million smartphones were in use, and the line between PDA and cellular phone became increasingly blurred [Oulasvirta et al. 2012]. As a result of this growing popularity recent research and our approach reveal that attackers have begun targeting mobile wireless technology, namely smartphones [Schultz et al. 2010]. Frequently, the most vulnerable mobile targets are devices which are *jailbroken* [Spaulding et al. 2012]. The term *jailbroken*, which is included as the 6th most influential keyword in the topic refers to devices owners have modified to gain root access and remove manufacturers' usage limitations. Since the problem of gaining root access is no longer a concern for malware these devices are especially desirable targets for cyber-attacks. The inclusion of other keywords elucidate additional vulnerabilities in mobile wireless technologies [Henry and Goldberg 2013]. Malicious applications or *apps* can steal personal infor-

mation such as account passwords and logins and send it back to the attackers. Users are vulnerable to downloading such applications because they are free and users are unaware that malicious applications can exist in the marketplace [Zhou et al. 2012]. *Bluetooth* technologies also create a unique vulnerability in current mobile devices. Unsolicited wireless devices can transmit executable malware to another device that has bluetooth enabled [Minar and Tarique 2012]. Each of these identified vulnerabilities is significantly different from those identified in 2006. Together they represent a dramatic transformation in the targets of cyber-attackers from 2006 to 2014.

4.2. The Degree Program Topic

Finally, the Degree Programs topic refers to the increasing number of universities offering their students the opportunity to pursue an advanced degree in cyber-security. In particular, during the month we collected Bing News data the Government Communications Headquarters' (GCHQ) in the United Kingdom announced their approval of six degree programs supporting their National Cyber-Security Strategy [Cornish et al. 2009]. This announcement occurred on August 4th, 2014, day 24 of our data gathering process. Figure 3 shows there was a significant spike in the documents related to this topic on days 24, 25 and 26. The are a number of similarities between the 2014 Degree Programs and 2006 NSA Database topics identified by our approach. Both represent emergent nationwide current events related to cyber-security and cyber-incidents separated and identified by our approach. The result is a capability that is actionable for intelligence analysts. We discuss our vision for actionable analysis along with the limitations of our work next.

5. DISCUSSION

In both our 2006 and 2014 evaluation, our approach was able to assess: (1) emergent cyber-security current events, targeted vulnerabilities and state of the art malware mechanisms. While the discovery of each of these artifacts is not novel, the ability of our two-tiered approach to automatically organize and isolate them from a corpus of unstructured text is novel. Intelligence analysts need proactive assessment of events, vulnerabilities and malware to enable preparedness, early identification and timely responses to cyber-incidents. Ideally, an analyst would have an automated daily report of current cyber-incident issues organized into topic areas. Our evaluation shows that our two-tiered approach is a promising avenue to achieving this capability.

Our use of the Bing News API to collect current data offers additional support. In our work, using the Bing News API to collect news data yields $\simeq$ 14,000 news articles per day. As a result, intelligence and security analysts could employ our approach daily to: (1) identify the set of news reports from the previous day related to cyber-incidents, (2) use topic modeling to organize the cyber-incident articles into current events, malware mechanisms and vulnerabilities and (3) gather $\simeq$ 14,000 new reports from the Bing News API for analysis the next day. We are confident that 14,000 articles is a large enough sample of news articles to yield actionable results. Recall, from Section 3.1 that the number of documents in our subset of the Nielsen Buzzmetrics dataset was similar (20,000) and that the performance of each classification algorithm stabilized using a 10,000 document dataset.

However, despite its ability to provide actionable assessment for intelligence analysts there a number of limitations to our approach. First, it requires a large amount of text-based reports. While the text within these reports can be unstructured, our results are only generalizable for data sets consisting of at least 10,000 documents. Furthermore, the content of the individual reports needs to be useful for the overall analysis to be useful. For example, adding a sufficient number of documents containing: (1) keywords related to cyber-incidents and (2) nonsense phrases would result in at least

one nonsensical topic being created by our approach. This vulnerability is amplified if instead of nonsensical phrases an attacker added convincing "red herring" phrases to direct attention to non-existent events, malware or vulnerabilities. Employing well vetted media outlets indexed by Bing reduces this risk, but does not completely eliminate it. Finally, the actions taken based on the topic keywords generated by of our approach are entirely left to the user(s). While our evaluation in 2006 and 2014 shows that an actionable assessment from generated keywords for each topic is straightforward, *we were created* the explanations forming those assessments *not the approach*. Ultimately, the approach limits the factors an analyst needs to consider when assessing the state of cyber-security but it does not automate the assessment process.

Despite these limitations it is important to note that our two-tiered can be applied to other domains besides cyber-security. Cyber-security is a particularly attractive domain for applying our approach because intelligence analysts need assistance in assessing current threats and trends in the face of massive data sets and volatile dynamics. However, our approach could be applied to other domains with large data sets and volatile dynamics including trading markets, traffic routing and weather forecasting.

6. RELATED WORK

A number of existing cyber-security research projects have influenced the design of our approach. The machine learning applications created to detect a variety of specific security threats from botnet traffic to phishing websites sparked our interest in automating intelligence analysis via machine learning [Ye et al. 2004; Livadas et al. 2006; Xiang et al. 2011]. Jiang et al.'s use of semantics in malware detection and Shon's hybrid approach to anomaly detection led us to consider text classification and topic modeling in combination [Shon and Moon 2007; Jiang et al. 2010]. The real-time capabilities of the ADS adaptive anomaly detector and Khoury & Tawbi's corrective enforcement runtime monitor influenced our vision of the timeliness needed in the intelligence analysis community [Ali et al. 2013; Khoury and Tawbi 2012]. Finally, the emphasis on data collection discussed in [Golle et al. 2008] guided our decision to make our approach applicable to any unstructured text.

7. CONCLUSIONS

Information has become the critical asset in the operation and management of virtually all modern organizations. Organizations embed information, communication, and networking technologies into their core mission processes as a means to increase their operational efficiency, exploit automation, reduce response times, improve decision quality, minimize costs, and/or maximize profit. However, this increasing dependence has resulted in an environment where cyber-incidents can be crippling. One of the difficulties in constructing effective cyber-security is the inability to recognize how, when and where cyber-incidents occur. Victims do not report when they have suffered an attack for fear of revealing vulnerabilities that can be further exploited. Furthermore, attackers do not reveal the details of their intrusion for fear of prosecution. The lack of information sharing makes it difficult to identify new attack mechanisms (e.g. malware) and targets (e.g. vulnerabilities). Given these unknowns it is difficult to assess the evolution of cyber-incidents in real time [Richardson and Director 2008].

Our two-tiered machine learning approach combining text classification and topic modeling directly addresses this difficulty. Given any text-based document corpus it is able to: (1) automatically classify those documents related to cyber-incidents from those documents unrelated to cyber-incidents and then (2) organize and decompose cyber-incident reports into topics and keywords. For past reports these capabilities can improve our ability to understand vulnerabilities and identify the means of attack.

For current data it offers an automated means to organize and identify new cyber-security threats and events from any corpus of unstructured text. Furthermore, we demonstrate that a sufficient amount of data can be collected for daily analysis by employing news search engine APIs.

There are significant implications from the ability to gather sufficient daily data and deploy our two tiered approach. Intelligence analysts need proactive assessment of events, vulnerabilities and malware to enable preparedness, early identification and timely responses to cyber-incidents. Ideally, an analyst would have an automated daily report of current cyber-incident issues organized into topic areas. Our evaluation shows that employing our two-tiered approach to achieve this capability is promising. In future work, we will explore case studies with intelligence analysts to transform our approach into a usable, actionable, automated daily cyber-incident assessment technology.

REFERENCES

Muhammad Qasim Ali, Ehab Al-Shaer, Hassan Khan, and Syed Ali Khayam. 2013. Automated Anomaly Detector Adaptation using Adaptive Threshold Tuning. *ACM Transactions on Information and System Security (TISSEC)* 15, 4 (2013), 17.

Drew Amorosi. 2011. Rotting Apple. *Infosecurity* 8, 4 (2011), 6–9.

Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology* 1, 1 (2010), 4–20.

David M Blei and John D Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics* (2007), 17–35.

David M Blei and John D Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications* 10 (2009), 71.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

Leslie Cauley. 2006. NSA has massive database of Americans? phone calls. *USA today* 11, 06 (2006).

Kian Ming Adam Chai, Hai Leong Chieu, and Hwee Tou Ng. 2002. Bayesian online classifiers for text classification and filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 97–104.

Changlong Chen, Min Song, and George Hsieh. 2010. Intrusion detection of sinkhole attacks in large-scale wireless sensor networks. In *Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference on*. IEEE, 711–716.

Yun Chen, Flora S Tsai, and Kap Luk Chan. 2008. Machine learning techniques for business blog search and mining. *Expert Systems with Applications* 35, 3 (2008), 581–590.

Robert M Clark. 2012. *Intelligence analysis: a target-centric approach*. CQ press.

William W Cohen. 1995. Text categorization and relational learning. In *ICML*. Citeseer, 124–132.

William W Cohen and Yoram Singer. 1999. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems (TOIS)* 17, 2 (1999), 141–173.

Paul Cornish, Rex Hughes, and David Livingstone. 2009. Cyberspace and the National Security of the United Kingdom. *Threats and Responses. Chatham House, London* (2009).

Adrian Crenshaw. 2011. Darknets and hidden servers: Identifying the true IP/network identity of I2P service hosts. *Black Hat DC* 201, 1 (2011).

Peter Dalgaard. 2008. *Introductory statistics with R*. Springer.

Stefan Frei, Martin May, Ulrich Fiedler, and Bernhard Plattner. 2006. Large-scale vulnerability analysis. In *Proceedings of the 2006 SIGCOMM workshop on Large-scale attack defense*. ACM, 131–138.

Stefan Frei, Bernhard Tellenbach, and Bernhard Plattner. 2008. 0-day patch exposing vendors (in) security performance. *BlackHat Europe* (2008).

Philippe Golle, Frank McSherry, and Ilya Mironov. 2008. Data collection with self-enforcing privacy. *ACM Transactions on Information and System Security (TISSEC)* 12, 2 (2008), 9.

Aleks Gostev. 2012. Cyber-threat evolution: the year ahead. *Computer Fraud & Security* 2012, 3 (2012), 9–12.

Ryan Henry and Ian Goldberg. 2013. Batch proofs of partial knowledge. In *Applied Cryptography and Network Security*. Springer, 502–517.

Kurt Hornik and Bettina Grün. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40, 13 (2011), 1–30.

Xuxian Jiang, Xinyuan Wang, and Dongyan Xu. 2010. Stealthy malware detection and monitoring through VMM-based "out-of-the-box" semantic view reconstruction. *ACM Transactions on Information and System Security (TISSEC)* 13, 2 (2010), 12.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Raphael Khoury and Nadia Tawbi. 2012. Corrective enforcement: a new paradigm of security policy enforcement by monitors. *ACM Transactions on Information and System Security (TISSEC)* 15, 2 (2012), 10.

Ron Kohavi and others. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, Vol. 14. 1137–1145.

Savio LY Lam and Dik Lun Lee. 1999. Feature reduction for neural network based text categorization. In *Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on*. IEEE, 195–202.

Wai Lam, Miguel Ruiz, and Padmini Srinivasan. 1999. Automatic text categorization and its application to text retrieval. *Knowledge and Data Engineering, IEEE Transactions on* 11, 6 (1999), 865–879.

Carl Livadas, Robert Walsh, David Lapsley, and W Timothy Strayer. 2006. Using machine learning techniques to identify botnet traffic. In *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*. IEEE, 967–974.

Xin Luo and Qinyu Liao. 2009. Ransomware: A new cyber hijacking threat to enterprises. *Handbook of research on information security and assurance* (2009), 1–6.

Fabio Massacci, Stephan Neuhaus, and Viet Hung Nguyen. 2011. After-life vulnerabilities: a study on firefox evolution, its vulnerabilities, and fixes. In *Engineering Secure Software and Systems*. Springer, 195–208.

Nateq Be-Nazir Ibn Minar and Mohammed Tarique. 2012. Bluetooth security threats and solutions: a survey. *International Journal of Distributed and Parallel Systems* 3, 1 (2012), 127–148.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Analyzing entities and topics in news articles using statistical topic models. In *Intelligence and Security Informatics*. Springer, 93–104.

Kok Wah Ng, Flora S Tsai, Lihui Chen, and Kiat Chong Goh. 2007. Novelty detection for text documents using named entity recognition. In *Information, Communications & Signal Processing, 2007 6th International Conference on*. IEEE, 1–5.

Nielson. 2006. Nielson BuzzMetrics Weblog Dataset. (March 2006). Retrieved July 7, 2014 from http://www.icwsm.org/data.html

Roy Oberhauser. 2010. Leveraging Semantic Web Computing for Context-Aware Software Engineering Environments. *Semantic Web, In-Tech* (2010), 157–179.

Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, and Eeva Raita. 2012. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing* 16, 1 (2012), 105–114.

Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman. 2011. *Mahout in action*. Manning.

Robert Richardson and CSI Director. 2008. CSI computer crime and security survey. *Computer Security Institute* 1 (2008), 1–30.

Brian Robinson. 2007. OS X Tiger Security Issues. http://www.itsecurity.com/features/top-12-os-x-tiger-security-issues-032007/. (2007). Accessed: 2014-06-30.

M Rogers. 1999. Psychology of hackers: Steps toward a new taxonomy. *online at http://www. infowar. com/hacker/99/HackerTaxonomy. shtml* (1999).

David Schultz, Barbara Liskov, and Moses Liskov. 2010. MPSS: Mobile proactive secret sharing. *ACM Transactions on Information and System Security (TISSEC)* 13, 4 (2010), 34.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 34, 1 (2002), 1–47.

Taeshik Shon and Jongsub Moon. 2007. A hybrid machine learning approach to network anomaly detection. *Information Sciences* 177, 18 (2007), 3799–3821.

Catarina Silva and Bernardete Ribeiro. 2003. The importance of stop word removal on recall values in text categorization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, Vol. 3. IEEE, 1661–1666.

John Spaulding, Alyssa Krauss, and Avinash Srinivasan. 2012. Exploring an open WiFi detection vulnerability as a malware attack vector on iOS devices. In *Malicious and Unwanted Software (MALWARE), 2012 7th International Conference on*. IEEE, 87–93.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis* 427, 7 (2007), 424–440.

Mervyn Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* (1974), 111–147.

Songbo Tan. 2005. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications* 28, 4 (2005), 667–671.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* 2 (2002), 45–66.

Flora S Tsai and Kap Luk Chan. 2007. Detecting cyber security threats in weblogs using probabilistic models. In *Intelligence and Security Informatics*. Springer, 46–57.

Guang Xiang, Jason Hong, Carolyn P Rose, and Lorrie Cranor. 2011. CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)* 14, 2 (2011), 21.

Christopher C Yang, Xiaodong Shi, and Chih-Ping Wei. 2006. Tracing the event evolution of terror attacks from on-line news. In *Intelligence and Security Informatics*. Springer, 343–354.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 42–49.

Nong Ye, Yebin Zhang, and Connie M Borror. 2004. Robustness of the Markov-chain model for cyber-attack detection. *Reliability, IEEE Transactions on* 53, 1 (2004), 116–123.

Haiyi Zhang and Di Li. 2007. Naive Bayes text classifier. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on*. IEEE, 708–708.

Yajin Zhou, Zhi Wang, Wu Zhou, and Xuxian Jiang. 2012. Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets.. In *NDSS*.

Received February 2999; revised March 2999; accepted June 2999