Automatic detection of cyber-recruitment by violent extremists

Article *in* Security Informatics · December 2014

DOI: 10.1186/s13388-014-0005-5

2 authors, including:

Matthew Steven Gerber
University of Virginia
70 PUBLICATIONS **2,856** CITATIONS

SEE PROFILE

Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Automatic detection of cyber-recruitment by violent extremists

Jacob R Scanlon* and Matthew S Gerber

Abstract

Growing use of the Internet as a major means of communication has led to the formation of cyber-communities, which have become increasingly appealing to terrorist groups due to the unregulated nature of Internet communication. Online communities enable violent extremists to increase recruitment by allowing them to build personal relationships with a worldwide audience capable of accessing uncensored content. This article presents methods for identifying the recruitment activities of violent groups within extremist social media websites. Specifically, these methods apply known techniques within supervised learning and natural language processing to the untested task of automatically identifying forum posts intended to recruit new violent extremist members. We used data from the western jihadist website *Ansar AlJihad Network*, which was compiled by the University of Arizona's Dark Web Project. Multiple judges manually annotated a sample of these data, marking 192 randomly sampled posts as recruiting (YES) or non-recruiting (NO). We observed significant agreement between the judges' labels; Cohen's $\kappa = (0.5, 0.9)$ at $p = 0.01$. We tested the feasibility of using naive Bayes models, logistic regression, classification trees, boosting, and support vector machines (SVM) to classify the forum posts. Evaluation with receiver operating characteristic (ROC) curves shows that our SVM classifier achieves an 89% area under the curve (AUC), a significant improvement over the 63% AUC performance achieved by our simplest naive Bayes model (Tukey's test at $p = 0.05$). To our knowledge, this is the first result reported on this task, and our analysis indicates that automatic detection of online terrorist recruitment is a feasible task. We also identify a number of important areas of future work including classifying non-English posts and measuring how recruitment posts and current events change membership numbers over time.

Keywords: Cyber; Recruitment; Extremist; Terrorism; Darkweb; Machine learning; Natural language processing

Introduction

In the last decade, the modern landscape of extremism has expanded to encompass the Internet and online social media [1,2]. In particular, extremist organizations have increasingly used these technologies to recruit new members. Recent research by Torok shows that cyber tools are most influential at the onset of a future member's extremist activity—the recruitment and radicalization phase [2]. Terrorist groups use the free and open nature of the Internet to form online communities [3] and disseminate literature and training materials without having to rely on traditional media outlets which might censor or change their message [2,4]. Terrorist organizations engage in directed communication and advertisement, recruiting members on social websites like Second Life, Facebook, and radicalized religious web forums [1,2,5]. The intelligence community would benefit from knowledge of how terrorist organizations conduct online recruitment and whom they may be targeting.

The investigation report on FBI counterterrorism intelligence failures leading up to the Ft. Hood shooting on November 7, 2009 cited a "data explosion" and "workload" as contributing factors to analyst and agent oversights. At "nearly 20,000 Aulaqi-related [electronic documents]," keeping up with workload demands was clearly a challenge for the two reviewers assigned to the case at the time [6]. Considering this large volume of possibly relevant text data requiring review by a limited number of FBI agents, automated classification methods would be useful for pre-screening text documents—reducing the workload of human analysts.

Within this article, a violent extremist (VE) group is an organization that uses violent means, like terrorism,

*Correspondence: jrs6du@virginia.edu
Predictive Technology Laboratory, Department of Systems and Information Engineering, University of Virginia, Charlottesville VA, USA

to disrupt a legitimate authority, whereas insurgents and terrorists are common types of violent extremist groups that act with the specific goal of influencing public opinion or inciting political change. A radical religious group organizing inflammatory yet peaceful protests or a politically motivated person engaging in civil disobedience are *not* considered violent extremists under these definitions. Many modern groups, like the Westboro Baptist Church, have radical religious views, but these beliefs are neither necessary nor sufficient to classify them as violent extremists without the intent to carry out or advocate for *specific acts of violence*. Within this article, VE recruitment is any attempt by a group or individual involved in VE to recruit, radicalize, or persuade another person to aid a violent movement. Cyber-recruitment is therefore recruitment activity that makes use of computers and the Internet.

This article presents data and analytic methods for automatically identifying recruitment activities of violent extremist organizations within online social media. Specifically, these methods identify messages recruiting individuals for participation in violent extremism. For these classification purposes, a VE cyber-recruitment message is any message that attempts to persuade the reader to join a violent extremist organization. These recruitment messages must assist readers in finding violent movements to join, or describe ways to become more active or provide material aid. By developing and evaluating an automatic system for identifying such messages, we demonstrate an important and feasible method for identifying the intention/incitement of violent activity within online communities.

The rest of this article is organized as follows. In the section "Related work", we compare offline violent extremist recruitment with the recent increase in cyber-recruitment efforts. Additionally, we discuss previous counterinsurgency efforts and contemporary research that outlines the challenges associated with analyzing VE activities, like recruitment; we emphasize the specific gaps that our article addresses. The "Data collection and annotation" section describes our data requirements and the specific data sources we used, followed by the pre-processing and annotation steps required for supervised learning of VE recruitment. Following the annotation steps, we present our agreement analysis results of the VE recruitment annotations. In the section "Analytic approach", we propose a probabilistic model employing natural language features for automatically classifying VE recruitment in forum posts. We also describe the classification functions used in our supervised learning experiments, such as naive Bayes, logistic regression, and support vector machines. A description of the results obtained from the experiments along with our interpretation is provided in the "Results and discussion" section. Finally, "Conclusions and future work" section discusses future directions and potential for our proposed techniques along side privacy concerns related to automated monitoring and analysis of ubiquitous communication.

Related work

Offline recruitment and manual social network analysis

The modern jihadist insurgencies in Iraq and Afghanistan operate among the local civilian population and engage in both legal and illegal activities in order to achieve their strategic and political goals. However, the illegal acts are only effective when carried out by an organized and well-manned group [7]. Recruiting new members is thus a critical activity for both daily operations and the underlying political cause. An average terrorist group has a life expectancy of less than a year, so groups wishing to extend their lifespan must replace members lost through arrests, deaths, and defections [3]. Several studies have tried to understand why some people join violent rebellions [8-12], while others only sympathize or cooperate in a non-violent capacity [13-16]. This article facilitates such understanding by providing methods that identify examples of active recruitment activity within a population of individuals who may passively sympathize with violent groups.

Ralph McGehee observed VE recruitment first hand during his 1967 work to identify communist insurgents in the rural villages along the northern border of Thailand. His efforts enabled the joint CIA-Thai counterinsurgency (COIN) to provide targeted aid to at-risk villages and persons, and in doing so simultaneously thwart communist recruitment efforts and improve regional support for the Thai government. The success of McGehee's program can be attributed to his intelligence teams collecting information on nearly every person in the villages, not just the communist sympathizers he was specifically targeting. This provided a more complete picture of the community and allowed this early social network analysis (SNA) effort to better infer the community's support for the communists and successfully identify active members of the insurgency [17]. Although our research problem specifically targets online communities, strong parallels exist between these virtual worlds and the physical communities addressed by McGehee because both contain violent extremist groups that operate within, hide among, and recruit from a passive majority population.

Cyber-recruitment, social network analysis, and data mining

The primary danger of cyber-recruitment is its ability to quickly expose large online communities to a substantial amount of engaging, multimedia content [2,18,19]. COIN experts are increasingly concerned with the potential of these cyber-communities for illegal purposes. Most

literature has focused on how violent extremist groups use legitimate social networking websites along with online discussion forums for recruitment and other activities. This prior research largely provides evidence and case studies of real online VE activity and suggests ways that virtual worlds may be used by these groups in the future [4,5,19-22]. Recent research has evaluated the use of political tools for shutting down websites or shaming material supporters [23]. Some researchers have suggested the use of web-crawling and analysis techniques to monitor for VE activities including recruitment [2,24,25]; however, we are not aware of any implementations of such techniques on recruitment specifically. This article presents new research that fills this gap, addressing the need to detect cyber-recruitment in online social media forums.

Computer-based social network analysis is a large field of research, one objective of which is to identify the organizational structure of VE networks [1,26-29]. With objectives similar to McGehee's manual SNA work, present research hopes to detect the presence of VE groups and their influence within large-scale networks based on the number of interconnections among VEs and influential community members. There have also been preliminary attempts to profile individual users using text mining techniques [30]. However, this prior research has typically focused on violent extremist activity in general without focusing on a particular activity like recruitment. Although much COIN literature has covered cyber-recruitment, and data/text mining techniques have been used in an early capacity to collect/analyze Internet data, no published research has applied such techniques to specifically examine the cyber-recruitment activities of extremist groups in online environments. The present research complements the research surveyed above by building on recent data collection efforts, focusing on online recruitment specifically, and applying current techniques from natural language processing to automatically identify recruitment activities.

Data collection and annotation

The need for cyber-COIN tools has increased interest in methods that analyze so-called "dark web" content. Dark web content is defined as information from typically private social websites where extremists interact. Many early efforts focused on locating, accessing, extracting, and storing data from dark web forums [1,22,24,31,32]. The present research builds on these key efforts. In the section "Data requirements and sources", we describe requirements that must be met by data sources supporting our objectives along with specific data sources used in our study. In the "Data pre-processing and annotation" section, we describe our manual annotation effort, which analyzed individual posts for recruitment content.

Data requirements and sources

This article leverages prior data collection efforts by using pre-compiled forum post data to model violent extremist recruitment within online social media. The following data requirements are needed to support our research objectives.

Violent extremist activity - The collected data should come from sources that are popular among violent extremist groups and their sympathizers and contain overt recruitment for such groups.

Contemporary time-frame - The collected data should cover a contemporary time-frame (e.g., the last decade) in order to be considered relevant to contemporary anti-extremist efforts.

Language - The collected data must use the English language or be translatable to English using an automatic process like Google's machine-translation service [33].

We identified the Dark Web Portal Project [31,34] as an ideal data source according to the requirements described above. The Dark Web Portal is a repository of social media messages compiled from 28 different online discussion forums. These forums focus on extremist religious (e.g., jihadist) and general Islamic discussions, many of which are sympathetic to radical Islamic groups. Most of the thirteen million collected messages come from Arabic sources, but the Dark Web Project provides translation services and compiles information from at least seven dedicated English-language forums. The most relevant forums come from the Ansar AlJihad Network, which we summarize in Table 1 and describe in more detail below.

The Ansar AlJihad Network is a set of invitation-only jihadist forums in Arabic and English that are known to be popular with western jihadists [35]. The Dark Web Project compiled 299,040 total messages posted on Ansar AlJihad between 2008-2012. Fewer posts are compiled from the English forums, called Ansar1, than from the Arabic portion of the site; however, the English subset was sufficiently large for our study and contained contemporary,

Table 1 Forums used in our study, extracted from the Ansar AlJihad Network via the Dark Web Portal

	AsAnsar	Ansar1
Time-frame	11/2008 - 5/2012	12/2008 - 1/2010
Messages	269,548	29,492
Members	5,034	382
Language	Arabic	English

original-English discussions between jihadists and jihadist sympathizers. We used this subset in all of our experiments. The structured data annotations discussed below are the only data elements not originating from this pre-compiled Ansar AlJihad source.

Data pre-processing and annotation

We collected and pre-processed the Ansar1 data as follows:

1. We read in a sample of raw Ansar1 forum posts and compiled the message text and respective message IDs into an initial corpus. We then automatically removed duplicates (same message ID) and empty documents (no message text) from the corpus.
2. Most posts contain exclusively English text as Asnar1 is the English-language forum for the Ansar AlJihad Network. However, occasional posts include non-English words or phrases; these are commonly Arabic passages from the Koran. In these cases the non-English passages were converted to English using Google Translate [33]. We left slang words written in latin characters intact under the assumption that they were meant to be readable by an English language speaker. For example, "Kuffar" is a derogatory Arabic term for unbeliever.

The Dark Web Portal project does not indicate which messages contain VE recruitment content and which do not. Thus, we manually annotated this information within the data. We provided two independent judges with the following instructions:

1. You have been provided with 192 forum posts sampled from a Jihadist forum.
2. Read each post carefully and determine whether that post has the intent to recruit violent extremists to some group or movement. For the purpose of annotation, violent extremist recruitment is defined as any attempt by a group or individual to recruit, radicalize, or persuade another person into aiding a violent movement aimed at disrupting a legitimate authority.
3. Annotate each post by marking it as either (a) contains violent extremist recruitment, or (b) does not contain violent extremist recruitment.

The forum posts' message text had a wide range of sizes with an average of 246 words among the samples (352-word standard deviation); examples of annotated posts are shown in Table 2. We then used Cohen's κ [36] to validate the labeled data for consistency. Agreement, κ, is the proportion of agreement between the judges after chance agreement has been removed. The value of κ is bounded on [−1,+1] with zero indicating that observed agreement equals chance agreement. Therefore a positive κ indicates non-random agreement between judges, and a negative κ indicates conflicting annotation between judges. The following terms are used to calculate κ:

$\kappa = \frac{p_o - p_c}{1 - p_c}$
p_o = proportion of observations for which judges agree (see Table 3),
p_c = proportion of observations for which agreement is expected by chance (see Table 3)

Table 2 Example text of Ansar1 forum posts and the respective annotations

Annotation	Sample text**
Recruitment	*A Golden chance to join Jihad in Somalia. Abo Dojana invited those who want to participate in jihad to join the militants in Somalia to form what he called a base of martyrdom-seekers who would from there spread to the entire world. Somalia could actually be an ideal base for physical and weapons training...*
Recruitment	*Representing the militant Islamic group Shebab, Abu Mansour makes a pitch for new overseas recruits after praising one militant fighter killed in an apparent ambush. 'So, if you can encourage more of your children and more of your neighbors and anyone around to send people like him to this jihad (holy war), it would be a great asset for us,' he says.*
Not recruitment	*I have now added him as a friend on Facebook. But something tells me that he isn't going to answer to my request. LOL, you had me rolling on the floor man!!!!! So this attack was done my 'Jaish al Mujihadeen' How it that possible? , did they have problems with bounced-checks from the US?*
Not recruitment	*A court in the German city of Koblenz sentenced a German of Pakistani origin to eight years in prison Monday on a conviction of assisting the international Al-Qaeda terror network. The man gave the group financial aid and tried to recruit new members in German territory, according to the indictment*
Not recruitment	*Did Mansoor join the emerat? I heard he is still fighting for Ichkira Republic*

**Incorrect spellings and grammar of original posts have been left as is throughout the table.

Table 3 Agreement matrix of proportions for recruitment categories

		Judge A		
	Category	**No**	**Yes**	p_{iB}
Judge	**No**	0.82 (0.74)*	0.04	0.86
B	**Yes**	0.03	0.11 (0.02)	0.14
	p_{iA}	0.85	0.15	$\sum p_i = 1.00$
	$p_o = 0.82 + 0.11 = 0.93$			
	$p_c = 0.74 + 0.02 = 0.76$		$\kappa = \frac{0.93 - 0.76}{1 - 0.76} = 0.70$	

*Parenthetical values are proportions expected due to chance association.

Table 3 shows the agreement results for our manually annotated Ansar1 data. As shown, the two judges found that approximately 11% of the posts contained VE recruitment. The two judges agreed on the labels for 93% of the posts, with an expected chance agreement of 76%, producing a κ of 70% (see Table 3 for details). Significant non-random agreement was observed with a confidence interval of $(0.5, 0.7)$ at $p = 0.01$; however, interpreting strength of agreement is a common problem with agreement metrics. Some studies have attempted to provide a scale and would describe $\kappa = 0.70$ as "substantial" strength of agreement [37,38], but despite considerable debate among statisticians this issue has never been definitively addressed. Considering both the significance and magnitude of agreement, these results adequately justify using the annotated Ansar1 messages in our analytic approach. To increase the final size of our experimental dataset, one of the judges annotated an additional 100 posts from the Ansar1 collection following the same protocol described above. In total, we observed that 13% of forum posts contained recruitment according to our definition.

Analytic approach

We developed a binary classifier that labels forum posts as either containing or not containing VE recruitment. We used the following probability model:

$$Pr[Recruitment = True \mid d_i] = F[w_1(d_i), ..., w_n(d_i)] \tag{1}$$

In Equation 1, *Recruitment* is a binary classification label, $d_i \in D$ is a forum post, and w_j is a feature function of d_i. In the following section, we discuss the features used and then we present different formulations of the classification function F.

Text classification features

We employed a bag-of-words, or unigram only, feature space by parsing each forum post in the corpus into a term-by-document matrix. This matrix of term frequency (*tf*) features was created using the *RTextTools* and *tm* text mining packages in R [39,40] which also performed basic normalization and feature reduction through the removal of URL web addresses, numbers, punctuation, stopwords, and whitespace. The number of features was further reduced through stemming using the Porter Stemming Algorithm [41]. Under this representation, $w_j(d_i)$ is equal to the raw frequency of a stemmed word form, with n (the number of feature functions) equal to the number of distinct words remaining after document processing.

We then normalized each term frequency (tf_j) and weighted each by its inverse document frequency (IDF_j), producing the logarithmically scaled TF-IDF feature function shown below [42]:

$$w_j(d_i) = \frac{(\log_2(tf_j) + 1) \cdot IDF_j}{\sqrt{\sum_{j'} w_{j'}(d_i)}} \tag{2}$$

where the denominator is a normalization of the feature vector for unit length, and the formula for *IDF* is shown below:

$$IDF_j = \log_2 \left(\frac{|D|}{\sum_{d_i \in D} \mathbb{I}[w_j \in d_i]} \right)$$

$|D|$, corpus cardinality, is the total number of posts in the training corpus, and the denominator represents the number of posts containing at least one occurrence of the *jth* feature (i.e. word). In order to keep the test data unbiased we used *IDF* terms computed only from posts in the training portion of the corpus.

Classification functions

We conducted supervised learning over our annotated posts using a variety of classification functions: naive Bayes, logistic regression, classification trees, boosting, and support vector machines (SVM).

Naive Bayes

Our application of a naive Bayes classifier is described below as an example of how we applied the probability model in Equation 1 to the various classification algorithms mentioned in this section. We calculated the posterior probability of VE recruitment $Pr(Rec_j \mid d_i)$, where $Rec_j \in \{+1, -1\}$, by building upon Bayes' rule and the generic probabilistic model defined above [43]:

$$\begin{aligned} Pr\left[Rec_j \mid \mathbf{w}(d_i)\right] &= \frac{Pr\left[\mathbf{w}(d_i) \mid Rec_j\right] Pr(Rec_j)}{Pr\left[\mathbf{w}(d_i)\right]} \\ &= \frac{Pr\left[\mathbf{w}(d_i) \mid Rec_j\right] Pr(Rec_j)}{\sum_{r \in Rec} Pr\left[\mathbf{w}(d_i) \mid Rec_r\right] Pr(Rec_r)} \end{aligned}$$

The naive Bayes independence assumption reduces the joint probability to the product of component probabilities $Pr[w_k(d_i) \mid Rec_j]$, giving us the posterior probability estimator $F[w_1(d_i), ..., w_n(d_i)]$ from Equation 1. Since the denominator is a constant with respect to class Rec_j, the posterior function F can be further reduced to the following proportion:

$$Pr\left[Rec_j \mid w_1(d_i), \ldots, w_n(d_i)\right] \propto Pr(Rec_j) \prod_{k=1}^{n} Pr\left[w_k(d_i) | Rec_j\right] \tag{3}$$

Our implementation of naive Bayes was adapted from the R package *e1071* for use with the sparse training data typical in a term-by-document matrix [44]. We fit a naive

Bayes model using the default settings of Laplace (add one) smoothing and priors taken from the training data.

Logistic regression

We used our probability model from Equation 1 with a two-class logistic regression model. Given VE recruitment class labels $Rec = \{+1, -1\}$, we applied the following generalized linear model (GLM) using the logit function [45].

$$Pr\left[Rec_j = \pm 1 \mid \mathbf{w}(d_i)\right] = \frac{1}{1 + \exp\left[-Rec_j\left(\beta_0 + \sum_{k=1}^{n} \beta_k \cdot w_k(d_i)\right)\right]}$$

We estimated parameters $\beta_0, \ldots, \beta_k$ from training data by minimizing the L2-regularized log-likelihood:

$$\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + C\sum_{i=1}^{n} \log\left(1 + e^{-Rec_i\,\boldsymbol{\beta}^T\boldsymbol{d_i}}\right) \tag{4}$$

where $C > 0$ is the regularization cost parameter.

We used the *LiblineaR* package in R to minimize Equation 4 and then to predict the VE recruitment classification of testing data [46]. All GLM results shown in the "Results and discussion" section are for L2-regularized logistic regression models fit with the default settings for this R package and an L2-regularization cost parameter C equal to the ratio of negative to positive class labels.

Classification trees

We applied the probability model in Equation 1 to a classification tree by calculating the posterior probability of the recruitment classes at each node of the tree. The R package *tree* was used to train classifiers grown using recursive partitioning with a deviance criterion to select features at each node [47]. We used the default package parameters to control tree growth, including: minimum within-node deviance $= 0.01(deviance_{root})$, minimum allowable node size $= 10$, and minimum observations to a candidate child node $= 5$.

Logit boosting

We applied an ensemble of the tree classifier to this recruitment classification problem using the LogitBoost algorithm implemented in the R package *caTools* [48]. The boosting results shown in the "Results and discussion" section were produced using the package's default weak learner, decision stumps, and 101 boosting iterations. Logit boosting is an application of the original boosting algorithm, AdaBoost, except with the binomial log-likelihood as the minimized loss function (logistic loss) shown below [49]:

$$\sum_{i=1}^{n} log\left(1 + e^{-2Rec_i\,F(d_i)}\right) \tag{5}$$

Support vector machines

Finally, we trained a recruitment classifier using the support vector machine (SVM) algorithm implemented in the R package *e1071* [44]. SVMs do not fit into a probability model like Equation 1; however, the R package provides a method for estimating class probabilities if they are required for things like performance comparisons with receiver operating characteristic (ROC) curves. All SVM results shown below were produced using default package parameters, constraint violation cost $= 100$, and a radial basis function as the kernel.

Results and discussion

To make full use of the annotated data available for training and testing, we randomly segmented the data into ten folds and applied cross-validation. The statistics shown in this section come from the aggregated results of those ten models trained on mutually exclusive training data. We evaluated the classification methods using ROC curves, which show trade-offs between the metrics in the contingency table shown in Table 4. Specifically, ROC curves show the trade-offs between the False Positive Rate (FPR) and True Positive Rate (TPR) at various classification thresholds θ, as generated using Equations 6 and 7. We also employed area under the ROC curve (AUC) to compare each method's performance along the entire curve using a single measure.

$$FPR(\theta) = \frac{FP}{FP + TN} = \frac{FP}{N} \tag{6}$$

$$TPR(\theta) = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{7}$$

A comparison of five VE recruitment classifiers using the annotated Ansar1 data can be seen in Figure 1. These are mean ROC curves averaged over the ten fold cross-validation experiment. Results show all the classifiers performing better than a random-guess model (the diagonal), with the SVM classifier performing best at an AUC of 0.89. A comparison of bootstrap results estimating the

Table 4 Confusion matrix used to assess the recruitment model's classification performance

	Model classification		
Observed classification	$Pr(Rec_j\|d_i) \geq \theta$	$Pr(Rec_j\|d_i) < \theta$	
Recruitment = True	True positives (TP)	False negatives (FN)	$P = TP + FN$
Recruitment = False	False positives (FP)	True negatives (TN)	$N = FP + TN$
	$\hat{P} = TP + FP$	$\hat{N} = FN + TN$	$I = P + N$

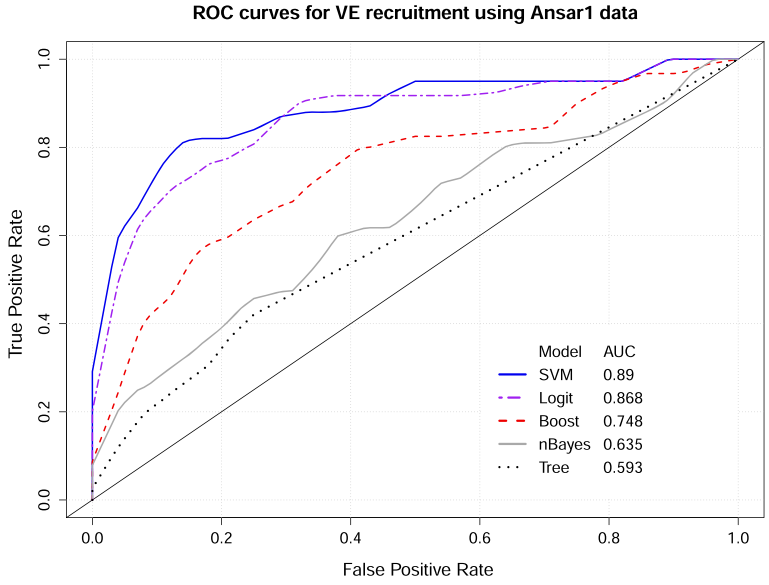

Figure 1 Comparison of VE recruitment classifiers using ROC curves. Curves are averaged over ten cross-validation folds showing classification results for the models.

mean cross-validated AUC can be seen in Figure 2. These bootstrap results were obtained by re-sampling 10,000 times from each testing fold, calculating the AUC on each bootstrapped sample, and then averaging each of those bootstrapped AUCs across the ten cross-validation folds. This resulted in 10,000 mean AUCs for each of the classification models described above. The box-plots provide a similar graphical performance comparison to Figure 1, but also provide a reference for how widely each method's accuracy varies. All the methods range between 0.2 and 0.3 AUC with SVM having the smallest performance variance and boosting having the widest. A statistical comparison of the five classification methods using the bootstrapped mean AUCs is shown in Table 5.

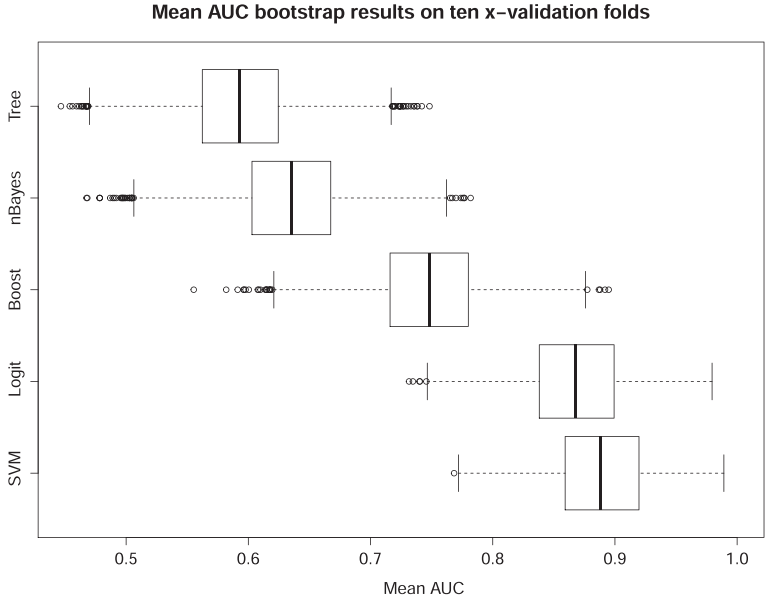

Figure 2 Comparison of VE recruitment classifiers using mean AUC bootstrap results. Box-plots of 10,000 bootstrapped AUCs averaged over ten cross-validation folds.

Table 5 95% confidence intervals for multiple comparisons of bootstrapped mean AUCs using Tukey's range test

	95% CI	
Tukey's test comparisons	**Upper**	**Lower**
SVM – Logit	0.019	0.023
SVM – Boost	0.139	0.143
SVM – nBayes	0.252	0.256
SVM – Tree	0.294	0.297
Logit – Boost	0.119	0.122
Logit – nBayes	0.232	0.235
Logit – Tree	0.273	0.276
Boost – nBayes	0.111	0.115
Boost – Tree	0.153	0.156
nBayes – Tree	0.040	0.043

We used Tukey's range test to determine if the difference in mean AUC between the models is statistically significant [50]. The 95% confidence intervals in Table 5 show a significant difference between each of the classification methods' mean AUC. Therefore, as shown by the ordering of model results in Figure 2 and Table 5, the SVM classifier returned the best performance and classification trees returned the worst mean AUC performance ($p < 0.05$).

The computational complexity of these methods is well documented; however, their runtime performance on this VE recruitment task is not well known. A comparison of time performance benchmarks for two different tasks can be seen in Table 6. The training task results are obtained from trials of each supervised learning algorithm applied to 294 annotated posts per trial. The classification task is the application of each pre-trained classifier to a set of posts resulting in an average classification time per post. Comparing the results in Table 6, we can see that all the methods train in under 30 seconds with SVM, Logistic Regression, and naive Bayes learning two orders of magnitude faster. SVM and logistic regression perform the best on the classification task, but all the methods classify posts in well under a second. Therefore any of these methods is a feasible VE recruitment classifier if the average time between posts is greater than a tenth of a second; a reasonable assumption for many online forums.

Table 6 Time performance benchmark results* (a) training time using 294 posts, and (b) mean classification time per post

	Training (s)	**Classification (ms/post)**
SVM	0.450	0.18
Logit	0.049	0.22
Boost	29.790	0.27
nBayes	0.987	128.57
Tree	11.140	28.18

*Hardware: Intel® Core™ i5 CPU M 480 @ 2.67 GHz; 8 GB RAM.

Typically classification research compares results against prior methods as a benchmark for improvements in accuracy; however, we were unable to find any previously published methods for the specific task of identifying violent extremist recruitment using text classification techniques. Thus, our results serve as initial performance benchmarks against which future methods can be compared.

To provide some understanding of how the best performing classification models are being trained to recognize VE recruitment, Table 7 shows a list of the top-weighted features in the logistic regression model. It is clear from the table that posts relating to the escalating conflicts in Nigeria and Somalia were primary topics in the Ansar1 data—an intuitive finding considering the 2009 timeframe of the sampled data. The importance of such terms hints at the Logit model's potential for over-fitting this particular time period. Running the same model on other time periods would likely produce lower performance scores, since different wording is likely. Other important terms like "jihad", "allah", and words stemming from "milit" exemplify the algorithm's ability to recognize typical features of Islamic violent extremism. Perhaps surprisingly, none of the top ten terms are particularly indicative of recruitment. This may be due to the abundant presence of terms like "recruit" and "join". These terms are among the top 30 most frequent, appearing in our corpus 178 and 126 times respectively. High frequency makes these terms more likely to occur in both recruitment and non-recruitment posts and therefore diminishes their discriminating power. Regardless of how they work, performance metrics show that the best models (SVM and Logit) detect VE recruitment with considerable accuracy (mean $AUC > 0.85$).

Conclusions and future work

This work was motivated by increasing online activities of violent extremist organizations along with the lack of automated approaches to analyze such activity.

Table 7 The most discriminating term features as weighted by the cross-validated logistic regression models

Feature	**Weight**	**Feature**	**Weight**
nigeria	0.90	jihad	0.61
hamas	0.72	alandalus	0.61
foreign	0.67	milit (ant,ary,...)	0.60
somalia	0.66	american	0.60
may	-0.65	allah	-0.54

Our research built upon recent data collection and analysis efforts to develop supervised learning and natural language processing methods that automatically identify cyber-recruitment by violent extremists. The results presented in this article support the conclusion that automatic VE recruitment detection is a feasible goal. As the first reported results on this task, our classifiers serve as initial performance benchmarks against which future VE recruitment classifiers can be compared.

In the future, our VE recruitment detection methods could be improved by including support for non-English languages. Whether such future methods use automatic translation or non-English features, support for other languages is an important task considering that violent extremist groups frequently operate in non-English speaking communities. Incorporating non-English text and features could be accomplished through the use of experts to perform the manual annotation. Expert judges might also improve annotation quality if agreement remains strong. Future work could also analyze classifier behavior in depth, and test the effectiveness of more advanced feature selection and modeling techniques. Methods like latent semantic analysis perform singular value decomposition transformations on the feature space and may be employed to further reduce both dimensionality and the effect of non-discriminating terms [51]. Latent Dirichlet allocation may be used to substitute the terms in a high-dimensional feature space with a smaller set of latent topics that represent the major subjects appearing in the corpus [52]. Such latent variable modeling techniques could serve as feature selection and replacement methods while preserving the statistical relationships that are essential for text classification tasks.

By testing the effectiveness of our methods in a proxy for real-world settings we demonstrated that such automated classification tools would clearly fit into the workflow of counterterrorism intelligence teams like the FBI's information review analysts. The current workflow tasks human analysts with manually reviewing and annotating "the ever-increasing [volume of investigative] information" stored in data warehouses like the Electronic Surveillance Data Management System (DWS-EDMS) used by the FBI [6]. An automated classification system using our methods for detecting VE recruitment could serve as a pre-screening step in the current review workflow tasked with reducing the volume of documents requiring human attention. Our automated approach could also complement current lead management systems like eGuardian by automatically detecting potential terrorist recruitment events so they can be efficiently compiled into leads for current investigations or used as evidence to open new terrorism-related investigations.

More generally our automated classification methods could be used as part of a VE recruitment identification and tracking methodology that would enable the study of recruitment efforts and the membership dynamics of violent organizations. Such a method might be able to measure the effectiveness of extremist and counterinsurgency efforts on new membership by correlating specific recruitment activities and current events with changes in the VE population of a community. As a future research path, this proposed methodology requires (1) an automated system for classifying whether a forum user is a member of a violent extremist group, and (2) time series methods for analyzing recruitment and membership along a timeline.

In light of the still unfolding news regarding the NSA's Boundless Informant and PRISM programs [53], we address some ethical implications of our work. Given that such a comprehensive and intrusive source of text data does exist, there is clearly a potential for abusing a recruitment and membership classification method to target non-combative individuals. Such tracking methods could thwart perfectly legal recruitment efforts of peaceful protesters, or radical yet law-abiding religious sects. These groups might fit the profile of a VE organization in every way except the critical ingredient of violence. Furthermore, recruitment alone rarely necessitates a violent act even though a recruiter may refer to or even encourage such acts. Because of these possible unethical repercussions, we proposed classification methods that target not just extremist groups, but specifically violent groups engaged in acts like terrorism. We hope that tuning the learning algorithms in this way will reduce some risk of misuse.

Abbreviations

AUC: Area under the curve (ROC curve); Boost: Boosting algorithm; COIN: Counterinsurgency; FPR: False positive rate; GLM: Generalized linear model; IDF: Inverse document frequency; Logit: Logistic regression algorithm; nBayes: Naive Bayes algorithm; Rec: Recruitment (of violent extremists); ROC: Receiver operating characteristic; SNA: Social network analysis; SVM: Support vector machines; tf: Term frequency; TPR: True positive rate; VE: Violent extremist.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JS developed the recruitment detection methodology, implemented and tested the learning algorithms used in this paper, and drafted the manuscript. MG provided theoretical guidance in the whole procedure and revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

This research was financially supported by a grant from the United States Army Research Laboratory (ARL).

Received: 16 January 2014 Accepted: 29 May 2014
Published online: 13 August 2014

References

1. LA Overbey, G McKoy, J Gordon, S McKitrick, Automated sensing and social network analysis in virtual worlds, in *Intelligence and Security Informatics (ISI)* (IEEE Vancouver, BC, Canada, 2010), pp. 179–184
2. R Torok, "Make A Bomb In Your Mums Kitchen": Cyber Recruiting And Socialisation of 'White Moors' and Home Grown Jihadists, in *Australian*

Counter Terrorism Conference (School of Computer and Infomation Science, Edith Cowan University Perth, Western Australia, 2010), pp. 54–61
3. M Rogers, ed. by A Silke, Chapter 4: The Psychology of Cyber-Terrorism, in *Terrorists, Victims and Society: Psychological Perspectives on Terrorism and its Consequences* (John Wiley & Sons Chichester, West Sussex, England, 2003), pp. 77–92
4. S O'Rourke, Virtual radicalisation: Challenges for police, in *8th Australian Information Warfare and Security Conference* (School of Computer and Infomation Science, Edith Cowan University Perth, Western Australia, 2007), pp. 29–35
5. S Mandal, E-P Lim, Second life: Limits of creativity or cyber threat, in *IEEE Conference on Technologies for Homeland Security* (IEEE Waltham, MA, 2008), pp. 498–503
6. WH Webster, DE Winter, L Adrian, J Steel, WM Baker, RJ Bruemmer, KL Wainstein, Final report of the William H.Webster Commission on the Federal Bureau of Investigation, counterterrorism intelligence, and the events at Fort Hood, Texas on November 5, 2009. Technical report, Federal Bureau of Investigation (2012)
7. RR Tomes, Waging war on terror relearning counterinsurgency warfare. Parameters. **34**(1), 16–28 (2004)
8. F Gutiérrez, Recruitment in a civil war: a preliminary discussion of the colombian case, in *Santa Fe Institute, Mimeo*, (2006)
9. M Humphreys, JM Weinstein, Who fights? the determinants of participation in civil war. Am. J. Pol. Sci. **52**(2), 436–455 (2008)
10. MI Lichbach, *The Rebel's Dilemma*. (University of Michigan Press, Ann Arbor, 1998)
11. K Peters, P Richards, 'Why we fight': Voices of youth combatants in Sierra Leone. Africa. **68**(02), 183–210 (1998)
12. JM Weinstein, *Inside Rebellion: The Politics of Insurgent Violence*. (Cambridge University Press, New York, 2007)
13. RD Petersen, *Resistance and Rebellion: Lessons from Eastern Europe*. (Cambridge University Press, New York, 2001)
14. S Popkin, The rational peasant. Theory Soc. **9**(3), 411–471 (1980)
15. JC Scott, *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. (Yale University Press, New Haven & London, 1976)
16. EJ Wood, *Insurgent Collective Action and Civil War in El Salvador*. (Cambridge University Press, New York, 2003)
17. RW McGehee, *Deadly Deceits: My 25 Years in the CIA*. (Sheridan Square Publications, Inc., New York, 1983), pp. 95–116
18. M Conway, Terrorism and the internet: new media–new threat? Parliamentary Aff. **59**(2), 283–298 (2006)
19. R Torok, Developing an explanatory model for the process of online radicalisation and terrorism. Secur. Informatics. **2**(1), 1–10 (2013)
20. L Bowman-Grieve, A psychological perspective on virtual communities supporting terrorist & extremist ideologies as a tool for recruitment. Secur. Informatics. **2**(1), 1–5 (2013)
21. EF Kohlmann, Al-Qaida's MySpace: terrorist recruitment on the internet. CTC Sentinel. **1**(2), 8–9 (2008)
22. LA Overbey, G McKoy, J Gordon, S McKitrick, Jr MH, L Buhler, L Casassa, S Yaryan, Virtual DNA: Investigating cyber-behaviors in virtual worldsL. Technical Report 33-09 E, Space and Naval Warfare System Center Atlantic Charleston, SC (2009)
23. GS McNeal, Cyber embargo: Countering the internet jihad. Case West. Reserv. Univ. J. Int. Law. **39**, 789–826 (2008)
24. H Chen, S Thoms, T Fu, Cyber extremism in web 2.0: An exploratory study of international jihadist groups, in *IEEE International Conference on Intelligence and Security Informatics (ISI)* (IEEE Taipei, 2008), pp. 98–103
25. M Yang, M Kiang, H Chen, Y Li, Artificial immune system for illicit content identification in social media. J. Am. Soc. Inf. Sci. Technol. **63**(2), 256–269 (2012)
26. A Basu, Social network analysis of terrorist organizations in India, in *North American Association for Computational Social and Organizational Science (NAACSOS) Conference* (NAACSOS Notre Dame, Indiana, 2005), pp. 26–28
27. KM Carley, Destabilization of covert networks. Comput. Math. Organ. Theory. **12**(1), 51–66 (2006)
28. M Chau, J Xu, Using web mining and social network analysis to study the emergence of cyber communities in blogs, in *Terrorism Informatics* (Springer New York, 2008), pp. 473–494
29. J Diesner, KM Carley, Using network text analysis to detect the organizational structure of covert networks, in *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference* (NAACSOS Pittsburgh, 2004)
30. Z Chen, B Liu, M Hsu, M Castellanos, R Ghosh, Identifying intention posts in discussion forums, in *Proceedings of NAACL-HLT* (Association for Computational Linguistics Atlanta, Georgia, 2013), pp. 1041–1050
31. H Chen, W Chung, J Qin, E Reid, M Sageman, G Weimann, Uncovering the dark web: A case study of jihad on the web. J. Am. Soc. Inf. Sci. Technol. **59**(8), 1347–1359 (2008)
32. T Fu, A Abbasi, H Chen, A focused crawler for dark web forums. J. Am. Soc. Inf. Sci. Technol. **61**(6), 1213–1231 (2010)
33. Google Inc., Google Translate (2014). http://translate.google.com/
34. H Chen, E Reid, J Sinai, A Silke, B Ganor (eds.), *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*. (Springer, New York, 2008)
35. Artificial Intelligence Laboratory, University Of Arizona, Dark Web Forum Portal: Ansar AlJihad Network English Website (2014). http://cri-portal.dyndns.org
36. J Cohen, A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
37. JR Landis, GG Koch, The measurement of observer agreement for categorical data. Biometrics. **33**(1), 159–174 (1977)
38. JL Fleiss, B Levin, MC Paik, The measurement of interrater agreement. Stat. Methods Rates Proportions. **2**, 212–236 (1981)
39. I Feinerer, K Hornik, *Tm: Text Mining Package*. (R Foundation for Statistical Computing, 2014). R package version 0.5-10. http://CRAN.R-project.org/package=tm
40. TP Jurka, L Collingwood, AE Boydstun, E Grossman, W van Atteveldt, RTextTools: Automatic Text Classification Via Supervised Learning (2014). R package version 1.4.2. http://CRAN.R-project.org/package=RTextTools
41. CJ Van Rijsbergen, SE Robertson, MF Porter, *New Models in Probabilistic Information Retrieval*. (British Library Research and Development Dept, 1980)
42. W Cavnar, ed. by DK Harman, Using an n-gram-based document representation with a vector processing retrieval model, in *Overview of the Third Text Retrieval Conference* (Computer Systems Laboratory, National Institute of Standards and Technology Gaithersburg, MD, 1995), pp. 269–277
43. RO Duda, PE Hart, DG Stork, *Pattern Classification*, 2nd edn. (John Wiley & Sons, Inc, New York, 2001), p. 62
44. D Meyer, E Dimitriadou, K Hornik, A Weingessel, F Leisch, E1071: Misc Functions of the Department of Statistics (e1071), TU Wien (2014). R package version 1.6-2. http://CRAN.R-project.org/package=e1071
45. C-J Lin, RC Weng, SS Keerthi, Trust region newton method for logistic regression. J. Mach. Learn. Res. **9**, 627–650 (2008)
46. T Helleputte, LiblineaR: Linear Predictive Models Based On The Liblinear C/C++ Library (2013). R package version 1.80-7. http://CRAN.R-project.org/web/packages/LiblineaR
47. B Ripley, Tree: Classification and Regression Trees (2014). R package version 1.0-35. http://CRAN.R-project.org/package=tree
48. J Tuszynski, caTools: ROC AUC Tools, Moving Window Statistics (2013). R package version 1.16. http://CRAN.R-project.org/package=caTools
49. J Friedman, T Hastie, R Tibshirani, Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**(2), 337–407 (2000)
50. AP Fenech, Tukey's method of multiple comparison in the randomized blocks model. J. Am. Stat. Assoc. **74**(368), 881–884 (1979)
51. S Deerwester, ST Dumais, GW Furnas, TK Landauer, R Harshman, Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
52. DM Blei, AY Ng, MI Jordan, Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
53. G Greenwald, NSA Collecting Phone Records of Millions of Verizon Customers Daily (2013). http://www.guardian.co.uk/world/2013/jun/06/nsa-phone-records-verizon-court-order

doi:10.1186/s13388-014-0005-5
Cite this article as: Scanlon and Gerber: **Automatic detection of cyber-recruitment by violent extremists.** *Security Informatics* 2014 **3**:5.