Automatic Detection and Forecasting of Violent Extremist Cyber-Recruitment

---

A Thesis
Presented to
the faculty of the School of Engineering and Applied Science
University of Virginia

In partial fulfillment

of the requirements for the degree

Master of Science in Systems Engineering

by

Jacob R. Scanlon

May 2014

Approval Sheet

The thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science in Systems Engineering.

Jacob R. Scanlon
Author

The thesis has been read and approved by the examining committee:

Prof. Donald E. Brown, Committee Chair
Systems and Information Engineering

Prof. Matthew S. Gerber, Advisor
Systems and Information Engineering

Prof. John M. Owen
Politics

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science
May 2014

Acknowledgments

I would like to thank my advisors Don Brown and Matt Gerber for their continued help over the past two years. You have been wonderful teachers, mentors, and friends. I would also like to thank my parents, Robert and Ilona, and my fiancé Janie. Without your support and encouragement I could never have come this far.

This research was funded by a grant from the Army Research Laboratory (ARL).

Outline

List of Tables

List of Figures

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 violent extremists 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 research presents methods for identifying and forecasting the recruitment activities of violent groups within extremist social media websites. Specifically, these methods employ techniques within supervised learning and natural language processing for automatically: (1) identifying forum posts intended to recruit new violent extremist members, and (2) forecasting recruitment efforts by tracking changes in an online community's discussion over time. 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; the confidence interval of Cohen's κ was $(0.5, 0.9)$ at $p = 0.01$. We used naive Bayes models, logistic regression, classification trees, boosting, and support vector machines (SVM) to classify the forum posts in a 10-fold cross-validation experimental setup. Evaluation with receiver operating characteristic (ROC) curves shows that our SVM classifier achieves 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 $p = 0.05$). The forecasting task uses time series regression analysis to model the daily count of extremist recruitment posts. Evaluation with mean absolute scaled error (MASE) shows that employing latent topics as predictors can reduce forecast error compared to a naive (random-walk) model and the baseline time series model. To our knowledge, these are the first results reported on these tasks, and our analysis indicates that automatic detection and forecasting of online terrorist recruitment are feasible tasks. This research could ultimately help identify the impact of violent organizations, like terrorist groups, within the social net-

work of an online community. There are also 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.

1 Introduction

1.1 Background

In the last decade, the modern landscape of extremism has expanded to encompass the Internet and online social media [47, 57]. 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 [57]. Terrorist groups use the free and open nature of the Internet to form online communities [54] and disseminate literature and training materials without having to rely on traditional media outlets that might censor or change their message [46, 57]. Terrorist organizations engage in directed communication and advertisement, recruiting members on social websites like Second Life, Facebook, and radicalized religious web forums [41, 47, 57]. 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 [61]. 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 [61]. Considering this large volume of possibly relevant text data requiring review by a limited number of personnel, automated methods would be useful for pre-screening text documents and reducing the workload of human analysts. Automated forecasting methods would also be useful for anticipating future workload and staffing requirements. Knowing the amount of recruitment activity likely to occur in the near future might allow intelligence experts to assign additional resources before analysts are overwhelmed by a surge in activity.

For this research, a **violent extremist (VE)** group is an organization that uses violent means to disrupt a legitimate authority. **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 thesis, **VE recruitment** is any attempt by a group or individual involved in VE to recruit, radicalize, or persuade another person to aid a violent extremist movement. VE cyber-recruitment is therefore VE recruitment activity that makes use of computers and the Internet.

1.2 Problem Statement

This research seeks automated methods for identifying and forecasting recruitment activities of violent extremist organizations within online social media. Specifically, this thesis has two research objectives, which are described below.

1. Develop methods that automatically identify messages recruiting individuals for participation in violent extremist groups. For these classification purposes, a **VE cyber-recruitment** message is any online 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, this thesis provides a novel method for identifying the incitement of violent activity within online communities.

2. Develop methods that automatically forecast the level of VE recruitment activity conducted in an online community over time. Automated VE recruitment forecasting methods will track online recruitment posts and predict the magnitude of future recruitment activity within a targeted cyber-community. Such a method might provide insight into the relationship between VE recruitment activity in a forum, other local insurgent activity, and even news

about external events. This research objective complements the first objective, and together they produce an integrated analytic framework compiling intelligence on current VE recruitment posts and the magnitude of future recruitment activities. Ultimately, addressing these objectives could help intelligence experts to allocate scarce resources (e.g., personnel) more efficiently and respond more effectively to VE recruitment efforts.

2 Literature Review

2.1 Offline Recruitment and Manual Social Network Analysis

The modern jihadist insurgencies in Iraq and Afghanistan operate among local civilian populations and engage in both legal and illegal activities in order to achieve their strategic and political goals [56]. However, the illegal acts are only effective when carried out by an organized and well-manned group [56]. 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 lifespans must replace members lost through arrests, deaths, and defections [54]. Several studies have tried to understand why some people join violent rebellions [28, 31, 38, 49, 62], while others only sympathize or cooperate in a non-violent capacity [50, 51, 55, 64]. This thesis 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 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 on 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 [42]. Although this thesis 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.

2.2 Cyber-Recruitment, Social Network Analysis, & 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 [16, 57, 58]. Counterinsurgency (COIN) experts are increasingly concerned with the potential use 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 [7, 36, 41, 46, 48, 58]. Recent research has evaluated the use of political tools for shutting down websites or shaming material supporters [43]. Some researchers have suggested the use of web-crawling and analysis techniques to monitor for VE activities including recruitment [13, 57, 65]; however, there do not appear to be any implementations of such techniques on recruitment specifically. This thesis 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 [3, 8, 10, 18, 47]. With objectives similar to McGehee's manual SNA work, contemporary research aims 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 [14]. 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 a preliminary way 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. This thesis 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 and forecast recruitment activities.

3 Identification of VE Cyber-Recruitment

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 [11, 13, 24, 47, 48]. This thesis builds on these efforts and develops automatic methods to (1) identify and (2) forecast the online recruitment activity of violent extremist (VE) groups. This section presents the methods and results for the first objective, identifying VE cyber-recruitment, while Section 4 addresses the second objective, forecasting VE recruitment. In Section 3.1.1, we define requirements that must be met by data sources supporting our objectives and describe specific data sources used in our study. In Section 3.1.2, we describe our manual annotation effort, which analyzed individual posts for recruitment content. In Sections 3.2 and 3.3, we define a probabilistic model employing natural language features for automatically classifying VE recruitment in forum posts, detail the classification functions used in our supervised learning experiments, and then interpret the results obtained from the experiments.

3.1 Data Collection and Annotation

3.1.1 Data Requirements and Sources

This research 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.

Time-frame coverage: The data time span should be wide enough to capture time series effects like seasonality and current enough to be relevant to contemporary anti-extremist efforts. As a result, training and testing corpora must be sampled from a dataset covering at least one continuous year within the last decade.

Language: The collected data must use the English language or be translatable to English using an automatic process like Google's machine translation service [25].

We identified the Dark Web Portal Project [11, 12] 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 [1]. The Dark Web Project compiled 299,040 total messages posted on Ansar AlJihad between 2008-2012. Fewer posts

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

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

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, 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.

3.1.2 Data Pre-processing and Annotation

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

1. We read in all 29,492 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), retaining 28,744 posts in 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 typically Arabic passages from the Koran. In these cases the non-English passages were converted to English using Google Translate [25]. 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.

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.

We provided two independent judges with the following instructions:

1. You have been provided with 192 forum posts randomly 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 annotated 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 [15] 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 agreement beyond chance between judges, and a negative κ indicates disagreement beyond chance 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)

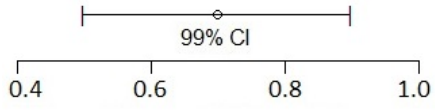
Table 3 shows the agreement results for our manually annotated Ansar1 data. As shown, the two judges found that approximately 11% of the posts contained VE recruitment. The two judges agreed on the labels for 93% of the posts, with an expected chance agreement of 76%, producing a κ of 70%. Significant non-random agreement was observed with a confidence interval of $(0.5, 0.7)$ at $p = 0.01$; however, interpreting the strength of agreement is a common problem with agreement metrics. Some studies have attempted to provide a

Table 3: Agreement matrix of proportions for recruitment categories [15].

		Judge A		
	Category	No	Yes	p_{iB}
Judge B	No	**0.82** (0.74)*	0.04	0.86
	Yes	0.03	**0.11** (0.02)	0.14
	p_{iA}	0.85	0.15	$\sum p_i = 1.00$

$p_o = 0.82 + 0.11 = 0.93$

$p_c = 0.74 + 0.02 = 0.76$ $\quad \kappa = \frac{0.93-0.76}{1-0.76} = 0.70$



* Parenthetical values are proportions expected due to chance association.

scale and would describe $\kappa = 0.70$ as "substantial" agreement [22, 37]. Considering both the significance and magnitude of agreement, we proceeded with our analysis using the annotated Ansar1 messages described above. 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.

3.2 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:

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

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.

3.2.1 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 [20, 35, 52], 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 [60]. 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 [9]:

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

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

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

$|D|$, 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 values computed only from posts in the training portion of the corpus.

3.2.2 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 $Pr(Rec_j \,|\, d_i)$, where $Rec_j \in \{+1, -1\}$, by building upon Bayes' rule and the generic probabilistic model defined above [19]:

$$\begin{aligned} Pr[Rec_j \,|\, \mathbf{w}(d_i)] &= \frac{Pr[\mathbf{w}(d_i) \,|\, Rec_j]\, Pr(Rec_j)}{Pr[\mathbf{w}(d_i)]} \\ &= \frac{Pr[\mathbf{w}(d_i) \,|\, Rec_j]\, Pr(Rec_j)}{\sum\limits_{r \in Rec} Pr[\mathbf{w}(d_i) \,|\, Rec_r]\, Pr(Rec_r)} \end{aligned}$$

The naive Bayes independence assumption reduces the joint probability $Pr[w_1, ..., w_n \,|\, Rec_j]$ to the product of component probabilities $Pr[w_k(d_i) \,|\, Rec_j]$, 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:

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

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 [45]. 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 [40].

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

We estimated parameters $\beta_0, ..., \beta_k$ 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 [30]. All GLM results shown in Section 3.3 are for L2-regularized logistic regression models fit with the default settings for this R package and a tuned 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 [53]. 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* [59]. The boosting results shown in Section 3.3 were produced using the package's default weak learner, decision stumps, and 101 boosting iterations. Logit boosting is an application of the original boosting

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

	Model Classification		
Observed Classification	$Pr(Rec_j \mid d_i) \geq \theta$	$Pr(Rec_j \mid d_i) < \theta$	
Recruitment = True	True Positives (TP)	False Negatives (FN)	$P = TP + FN$
Recruitment = False	False Positives (FP)	True Negatives (TN)	$N = FP + TN$
	$\hat{P} = TP + FP$	$\hat{N} = FN + TN$	$I = P + N$

algorithm, AdaBoost, except with the binomial log-likelihood as the minimized loss function (logistic loss) shown below [23]:

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

Support Vector Machines

Finally, we trained a recruitment classifier using the support vector machine (SVM) algorithm implemented in the R package *e1071* [45]. 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.

3.3 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

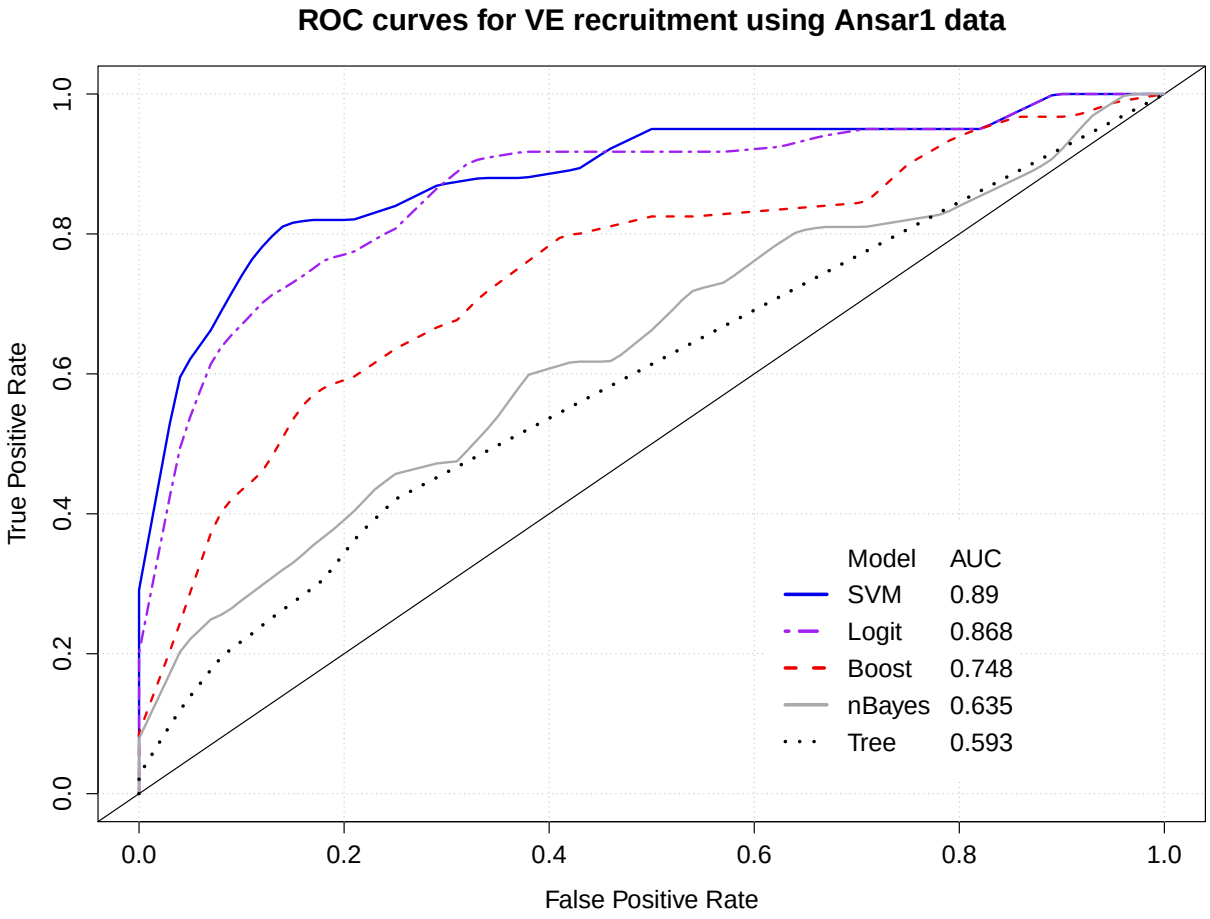


Figure 1: Comparison of VE recruitment classifiers using ROC curves. Curves are averaged over ten cross-validation folds showing classification results for the models.

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

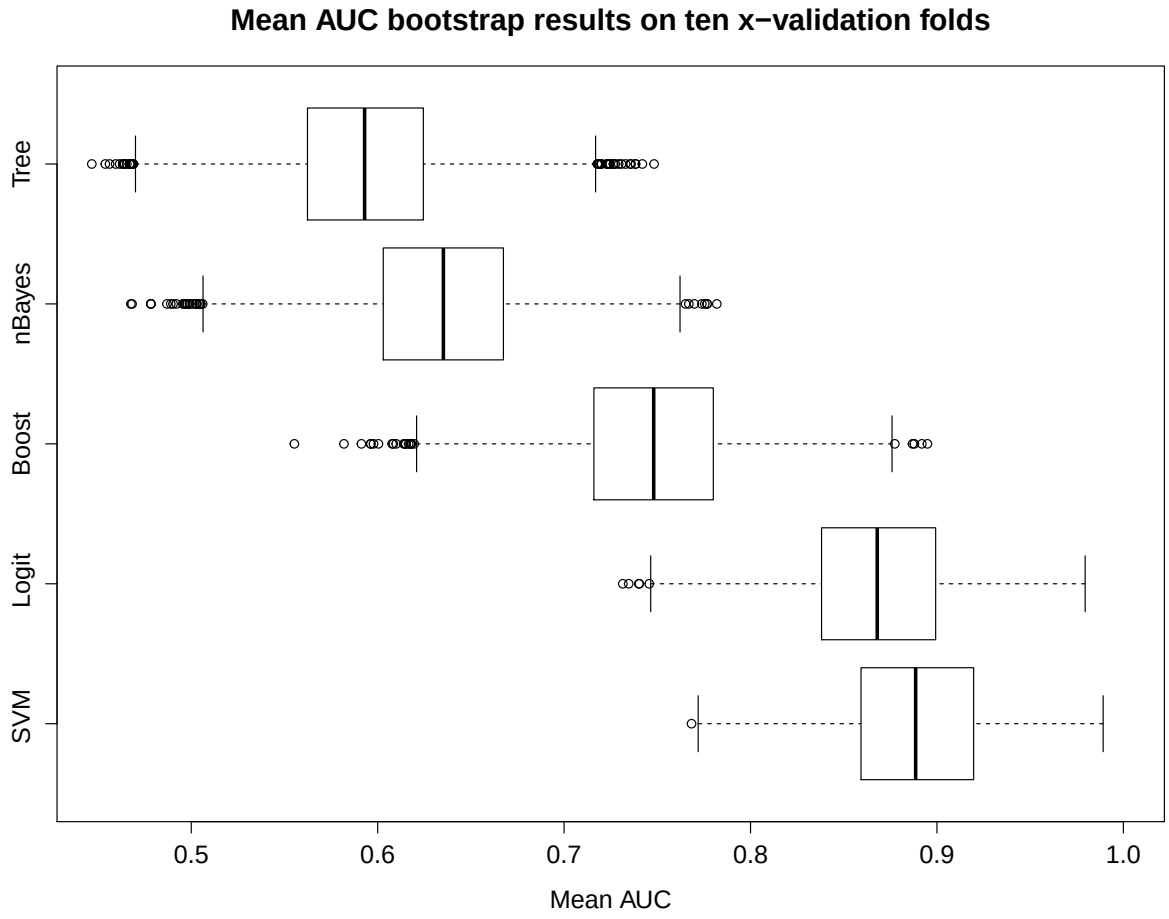


Figure 2: Comparison of VE recruitment classifiers using mean AUC bootstrap results. Box-plots of 10,000 bootstrapped AUCs averaged over ten cross-validation folds.

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

Tukey's Test	95% CI	
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

(the diagonal), with the SVM classifier performing best at an AUC of 0.89. A comparison of bootstrap results estimating the 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. We used Tukey's range test to determine if the difference in mean AUC between the models was statistically significant [21]. 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 [2, 4, 34, 39]; however, the runtime performance on this VE recruitment task is not well known. A comparison

Table 6: Time performance benchmark results: training time using 294 posts and 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 @ 2.67GHz; 8 GB RAM

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.

Classification research typically 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 related 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

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

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 and appear in our annotated 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 (reflected in low IDF scores). Regardless of how they work, performance metrics show that the best models (SVM and Logit) detect VE recruitment with considerable accuracy (mean $AUC > 0.85$).

4 Forecasting VE Cyber-Recruitment

The methods developed in this section build on the classification results and models developed above in order to forecast the magnitude of VE recruitment efforts in an online community. In Section 4.1, we describe the time series recruitment data along with the pre-processing steps used to obtain the recruitment response. In Section 4.2, we define our time series regression analytic approach and describe the text features and models used in our VE recruitment forecasting experiments. Finally, in Section 4.3, we compare the results

obtained from the forecasting experiments against our forecasting benchmarks.

4.1 Time Series Data

We modeled two separate response variables representing the amount of VE recruitment per time period: (1) the number of VE recruitment posts per day, and (2) the percentage of posts containing VE recruitment per day. Figure 3 shows a timeline of these response variables compared to the total amount of activity on the Ansar1 forum during the 2009 calendar year. In the following sections, we discuss the time series data, required pre-processing steps, and then we present different analytical approaches to the time series regression model used to forecast VE recruitment.

Time Series Data Pre-processing

We collected and pre-processed the time series data as follows:

1. We used the best-performing VE recruitment classifier (the SVM) developed in Section 3.2 to classify the non-annotated Ansar1 forum posts previously collected and processed in Section 3.1.2, merging them with the sample of annotated forum posts. This resulted in a total of 28,744 posts in the timeline.

2. We sorted the classified posts in ascending order by date and compiled a count of the number of posts and the number of recruitment posts each day. These daily counts were used to derive the continuous response variables described above (# recruitment, % recruitment).

3. Due to the large data gaps seen in Figure 3, we created a continuous set of time series data by removing all posts from dates prior to January 14, 2009.

Due to the time series nature of the data and the need to preserve the data order, we used a moving horizon to train the regression models and calculate forecasting error on unbiased

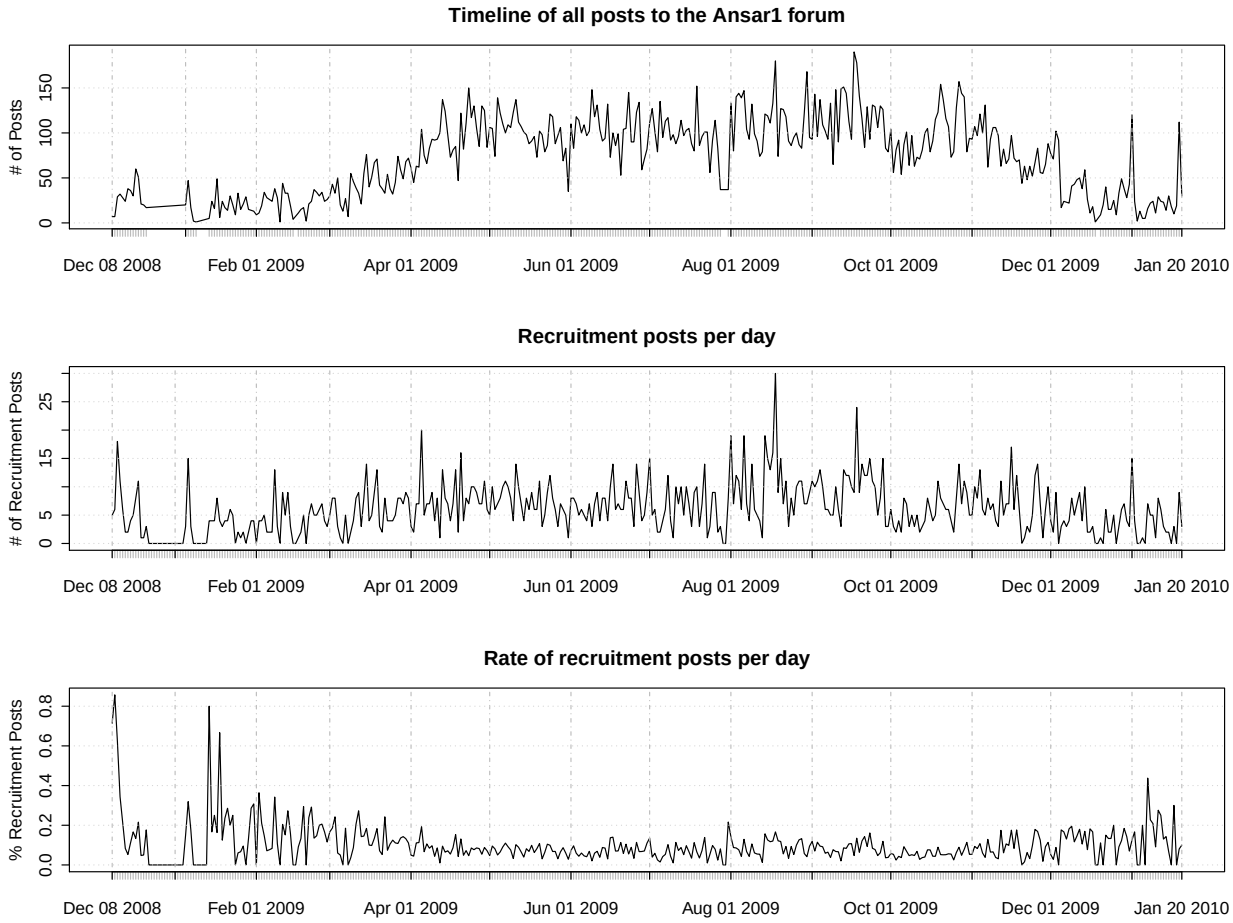


Figure 3: Timeline of forum posts compared to the response variables. A graphical comparison of (top) Ansar1 forum activity showing data gaps, (middle) the daily count of VE recruitment posts, and (bottom) the daily percentage of VE recruitment posts.

test observations. In our moving horizon, models are iteratively trained on the first $t-1$ time periods and tested on time period t until the last time period is reached; January 20, 2010 in our case.

4.2 Analytic Approach

Building on results obtained from the recruitment classifiers described in Section 3.2, we developed methods to forecast the count and percentage of VE recruitment posts occurring within a future time period. We used the following general regression model as a basis for this analytic approach:

$$E[Recruitment_t \mid d_{t-1}] = F[w_1(d_{t-1}), ..., w_n(d_{t-1})] \tag{8}$$

In contrast to classification's binary label in Equation 1, $Recruitment_t$ in Equation 8 is a continuous response representing the amount of VE recruitment expected at the current time period. Since forecasting is performed on discrete time periods rather than individual forum posts, $d_{t-1} \in D$ represents all forum post within the previous time periods, and w_j represents a feature function of d_{t-1}. In the following sections, we discuss the features used in forecasting and present different formulations of the regression function F.

4.2.1 Text and Topic Model Features

Since the forecasting approach uses discrete time periods rather than individual forum posts, we grouped the posts by day and concatenated the message text of each post into daily documents comprising all text posted on that day. These daily forum post documents are used in the topic model features described in this section.

In some of the regression models described below in Section 4.2.2, we employ a set of latent topic variables generated from the daily text as the predictors, $\mathbf{w}(d_{t-1})$, of VE recruitment. A natural language processing technique known as topic modeling was used to produce the set of latent variables. Similar to the classification task, our topic models

used a bag-of-words, or unigram-only, feature space by parsing the daily forum posts 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, which also performed basic normalization and feature reduction through the removal of URL web addresses, numbers, punctuation, stopwords, and whitespace [20, 35, 52]. The number of term-frequency features was further reduced through stemming using the Porter Stemming Algorithm [60].

Latent Dirichlet allocation (LDA) was used to substitute unigram terms in the high-dimensional feature space for a smaller set of latent topics that represent the major subjects appearing in the corpus [6, 27]. We generated $k = 30$ latent topics using the LDA algorithm implemented in the R package *topicmodels*, with the default prior topic weight of $\alpha_k = 1.67$ for all topics. Such latent variable modeling techniques serve as feature selection and replacement methods while preserving the statistical relationships that are essential for time series forecasting.

4.2.2 Time Series Models and Regression Functions

We conducted supervised learning over our moving horizon of training data using a variety of time series forecasting and regression techniques, including: naive model, autoregressive integrated moving average (ARIMA), ordinary least squares (OLS), and principle components regression (PCR).

Naive Model

Our naive model, also called a random-walk method, is a trivial forecasting method in which the forecast for each future period is the actual value for the current period. It is called a naive model because it assumes that there is no change from one period to the next. We use this naive model as a basic benchmark for comparison against more advanced forecasting techniques.

Autoregressive Integrated Moving Average (ARIMA)

We used our regression model from Equation 8 with an ARIMA model using the previous VE recruitment response values as predictors of future daily recruitment posts. Prior to fitting the model, integration (differencing) was applied if the time series data were not stationary (i.e. centered at zero). Given the stationary prior recruitment response values $Rec_1, ..., Rec_{t-1}$, we applied the following ARMA model with p autoregressive (AR) terms and q moving-average (MA) terms [63]:

$$E[Rec_t \,|\, \mathbf{Rec}_{1,...,t-1}] = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i Rec_{t-i} + \sum_{i=1}^{q} m_i \varepsilon_{t-i} \tag{9}$$

where c is a constant intercept term, $\varepsilon_t, \varepsilon_{t-1}, ...$ are residual error terms after removing trend and seasonality, and α_i, m_i are the parameters for the autoregressive and moving-average portions of the model, respectively.

We used the *forecast* package in R to automatically select the optimal number of AR and MA terms and then fit the model using OLS regression [32]. With the recruitment data used as predictors, only one order of differencing was required to make the time series stationary, and no significant seasonal patterns were detected. All results labeled *Baseline ARIMA* in Section 4.3 were produced with an ARIMA model fit with prior recruitment response values. In addition to the naive model, this basic ARIMA time series model is another benchmark against which more advanced models are compared to determine whether additional predictors, like latent topics, improve forecasting accuracy.

Principle Components Regression (PCR)

We applied the general regression model in Equation 8 to our recruitment forecasting problem using the probabilities of the 30 topics generated with LDA as our feature space $\mathbf{w}(d_{t-1})$. We used PCR to fit this model in order to address problems with multicollinearity and feature selection that occurred when using basic fitting methods like OLS. Our implementation of PCR used the principle components of the 30 latent topics as regressors, thereby eliminating

multicollinearity caused by regressing the highly correlated topics directly. Additionally, we employ cross-validation to determine the optimal number of components, thus further reducing model dimensions while retaining the highest variance components [29]. The PCR results shown in Section 4.3 were produced using the R package *pls* which generated the principle components of the 30 latent topics and selected the optimal number of components at each iteration of the moving horizon [44]. The OLS results are also shown in Section 4.3 as a comparison to PCR.

Time Series on Topic Model Features

Finally, we combined time series modeling techniques with the latent topic features to produce a recruitment forecasting method we call "topic time series modeling." In each iteration of the moving horizon the following actions were performed:

1. We fit a PCR model using the 30 latent topics as described above. However, this model is designed to predict the current day's VE recruitment using the current day's topic probabilities instead of the prior topics.

2. We generate the current day's 30 topic probabilities using time series models of the previous $t-1$ days' latent topics. In this step, an ARIMA model for each of the 30 latent topics is automatically produced from the prior data and then used to predict its respective topic probability for the current time period.

3. We then forecast the current time period's recruitment by using the 30 predicted topic probabilities from step 2 as inputs to the model produced in step 1.

The results labeled *Topic.ts* in Section 4.3 below are all produced using this topic time series modeling method.

Table 8: MASE and RMSE results for time series forecasting of total recruitment posts per day (# per day), and percentage of recruitment posts per day (% per day).

	MASE		RMSE	
	# per day	% per day	# per day	% per day
Naive Model	1.08	1.65	5.23	0.10
Baseline ARIMA	1.01	11.22	5.29	0.72
OLS	5.41	9.72	150.40	6.50
PCR	0.86	0.99	4.44	0.06
Topic.ts	0.82	1.03	4.34	0.08

4.3 Results and Discussion

We evaluated the VE recruitment forecasting methods using the two measures of forecasting accuracy: root-mean-square error (RMSE) and mean absolute scaled error (MASE), depicted in Equations 10 and 11:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(F_t - A_t)} \tag{10}$$

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|F_t - A_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|A_i - A_{i-1}|}\right) \tag{11}$$

Where A_t and F_t are the actual and forecasted values at time t, and the denominator of Equation 11 is the average absolute forecast error of the naive model. We used MASE because it can compare forecasting methods while not returning extremely high or even infinite error rates when forecasts are near zero [33]. RMSE is also helpful for understanding forecasting accuracy because the error rate is in the same units as the response.

Table 8 shows a comparison of five VE recruitment forecasting methods on our response variables. These results were obtained by applying the RMSE and MASE metrics to the set of forecasts predicted at each iteration of the moving horizon. Results show that topic model features in PCR and Topic.ts improve forecast accuracy over the naive model and ARIMA benchmarks. Examining RMSE shows that models of both total recruitment posts per day and percentage of recruitment per day improve by about 20% with the use of latent

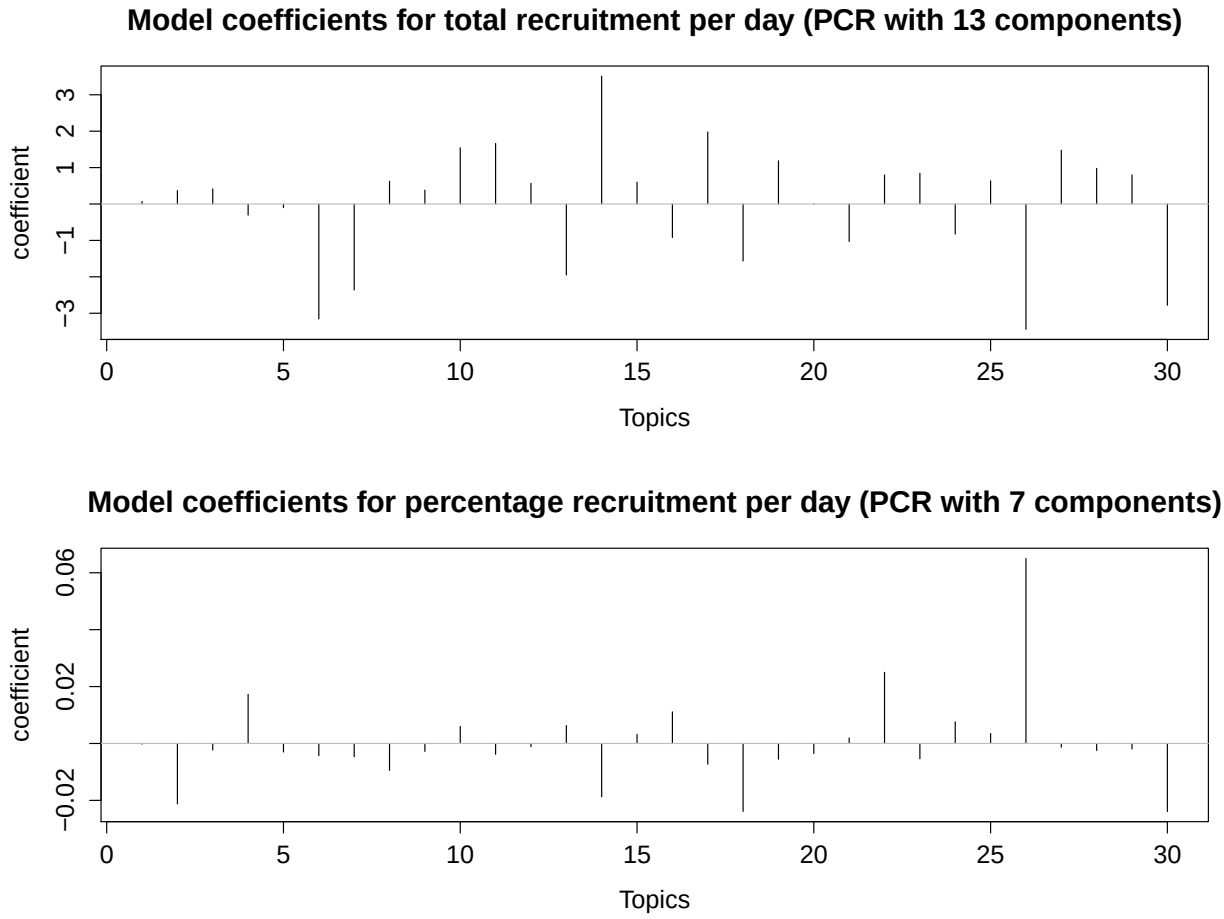


Figure 4: Comparison of topic feature coefficients for the PCR models. Feature weights of the PCR models forecasting (top) total recruitment posts per day, and (bottom) percentage of recruitment posts per day.

predictor variables. The full OLS models are shown here as a comparison with PCR and demonstrate that multicollinearity and a lack of feature selection can dramatically increase prediction error.

To provide some understanding of how the best-performing regression models are being trained to predict the amount of VE recruitment, Figure 4 compares the PCR model coefficients of the latent topic features. The models in Figure 4 represent the last moving horizon iteration which uses the largest set of training data. The two response variables produce distinct models with the number of components being an obvious difference; however, the coefficients show that topics 14, 26, and 30 are clearly important in both instances. We used

Table 9: The most common terms contributing to the highest-weighted topics in the representative PCR models.

Topic 6	Topic 7	Topic 14	Topic 18	Topic 26	Topic 30
said	said	kill	kill	allah	said
allah	allah	attack	allah	will	kill
attack	islam	islam	said	one	attack
kill	will	said	attack	muslim	police
quot(ed,ing,...)	god	soldier	correspond	people	milit(ant,ary,...)
brother	govern	taliban	arm(y,ed,...)	jihad	taliban
mujahideen	muslim	offic(ial,er,...)	will	brother	forc(e,ed,...)

the latent variables produced by the topic model in Section 4.2.1 as features in most of our forecasting methods. Since topics represent the major subjects appearing in a corpus, Table 9 provides some understanding of the most important topics for the dataset and therefore this forecasting task. From the terms in the table, it is clear that violence is a major theme being modeled in the corpus. The words "attack" and "kill" appear in all but two of the topics in Table 9 and "jihad" is common to Topic 26, making Topic 7 the only feature whose most common terms do not imply violence. The terms also highlight the importance of religion in the corpus, with Topic 30 being the only feature in Table 9 without a reference to god or Islam. Perhaps surprisingly, the thirty topics do not seem to be particularly indicative of recruitment. This may be due to LDA's unsupervised learning algorithm and the scarcity of VE recruitment posts, which comprise less than 10% of the Ansar1 corpus. Regardless of how they work, performance metrics show that latent topic features improve VE recruitment forecast accuracy over the benchmarks.

5 Conclusions and Future Work

This work was motivated by the explosive and continuing increase in online activities of violent extremist organizations along with the lack of automated tools to identify and predict such activity. Our research built upon recent data collection and analysis efforts to develop supervised learning and natural language processing methods that automatically identify

and forecast cyber-recruitment by violent extremists. The results presented in this thesis support the conclusion that automatic VE recruitment detection and forecasting are feasible goals. As the first reported results on these tasks, our classifiers and time series models serve as initial performance benchmarks against which future VE recruitment models can be compared.

In the future, our VE recruitment detection and forecasting 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 model 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 [17]. Latent Dirichlet allocation (LDA) 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 [6]. Although LDA topics were used in the time series forecasting models, such latent variable modeling techniques could also serve as feature selection and replacement methods for future VE recruitment classifiers. Additionally, our forecasting methods might be improved by generating further latent topic predictors and employing more sophisticated regression and feature selection methods like gradient boosting or random forests. Future forecasting methods should also consider more advanced natural language processing techniques like supervised latent Dirichlet allocation (sLDA), which has benefits over using unsupervised LDA to generate a topic model followed by separate regression analysis [5].

By testing the effectiveness of our methods in a proxy for real-world settings, we demon-

strated that such automated classification and forecasting tools would 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 [61]. 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 [61]. An automated VE recruitment forecasting system could be used to predict future staffing requirements for analysts tasked with manually reviewing the leads and posts identified by a VE recruitment classification tool.

More generally, our automated classification and time series 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) additional time series methods for analyzing the relationships between recruitment and membership along a timeline.

In light of the still unfolding news regarding the NSA's Boundless Informant and PRISM programs [26], 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-violent 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 component 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 and forecasting 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 the risk of misuse.

Bibliography

[1] Artificial Intelligence Laboratory, University Of Arizona, *Dark Web Forum Portal: Ansar AlJihad Network English website*.

[2] A. Ashari, I. Paryudi, and A. M. Tjoa, *Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool*, International Journal of Advanced Computer Science and Applications (IJACSA), 4 (2013), pp. 33–39.

[3] A. Basu, *Social network analysis of terrorist organizations in India*, in North American Association for Computational Social and Organizational Science (NAACSOS) Conference, 2005, pp. 26–28.

[4] E. Bauer and R. Kohavi, *An empirical comparison of voting classification algorithms: Bagging, boosting, and variants*, Machine learning, 36 (1999), pp. 105–139.

[5] D. M. Blei and J. D. McAuliffe, *Supervised topic models.*, in NIPS, vol. 7, 2007, pp. 121–128.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, The Journal of Machine Learning Research, 3 (2003), pp. 993–1022.

[7] L. Bowman-Grieve, *A psychological perspective on virtual communities supporting terrorist & extremist ideologies as a tool for recruitment*, Security Informatics, 2 (2013), pp. 1–5.

[8] K. M. Carley, *Destabilization of covert networks*, Computational & Mathematical Organization Theory, 12 (2006), pp. 51–66.

[9] W. Cavnar, *Using an n-gram-based document representation with a vector processing retrieval model*, NIST Special Publication, (1995), pp. 269–277.

[10] M. Chau and 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.

[11] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann, *Uncovering the dark web: A case study of jihad on the web*, Journal of the American Society for Information Science and Technology, 59 (2008), pp. 1347–1359.

[12] H. Chen, E. Reid, J. Sinai, A. Silke, and B. Ganor, eds., *Terrorism informatics: knowledge management and data mining for homeland security*, Springer, New York, 2008.

[13] H. Chen, S. Thoms, and 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), 2008, pp. 98–103.

[14] Z. Chen, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, *Identifying intention posts in discussion forums*, in Proceedings of NAACL-HLT, 2013, pp. 1041–1050.

[15] J. Cohen, *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, 20 (1960), pp. 37–46.

[16] M. Conway, *Terrorism and the internet: New media–new threat?*, Parliamentary Affairs, 59 (2006), pp. 283–298.

[17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, 41 (1990), pp. 391–407.

[18] J. Diesner and K. M. 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, Pittsburgh, 2004.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, Inc., New York, 2nd ed., 2001.

[20] I. Feinerer and K. Hornik, *tm: Text Mining Package*, R Foundation for Statistical Computing, 2014. R package version 0.5-10.

[21] A. P. Fenech, *Tukey's method of multiple comparison in the randomized blocks model*, Journal of the American Statistical Association, 74 (1979), pp. pp. 881–884.

[22] J. L. Fleiss, B. Levin, and M. C. Paik, *The measurement of interrater agreement*, Statistical methods for rates and proportions, 2 (1981), pp. 212–236.

[23] J. Friedman, T. Hastie, and R. Tibshirani, *Additive logistic regression: a statistical view of boosting*, The annals of statistics, 28 (2000), pp. 337–407.

[24] T. Fu, A. Abbasi, and H. Chen, *A focused crawler for dark web forums*, Journal of the American Society for Information Science and Technology, 61 (2010), pp. 1213–1231.

[25] Google Inc., *Google Translate*.

[26] G. Greenwald, *NSA collecting phone records of millions of Verizon customers daily*, June 2013.

[27] B. Grün and K. Hornik, *topicmodels: An R package for fitting topic models*, Journal of Statistical Software, 40 (2011), pp. 1–30.

[28] F. Gutiérrez, *Recruitment in a civil war: A preliminary discussion of the colombian case*, in Santa Fe Institute, Mimeo, 2004.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York, second ed., 2009.

[30] T. Helleputte, *LiblineaR: Linear Predictive Models Based On The Liblinear C/C++ Library*, 2013. R package version 1.80-7.

[31] M. Humphreys and J. M. Weinstein, *Who fights? the determinants of participation in civil war*, American Journal of Political Science, 52 (2008), pp. 436–455.

[32] R. J. Hyndman, G. Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, Y. Khan, C. Bergmeir, and E. Wang, *forecast: Forecasting functions for time series and linear models*, 2014. R package version 5.3.

[33] R. J. Hyndman and A. B. Koehler, *Another look at measures of forecast accuracy*, International journal of forecasting, 22 (2006), pp. 679–688.

[34] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*, Springer, Berlin, 1998.

[35] T. P. Jurka, L. Collingwood, A. E. Boydstun, E. Grossman, and W. van Atteveldt, *RTextTools: Automatic Text Classification via Supervised Learning*, 2014. R package version 1.4.2.

[36] E. F. Kohlmann, *Al-Qaida's MySpace: Terrorist recruitment on the internet*, CTC Sentinel, 1 (2008), pp. 8–9.

[37] J. R. Landis and G. G. Koch, *The measurement of observer agreement for categorical data*, biometrics, 33 (1977), pp. 159–174.

[38] M. I. Lichbach, *The Rebel's Dilemma*, University of Michigan Press, Ann Arbor, 1998.

[39] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms*, Machine learning, 40 (2000), pp. 203–228.

[40] C.-J. Lin, R. C. Weng, and S. S. Keerthi, *Trust region newton method for logistic regression*, The Journal of Machine Learning Research, 9 (2008), pp. 627–650.

[41] S. Mandal and E.-P. Lim, *Second life: Limits of creativity or cyber threat?*, in IEEE Conference on Technologies for Homeland Security, 2008.

[42] R. W. McGehee, *Deadly Deceits: My 25 Years in the CIA*, Sheridan Square Publications, Inc., New York, 1983.

[43] G. S. McNeal, *Cyber embargo: Countering the internet jihad*, Case Western Reserve University Journal of International Law, 39 (2008), pp. 789–826.

[44] B.-H. Mevik, R. Wehrens, and K. H. Liland, *pls: Partial Least Squares and Principal Component regression*, 2013. R package version 2.4-3.

[45] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-2.

[46] S. O'Rourke, *Virtual radicalisation: Challenges for police*, (2007).

[47] L. A. Overbey, G. McKoy, J. Gordon, and S. McKitrick, *Automated sensing and social network analysis in virtual worlds*, in Intelligence and Security Informatics (ISI), 2010, pp. 179–184.

[48] L. A. Overbey, G. McKoy, J. Gordon, S. McKitrick, M. H. Jr., L. Buhler, L. Casassa, and S. Yaryan, *Virtual DNA: Investigating cyber-behaviors in virtual worlds*, Technical Report 33-09 E, Space and Naval Warfare System Center Atlantic, Charleston, SC, 2009.

[49] K. Peters and P. Richards, *'Why we fight': Voices of youth combatants in Sierra Leone*, Africa, (1998), pp. 183–210.

[50] R. D. Petersen, *Resistance and rebellion: lessons from Eastern Europe*, Cambridge University Press, New York, 2001.

[51] S. Popkin, *The rational peasant*, Theory and society, 9 (1980), pp. 411–471.

[52] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[53] B. Ripley, *tree: Classification and regression trees*, 2014. R package version 1.0-35.

[54] M. Rogers, *The psychology of cyber-terrorism*, Terrorists, Victims and Society: Psychological Perspectives on Terrorism and its Consequences, (2003), pp. 77–92.

[55] J. C. Scott, *The moral economy of the peasant: Rebellion and subsistence in Southeast Asia*, Yale University Press, New Haven & London, 1976.

[56] R. R. Tomes, *Waging war on terror relearning counterinsurgency warfare*, Parameters, 34 (2004), pp. 16–28.

[57] 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, November 2010, pp. 54–61.

[58] R. Torok, *Developing an explanatory model for the process of online radicalisation and terrorism*, Security Informatics, 2 (2013), pp. 1–10.

[59] J. Tuszynski, *caTools: ROC AUC Tools, moving window statistics*, 2013. R package version 1.16.

[60] C. J. Van Rijsbergen, S. E. Robertson, and M. F. Porter, *New models in probabilistic information retrieval*, British Library Research and Development Dept, London, 1980.

[61] W. H. Webster, D. E. Winter, J. Adrian L. Steel, W. M. Baker, R. J. Bruemmer, and K. L. 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*, tech. report, Federal Bureau of Investigation, 2012.

[62] J. M. Weinstein, *Inside rebellion: The politics of insurgent violence*, Cambridge University Press, New York, 2007.

[63] P. Whitle, *Hypothesis testing in time series analysis*, vol. 4, Almqvist & Wiksells, 1951.

[64] E. J. Wood, *Insurgent collective action and civil war in El Salvador*, Cambridge University Press, New York, 2003.

[65] M. Yang, M. Kiang, H. Chen, and Y. Li, *Artificial immune system for illicit content identification in social media*, Journal of the American Society for Information Science and Technology, 63 (2012), pp. 256–269.