School of Information Studies - Faculty
Scholarship

School of Information Studies (iSchool)

2005

Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content

Ozgur Yilmazel
Syracuse University

Svetlana Symonenko
Syracuse University

Niranjan Balasubramanian
Syracuse University

Elizabeth D. Liddy
Syracuse University

Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content

Ozgur Yilmazel, Svetlana Symonenko, Niranjan Balasubramanian,
Elizabeth D. Liddy

Center for Natural Language Processing
School of Information Studies – Syracuse University
Syracuse, NY 13244
{oyilmaz, ssymonen, nbalasub, liddy}@syr.edu

Abstract.

Experiments were conducted to test several hypotheses on methods for improving document classification for the malicious insider threat problem within the Intelligence Community. Bag-of-words (BOW) representations of documents were compared to Natural Language Processing (NLP) based representations in both the typical and one-class classification problems using the Support Vector Machine algorithm. Results show that the NLP features significantly improved classifier performance over the BOW approach both in terms of precision and recall, while using many fewer features. The one-class algorithm using NLP features demonstrated robustness when tested on new domains.

1 Introduction

This paper reports on further developments in the research [1, 2] that leverages Natural Language Processing (NLP) and Machine Learning (ML) technologies to improve one aspect of security within the Intelligence Community (IC). This would be done by monitoring insiders' workflow documents and alerting the system assurance administrator if the content of the documents shifts away from what is expected, given the insiders' assignments. This capability is being implemented as one piece of a tripartite system prototype within the context of the ARDA-funded project, *A Context, Role and Semantic-based Approach for Countering Malicious Insider* [3]. In particular, we evaluate the applicability of a one-class categorization algorithm - Support Vector Machines (SVM) - which, unlike a regular classifier, is trained on 'typical' examples only and then used to detect both 'typical' and 'atypical' data. This is warranted by the context of the problem where the potential subject domain of interest to the malicious insider is unknown in advance and, therefore, it is not feasible to provide 'off-topic' examples to train a classifier.

2 Problem Background

It is known from Subject Matter Experts (SMEs) from the IC that analysts operate within a mission-based context, focused mainly on specific topics of interest (TOIs)

and geo-political areas of interest (AOIs). The information accessed by analysts ranges from news articles to analyst reports, official documents, emails, queries, and the role and the task assigned to the analyst dictates the scope of their TOI/AOI. Within this mission-focused context, our hypothesis is that the ML-based text categorization of documents produced by the NLP-based semantic analysis of texts will enable a system to measure the extent to which an insider's document workflow is within the scope of the assigned task, in terms of TOI and AOI.

To illustrate the problem, consider the following "Threat Scenario", which is one of the six developed by the project team, based on a review of known malicious insider cases and consultations with the IC. An analyst with appropriate security clearance works on problems dealing with the Biological Weapons Program (TOI) in Iraq (AOI). For some reason, the analyst begins collecting information on ballistic missiles in North Korea. Since the topic is beyond his assigned task, these actions are covert, interspersed with his 'normal', 'on-topic' communications. Now and then he would query a database and retrieve documents on North Korea's missiles; occasionally, he would send a question to another analyst from the North Korea shop and receive documents via email; to pass the information to his external partners, he would copy data to a CD or print documents out. As these actions involve such textual artifacts as documents, database queries, and emails, analysis of their semantic content should be indicative of which topics are of interest to the analyst. Further comparison of these topics to what is *expected*, given the analyst's task, would reveal whether they are beyond the expected scope.

In addition to monitoring insider's communications, semantic analysis can be run *ex-post-facto*, if an information assurance engineer grew suspicious of an individual. Alternatively, it can help quickly characterize large collections of documents by separating them into semantic-driven categories for a wide range of applications.

It is important to note that the system will not replace human supervisors, but assist them by reducing the data to analyze to just the detected 'anomalies'.

3 Related Work

Until recently, the problem of detecting malicious insider activity was mainly approached from the *cyber security* standpoint, with systems as the main object of potential attack [4, 5]. The 2003 and 2004 Symposia on Intelligence and Security Informatics (ISI) demonstrated an increased appreciation of information as an important factor of national security. As information is often represented through textual artifacts, linguistic analysis has been applied to the problems of cyber security. Sreenath [6] showed how reconstruction of users' queries from their online logs with latent semantic analysis can be applied to detect malicious intent. Studies also looked at linguistic indicators of deception in interview transcripts [7], email messages [8], and online chat [9]. Bengel [10] applied classification algorithms to the task of chat topic detection.

Another line of text classification research addresses situations when providing 'negative' examples for training is not feasible, for example, in intrusion detection [11], adaptive information filtering [12, 13], and spam filtering [14]. Recently, research effort has focused on application of a one-class categorization algorithm,

which is trained on positive examples only and then tested on the data that contain both positive and negative examples. Conceptually, the task is to acquire all possible knowledge about one class and then apply it to identify examples that do *not* belong to this class. As the one-class Support Vector Machines (SVM) [15] was shown to outperform other algorithms [12, 13, 16], it was chosen for our experiments. The novelty of our approach is in evaluating its effectiveness on various sets of features selected to represent documents. In particular, we compared the BOW representation with different combinations of linguistic features generated using NLP techniques.

4 Proposed Solution

The task of identifying 'off–topic' documents is modeled as a text categorization problem. Categorization models of expected topics are first built from the semantic content of a given set of documents, representing the analyst's 'normal' workflow. New documents are then categorized as on- or off-topic based on their semantic similarity to this Expected Model. The effectiveness of the solution is dependent on how well we can model expected communications, as well as on the accuracy of the categorization model and its generalizability to new documents. The most commonly used document representation has been the BOW [17, 18]. It has been shown that the knowledge of statistical distribution of terms in texts is sufficient to achieve high classification performance. However, in situations where the available training data is limited (as is frequently true in real-life applications), classification performance on BOW suffers. Our hypothesis is that the use of fewer, more discriminative linguistic features can outperform the traditional bag-of-words representation, particularly in the case of limited training data.

The novelty of the proposed approach is in using linguistic features either extracted or assigned by our NLP-based system [19]. Such features include entities (nouns and noun phrases), named entities (proper names), and their semantic categories (i.e. PERSON, ORGANIZATION). Furthermore, the system can map these features into higher-level concepts from external knowledge sources, particularly, those indicative of TOI and AOI. By utilizing these more abstract features, the system can produce document vectors that are well separated in the feature space.

The NLP analysis is performed by TextTagger™, a text processing system built at the Center for Natural Language Processing (CNLP)[20]. The system employs a part-of-speech tagger and a sequence of rule-based shallow parsing phases that use lexico-semantic and syntactic clues to identify and categorize entities, named entities, events, as well as relations among them. Next, individual topics and locations are mapped to appropriate categories from knowledge bases. The choice of knowledge bases was driven by the project context. Concept inference for TOI is supported by an ontology developed for the Center for Nonproliferation Studies' (CNS)[21] collection of documents from the weapons of mass destruction (WMD) domain. For the conceptual organization of AOI, we utilize the SPAWAR Gazetteer [22]. Given that analysts usually operate on the country-level of AOI, the inference for geographical concepts is set to the 'Country' level, but other levels of granularity are possible. The entity and event extractions are output as frames, with relation extractions as frame slots.

Figure 1 shows sample extractions for the named entity 'Bavarian Liberation Army' with inferred AOI ('Country') and TOI ('CNS_Superclasses').

Authorities suspect the Bavarian Liberation Army, an extreme right-wing organization, may be responsible.

Bavarian Liberation Army
    Country=Austria
    CNS_Superclasses=Terrorist-Group

Fig. 1. A sample extraction and concept inference.

The NLP-extracted features are then used to generate document vectors for machine learning algorithms.

5 Experimentation

5.1 Experimentation Dataset

Experiments were run on a subset of the larger Insider Threat collection created for the project. Its core comes from the CNS collection and covers such topics as WMD and Terrorism, and such genres as newswires, articles, analytic reports, international treaties, emails, and so on. Training and Testing document sets were drawn from the collection based on the project scenarios. These scenarios are synthetic datasets that represent the insiders' workflow through atomic actions (e.g. 'search database', 'open document'), some of which are associated with documents. The scenarios span a period of six months each and include a baseline case (with no malicious activity) and six threat cases. The scenarios cover the workflow of hundreds of insiders with different work roles and tasks; for our experiments, we focused on one analyst from the Iraq/Biological Weapon shop. The above described Threat Scenario set the base for the Training and Testing datasets.

The documents were retrieved in a manner simulating the analysts' work: manually constructed task-specific queries (Figure 2) were run against the Insider Threat collection. Sets of such queries were also included in the Training and Testing datasets.

(a)   *+UNMOVIC +inspect* +biolog* +Iraq**

(b)   *+missile +test* North +Korea*

Fig. 2. Sample queries on topics of 'Biological weapons program in Iraq' (a) and 'Missile program in North Korea' (b).

Both sets included 'noise' (webpages on topics of general interest) as it is realistic to assume that, in the course of their workday, analysts may use the Web for personal reasons as well.

Documents retrieved by the 'North Korea' queries were labeled as OFF-topic. All other documents were labeled as ON-topic, since, for the purposes of the project, it will suffice if the classifier distinguishes the 'off-topic' documents from the rest. The Training set contained only ON-topic documents, whereas the Testing set also included OFF-topic documents. Table 1 shows the content and the volume of the resulting Training and Testing datasets. The relatively small share of OFF-topic documents in the Testing set (only 8.4%), though realistic given the context of the project, represented yet another challenge, as classification algorithms tend to favor more populated classes.

Table 1. Training and Testing datasets.

	Training	Testing	
	ON (Iraq/Bio)	ON (Iraq/Bio)	OFF (NK/Missile)
Documents	6382	3194	183
Queries	461	222	135
Total Class	*6843*	*3416*	*318*
Total Set	**6843**		**3734**

5.2 Classification experiments

For classification experiments, we used an SVM classifier not only because it has been shown to outperform kNN, Naïve Bayes, and other classifiers on the Reuters Collection [23, 24], but also because it can handle one-class categorization problems as well. Experiments were run in LibSVM [25], modified to handle file names in the feature vectors, and to compute a confusion matrix for evaluation.

We experimented with the following feature sets:

1. Bag-of-words representation (BOW): each unique word in the document is used as a feature in the document vector.
2. Categorized entities (CAT): only words identified as entities or named entities constitute features in the document vector.
3. TOI/AOI extractions (TOI/AOI): document vector includes only terms assigned as TOI/AOI indicators
4. TOI/AOI extractions + important categories (TOI/AOI_cat): document vector uses TOI/AOI features (as in 3) plus all entities and named entities categorized as geographical or domain-relevant concepts (e.g. 'WMD', 'missile', 'terrorism')

We applied stemming, a stop-word filter, and lower case conversion to all of the representations. The associated value for each term in the document representation is the frequency of the term in that document. The experiments reported herein were run with the linear kernel SVM, all parameters set to default.

The results of the experiments can be represented in a confusion matrix (Table 2), where TrueON are documents correctly classified as ON-topic; FalseON are OFF-topic documents assigned to the ON-topic class; TrueOFF are correctly detected OFF-topic documents, and FalseOFF are ON-topic documents misclassified into the OFF-topic class.

Table 2. Confusion matrix.

	True ON	False ON	True OFF	False OFF
BOW	2436	67	251	1070
CAT	1954	80	238	1462
AOI/TOI	1819	51	267	1597
AOI/TOI_cat	2412	16	302	1004

Classifier performance was assessed using standard metrics of precision and recall [26] and a weighted F-score, calculated for each class. Figure 3 shows sample formulas for precision on the ON-topic (1) and the recall of the OFF-topic (2) classes.

$$\text{Precision (ON)} = \frac{\text{TrueON}}{\text{TrueON} + \text{FalseON}} \tag{1}$$

$$\text{Recall (OFF)} = \frac{\text{TrueOFF}}{\text{TrueOFF} + \text{FalseON}} \tag{2}$$

Fig. 3. Sample formulas for Precision and Recall.

In mainstream text categorization research, the performance focus is usually on the 'positive' class, so the scores (precision, recall, F-measure) are often reported for this class only. The context of our project, however, gives much greater importance to detecting the 'negative' (i.e. potentially malicious) cases, while keeping the rate of 'false alarms' (FalseOFF) down. This provided a rather uncommon task for training the classifier: to aim not only for higher precision on ON-topic, but also for greater recall of OFF-topic. Therefore, in evaluating the classifier, we focused on the scores for the OFF-topic class, therefore, for the OFF-topic class, the F-measure was calculated with the weight $\beta=10$ (i.e. the Recall was weighted 10 times as important as Precision). The F-score for the ON-topic class was calculated using the standard weight $\beta=1$. Figure 4 shows the F-measure formula used. The actual value of β is not significant as long as it is greater than zero, since it places a higher emphasis on the precision than recall and F-score is not used to tune parameters of the learning algorithm.

$$\text{F-score} = \frac{(\beta+1) * \text{Precision} * \text{Recall}}{\beta * \text{Precision} + \text{Recall}} \tag{3}$$

Fig. 4. Weighted F-score.

The results (Table 3) demonstrate that, similarly to what was observed in experiments with the regular SVM classifier [2], document representations using TOI/AOI features only (TOI/AOI) or in combination with domain-important categories

(AOI/TOI_cat) improve the classifier performance over the baseline (BOW), while using many fewer features. In particular, AOI/TOI shows over 5% improvement in Recall (OFF) while using forty-nine times fewer features. Using a combination of AOI/TOI and category information (AOI/TOI_cat) achieves 16% improvement on Recall (OFF) and over 12% improvement on the weighted F-OFF over the baseline with nine times fewer features than BOW.

Table 3. Experimental results.

	Features	Prec ON	Rec ON	F ON, $\beta=1$	Prec OFF	Rec OFF	F OFF, $\beta=10$
BOW	19774	97.22	68.68	80.50	19.0	78.93	61.34
CAT	10682	96.07	57.20	71.71	14.0	74.84	53.65
AOI/TOI	403	97.27	53.25	68.82	14.32	83.96	58.22
AOI/TOI_cat	2149	99.34	70.61	82.55	23.12	94.97	74.05

Although the decision to switch from the regular to the one-class SVM was guided by the context of our project, it was supported by the significantly higher performance of the one-class SVM on the OFF-topic class (Table 4). Regular SVM suffered from training on a weakly representative set for the OFF-topic class. Considering that the one-class SVM was able to achieve up to 94% of recall of 'off-topic' examples with no prior knowledge of what constitutes 'off-topic', the improvement is impressive. The downside of such a high recall of the OFF-topic, however, was the deteriorated recall of the ON-topic. In other words, the one-class SVM errs in favor of the previously unknown 'negative' class, thus, causing 'false alarms'.

Table 4. Recall of the OFF-topic class: Regular vs. One-Class SVM.

	Regular SVM		One-Class SVM	
	Recall OFF	F OFF, $\beta=10$	Recall OFF	F OFF, $\beta=10$
BOW	48.11	50.49	78.93	61.34
CAT	27.0	28.92	74.84	53.65
AOI/TOI	38.99	41.28	83.96	58.22
AOI/TOI_cat	38.68	40.96	94.97	74.05

Next, as in our experiments with the regular SVM [2], we wanted to assess how the one-class SVM will perform on a different 'off-topic' domain. We used the same Training set, and the ON-topic part of the Testing set. For the OFF- part of the Testing set, the documents were retrieved from the Insider Threat dataset with queries on the topic of 'China/Nuclear weapons' (Table 5):

Table 5. Testing dataset with OFF-topic documents drawn from the 'China/Nuclear' domain.

Testing China/Nuclear		
	ON-topic (Iraq/Bio)	OFF-topic (China/Nuclear)
Documents	3194	181
Queries	222	129
Total Class	*3416*	*310*

Total Set	3726	

Experimental results (Tables 6 and 7) support the trend observed in the prior experiments. One-class categorization on the NLP-enhanced document representations achieves superior performance, particularly on the 'off-topic' class, compared to the baseline (BOW). Besides, the domain change for the 'off-topic' documents does not seem to impact the classifier performance to a significant extent, which was the case with the regular SVM. Such robustness is quite reasonable, since the one-class SVM is not biased (via training) towards a particular kind of 'negative' data.

Table 6. Confusion Matrix (OFF-topic documents drawn from the 'China/Nuclear' domain).

	True ON	False ON	True OFF	False OFF
BOW	2346	134	176	1070
CAT	1954	83	227	1462
AOI/TOI	1819	96	214	1597
AOI/TOI_cat	2412	79	231	1004

Table 7. Experimental results (OFF-topic documents drawn from the 'China/Nuclear' domain).

	Features	Prec ON	Rec ON	F ON, β=1	Prec OFF	Rec OFF	F OFF, β=10
BOW	19774	94.60	68.68	79.58	14.13	56.77	44.55
CAT	10682	95.93	57.20	71.67	13.44	73.23	52.14
AOI/TOI	403	94.99	53.25	68.24	11.82	69.03	47.94
AOI/TOI_cat	2149	96.83	70.61	81.67	18.70	74.52	58.61

Overall, the results show that the one-class SVM performs impressively well, especially, on recall of the OFF-topic class. Another important point is that the algorithm appears to be robust to handle different subject domains of 'negative' examples. We believe, therefore, that it can be effectively applied to categorization problems where only 'positive' examples are available. The results also demonstrate that the use of NLP-based features achieves better performance in categorization while using many fewer features than the commonly used bag-of-words representation.

6 Conclusion and directions for future research

The experiments described herein show that leveraging one-class SVM with the NLP-extracted features for document representation improves classification effectiveness and efficiency. In future research we will seek to evaluate the impact of different combinations of linguistic features, extractions from text, and concepts inferred from external knowledge bases on categorization accuracy. In addition, to further explore the robustness of the one-class classifier, we plan to test it on a combination of different subject domains for the 'off-topic' class.

The one-class approach fits particularly well the situations where it is not feasible to provide 'atypical' examples. Overall, the research reported herein holds potential for providing the IC with the analytic tools to recognize anomalous insider activity; as

well as to build content profiles of vast document collections when applied in a broader context.

7 Acknowledgements

References

[1]   S. Symonenko, E. D. Liddy, O. Yilmazel, R. DelZoppo, E. Brown, and M. Downey, "Semantic Analysis for Monitoring Insider Threats," presented at The Second NSF/NIJ Symposium on Intelligence and Security Informatics (ISI2004). Tucson, AZ, 2004.

[2]   O. Yilmazel, S. Symonenko, E. D. Liddy, and N. Balasubramanian, "Improved Document Representation for Classification Tasks For The Intelligence Community (Forthcoming)," presented at AAAI, CA, 2005.

[3]   R. DelZoppo, E. Brown, M. Downey, E. D. Liddy, S. Symonenko, J. S. Park, S. M. Ho, M. D'Eredita, and A. Natarajan, "A Multi-Disciplinary Approach for Countering Insider Threats," presented at Workshop on Secure Knowledge Management (SKM), Amherst, NY, 2004.

[4]   J. Anderson, "Computer Security Threat Monitoring and Surveillance," James P. Anderson Co., Fort Washington, PA 15 April 1980 1980.

[5]   R. H. Lawrence and R. K. Bauer, AINT misbehaving: A taxonomy of anti-intrusion techniques, http://www.sans.org/resources/idfaq/aint.php.

[6]   D. V. Sreenath, W. I. Grosky, and F. Fotouhi, "Emergent Semantics from Users' Browsing Paths," presented at First NSF/NIJ Symposium on Intelligence and Security Informatics, Tucson, AZ, USA, 2003.

[7]   J. Burgoon, J. Blair, T. Qin, and J. Nunamaker, Jr., "Detecting Deception Through Linguistic Analysis," presented at First NSF/NIJ Symposium on Intelligence and Security Informatics, Tucson, Arizona, 2003.

[8]   L. Zhou, J. K. Burgoon, and D. P. Twitchell, "A Longitudinal Analysis of Language Behavior of Deception in E-mail," presented at First NSF/NIJ Symposium on Intelligence and Security Informatics., Tucsona, AZ, USA, 2003.

[9]   D. P. Twitchell, J. F. Nunamaker Jr., and J. K. Burgoon, "Using Speech Act Profiling for Deception Detection," presented at Second NSF/NIJ Symposium on Intelligence and Security Informatics (ISI2004), Tucson, AZ, 2004.

[10]   J. Bengel, S. Gauch, E. Mittur, and R. Vijayaraghavan, "ChatTrack: Chat Room Topic Detection Using Classification," presented at Second NSF/NIJ Symposium on Intelligence and Security Informatics (ISI2004), 2004.

[11]   K. A. Heller, K. M. Svore, A. Keromytis, D.,, and S. J. Stolfo, "One Class Support Vector Machines for Detecting Anomalous Windows Registry Accesses," presented at The Third IEEE International Conference on Data Mining, Melbourne, Florida, USA, 2003.

[12]   H. Yu, C. Zhai, and J. Han, "Text classification from positive and unlabeled
         documents," presented at The Twelfth International Conference on Informa-
         tion and Knowledge Management, New Orleans, LA, USA, 2003.
[13]   L. M. Manevitz and M. Yousef, "One-class svms for document classifica-
         tion," *The Journal of Machine Learning Research*, vol. 2, pp. 139-154,
         2002.
[14]   K.-M. Schneider, "Learning to Filter Junk E-Mail from Positive and Unla-
         beled Examples," 2004.
[15]   B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. William-
         son, "Estimating the Support of a High-Dimensional Distribution," Micro-
         soft Research Technical Report 99-87, 1999.
[16]   B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building Text Classifiers
         Using Positive and Unlabeled Examples," presented at The Twelfth Interna-
         tional Conference on Information and Knowledge Management, New Or-
         leans, LA, USA, 2003.
[17]   S. Dumais, P. John, D. Heckerman, and M. Sahami, "Inductive learning al-
         gorithms and representations for text categorization," presented at The Sev-
         enth International Conference on Information and Knowledge Management,
         Bethesda, Maryland, United States, 1998.
[18]   F. Sebastiani, "Machine learning in automated text categorization," *ACM
         Computing Surveys*, vol. 34, pp. 1-47, 2002.
[19]   E. D. Liddy, "Natural Language Processing," in *Encyclopedia of Library
         and Information Science*, 2nd ed. New York: Marcel Decker, Inc., 2003.
[20]   Center for Natural Language Processing (CNLP), www.cnlp.org.
[21]   Center for Nonproliferation Studies (CNS), http://cns.miis.edu/.
[22]   B. Sundheim and R. Irie, "Gazetteer Exploitation for Question Answering:
         Project Summary," 2003.
[23]   Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods,"
         presented at 22nd Annual International SIGIR, Berkley, CA, 1999.
[24]   T. Joachims, *Learning to Classify Text using Support Vector Machines:
         Ph.D. Thesis*: Kluwer Academic Publishers, 2002.
[25]   C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector ma-
         chines," 2001.
[26]   C. J. van Rijsbergen, *Information Retrieval*. London: Butterworth, 1979.