Improved Document Representation for Classification Tasks For The Intelligence Community

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

Research within a larger, multi-faceted risk assessment project for the Intelligence Community (IC) combines Natural Language Processing (NLP) and Machine Learning techniques to detect potentially malicious shifts in the semantic content of information either accessed or produced by insiders within an organization. Our hypothesis is that the use of fewer, more discriminative linguistic features can outperform the traditional bag-of-words (BOW) representation in classification tasks. Experiments using the standard Support Vector Machine algorithm and the LibSVM algorithm compared the BOW representation and two NLP representations. Classification results on NLP-based document representation vectors achieved greater precision and recall using forty-nine times fewer features than the BOW representation. The NLP-based representations improved classification performance by producing a lower dimensional but more linearly separable feature space that modeled the problem domain more accurately. Results demonstrate that document representation using sophisticated NLP-extracted features improved text classification effectiveness and efficiency with the SVM and LibSVM algorithms.

Introduction

This research addresses the question of whether the AI technologies of Natural Language Processing (NLP) and Machine Learning (ML) based text categorization can be applied to the problem of monitoring insider activity with the goal of detecting malicious insiders within the Intelligence Community (IC) (Symonenko et al. 2004). This would be done by monitoring insiders' work flow documents and emitting an alert to the central risk assessor monitored by a system assurance administrator if the documents accessed or produced by an IC analyst are not semantically appropriate to the domain of the analyst's assigned tasks. NLP-driven information extraction and ML-based The capability is being implemented and tested as one piece of a tripartite solution in a system prototype within the context of a larger project being conducted under ARDA's Information Assurance for the Intelligence Community Program. The project, *A Context, Role and Semantic (CRS)-based Approach for Countering Malicious Insider Threats* (further in the text referred to as the Insider Threat project) is focused on advancing the state of the art in Insider Threat countermeasures by developing techniques to model behavior of insiders operating in an IC context and to distinguish between expected and anomalous user behavior. The Semantic Analysis (SA) of content being described here, is coupled in the prototype with Social Network Analysis (SNA) which monitors and detects anomalies in social behavior, and Composite, Role-based Monitoring (CRBM) which analyzes insider activity based on organizational, application, and operating system roles (DelZoppo et al. 2004).

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) that are assigned based on the analysts' expertise and experience. The information that is accessed or produced by analysts ranges from news articles to analyst reports, official documents, email communications, query logs, etc, and the role and the task assigned to the analyst dictates their domain of interest in terms of TOI/AOI, their communication patterns, intelligence products and information systems needed, as well as the intelligence work products created. Within this mission-focused context, our hypothesis is that NLP-based semantic analysis of text, combined with ML-driven text categorization based on features produced by the NLP, will enable a system to measure the extent to which an insider's text-based communications are "off-topic" in terms of their TOI and AOI for the task they have been assigned.

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 (insider), 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 apparently 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 contact a friend, another analyst from the North Korea workshop about the country's missile capabilities and receive documents via email; to pass the information to his external partners, he would copy data to a CD or print a few documents out. As these actions involve textual artifacts, such as documents, database queries, emails (both messages and file attachments), an analysis of their semantic content should be indicative of which topics are currently of interest to the analyst. Further comparison of these topics to what is *expected* from the analyst, given the assigned task, would reveal whether they fall within the scope of the expected TOI/AOI.

Obviously, such actions as making phone calls or sending documents via regular mail are beyond the scope of the semantic analysis. To some extent, they can be detected by other system components: the SNA monitor, for example, will detect communication patterns, which are unusual for the insider in terms of frequency or communication partner. In another plausible scenario, the insider can pose a threat by exfiltrating information to external sources. In this case, the content of the involved documents falls within the expected scope, so no anomaly would be detected by the semantic analysis. However, an unusually high volume of printing, copying or emailing to an external source will be reported by the CRBM and SNA monitors. Such combination of evidence of different types and from different sources ensures the sensitivity and robustness of the insider threat detection system which is currently being developed.

In addition to monitoring insider's communications, semantic analysis can be run *ex-post-facto*, for example, if an information assurance engineer's intuition leads them to suspect an individual. Alternatively, it can be applied to characterize large collections of documents by generating their conceptual models.

It is important to note that the system will not replace human supervisors (e.g. information assurance engineers), but rather assist them by directing their attention to the detected 'anomalies'. Obviously, some of these 'anomalies' will be 'false alarms', but such filtering should still save human time otherwise spent on intuition-guided analysis of much more data. System optimization should then aim for the high recall of truly 'anomalous' indicators, while keeping the rate of 'false alarms' low.

Related Work

To date, the research addressing the problem of detecting malicious insider activity has placed greater emphasis on *cyber security,* with systems and networks as the main object of such attacks. Until recently, semantics has been mainly applied to describe the role-based access policy of an organization (Anderson 1999; Upadhyaya, Chinchani, and Kwiat 2001). The 2003 and 2004 Symposia on Intelligence and Security Informatics (ISI) demonstrated an increased appreciation of information itself as an important factor of national security and a potential object of attack. As information is often represented through textual artifacts of various genres, linguistic analysis has been actively applied to the problems related to cyber security. Stolfo et al. (2003) mined subject lines of email messages for patterns typical for particular user groups. Other studies looked at linguistic indicators of deception in interview transcripts (Burgoon et al. 2003), email messages (Zhou, Burgoon, and Twitchell 2003), and online chat room conversations (Twitchell, Nunamaker Jr., and Burgoon 2004). Bengel et al. (2004) applied classification algorithms to the task of chat room topic detection. Sreenath et al. (2003) employed latent semantic analysis to reconstruct users' original queries from their online browsing paths and applied this technique to detecting malicious (terrorist) trends.

Our work is conceptually related to anomaly detection research (Anderson 1980; Lawrence and Bauer 2000). The novelty of our approach is in its focus on the insider, who, unlike an intruder, may possess required system security clearance. Another important point of distinction is in that the content of information accessed eludes the existing intrusion detection techniques, as little can be inferred from the available resource metadata. To address this need, we have taken a document-driven approach that focuses on the content. This paper reports on further developments in the ongoing research (DelZoppo et al. 2004; Symonenko et al. 2004) that combines NLP and ML techniques to detect potentially threatening shifts in the content of information accessed by an insider.

Proposed Solution

The problem of identifying documents that are off – or on – topic can be modeled as a text categorization problem. Categorization models of expected topics are first built from the semantic content of a given set of documents that reflect the analyst's assignment. New documents are then categorized as on-topic or off-topic based on the similarity of their semantic content to what is expected. When the level of risk based on off-topic documents accessed and/or produced exceeds a pre-defined threshold, a risk indicator

is sent to the central risk assessor, which merges this information with evidence indicators from anomaly detectors of other cyber-observables for review and action by an information assurance engineer.

The effectiveness of such a solution is dependent on how well we can model expected communications and the accuracy of the categorization models in assigning documents that are accessed and produced to the categories of on-topic and off-topic and as well as the generalizability of the model to new documents. The most commonly used document representation has been the simple bag-of-words (BOW) (Dumais et al. 1998; Sebastiani 2002). It has been shown that in many text classification problems, the vocabulary of the categories and its statistical distribution is sufficient to achieve high performance results. However, in situations where the available training data is limited (as is frequently true in real-life applications), classification performance suffers. Our hypothesis is that the use of fewer, more discriminative linguistic features can outperform the traditional bag-of-words representation.

Advantages of using NLP-driven representations are two-fold. First, domain-specific feature selection, utilizing external (domain) knowledge, reduces the feature space to the more content-laden terms. In the IC domain, terms indicative of AOI and TOI are more important for describing the document content than other named entities (for example, the name of the news agency). Therefore, limiting document representations to AOI/TOI features only should still provide for good conceptual separation of documents and, thus, perform as good as, if not better than, the BOW representation. Next, NLP-based concept inference adds dimensions of (dis)similarity. For example, a document reporting on a terrorist bombing in Moscow and a document discussing a terrorist activity in Petersburg will be considered more similar when the broader AOI concept of 'country' ("Russia") is added to both.

In brief, utilizing NLP features to produce a semantic representation of the documents' content makes it possible to utilize both real world knowledge and domain knowledge available in resources such as ontologies to even further improve the representation that provides the basis of the categorization.

The novelty of the proposed approach is in using linguistic features either extracted or assigned by our NLP-based system (Liddy 2003) for document representation. Such features include part-of-speech (POS) tags, entities (nouns and noun phrases), named entities (proper names) and categories of entities and named entities. Furthermore, the system can utilize these document-based NLP features to map into and inference about higher-level concepts in external knowledge sources that are indicative of topics of interest (TOI) and geo-political areas of interest (AOI). 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). The text processing phases, fairly standard for NLP systems, include a probabilistic part-of-speech tagger and a sequence of shallow parsing phases using symbolic rules in a regular expression language. These phases employ 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 in knowledge bases by linguistic rules and an automated querying of these knowledge bases.

The choice of knowledge bases was driven by the context of our project – the IC with its focus on TOI and AOI. Concept inference for TOI is supported by an ontology developed for the Center for Nonproliferation Studies' (CNS) collection of documents from the nonproliferation of weapons of mass destruction (WMD) domain. The process of TOI inference begins when the system recognizes that a term from a document exists in the knowledge base. It then augments the term extraction by all classes it belongs to. We also utilize information about the entity, found in the "gloss"-like ontology attributes, to enhance the term extraction with related terms[1]. For the conceptual organization of AOI, we utilize the SPAWAR Gazetteer, which combines resources of four publicly available gazetteers: NGA (NIMA); USGS; CIA World Factbook; and TIPSTER (Sundheim and Irie 2003). Given that analysts usually operate on the country-level of AOI, the concept inference for geographical terms is set to the 'Country' level, but it allows for different levels of inference granularity. 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': inferred AOI ('Country') and TOI ('CNS_Superclasses'), as well as named entities found in the ontology "glosses" ('CNS_Namedentity').

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

Bavarian Liberation Army
Country=Austria
CNS_Namedentity=Graz
CNS_Superclasses=Terrorist-Group

Figure 1. A sample extraction and concept inference

[1] this simulates analyst's utilizing background knowledge or coming up with useful associations

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

Experimentation

Experimentation Dataset

To assess the effectiveness of using NLP-extracted features vs. bag-of-words document representations, we ran experiments on a subset of the collection created for the purposes of the Insider Threat project. With the known constraints on procuring actual data from IC operational settings, we gathered resources that would best represent the spectrum of textual data accessed during the analyst's work processes. CNS documents constitute the core of the collection, which covers such topics as WMD and Terrorism and is diverse in genre, encompassing, among others, newswires, topic articles, analytic reports, international treaties, and emails.

Training and Testing document sets were drawn from the Insider Threat collection based on the project scenarios. The scenarios are synthetic datasets that represent the insiders' workflow through atomic actions (e.g. 'search database', 'make phone call', 'open document'), some of which are associated with the documents. The scenarios include a baseline case (with no insider threat activity) and six different threat cases. The baseline scenario was used to develop the Training set, and the above described threat scenario set the base for the Testing set. The scenarios describe the workflow of hundreds of insiders with different work roles and tasks; for the purposes of the experiment, we focused on the workflow of one junior analyst from the Iraq/Biological Weapon workshop.

The documents were retrieved in a manner simulating the analysts' work processes: manually constructed topic-specific queries were run against the Insider Threat collection. Figure 2 shows sample queries for the task of collecting information on the topic of 'Biological weapons program in Iraq'.

+Iraq +biologic* weapon* program*
+Iraq +biologic* manufacturing facilit* site* locat**

Figure 2. Sample queries on Iraq/Biological weapons.

For our initial experiments, we assumed that off-topic examples can be supplied. In this particular case, 'North Korea/Missiles' was chosen as malicious topic shift. 'Off-topic' documents were retrieved in a similar fashion by running a set of queries on the topic of 'missile program in North Korea' against the above collection. Sets of such queries were also included in the Training and Testing datasets. Since 'negative' ('off-topic') examples were not part of the Baseline scenario, but the classifier needed to be trained on both 'positive' and 'negative' documents, the latter were added to the Training set. An additional group of documents constituted so-called "white noise": web pages on topics of general interest (news, technology, sports, finance), as it is expected that, in the course of their workday, analysts access the Web for personal reasons as well.

Documents retrieved by '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. For each project scenario, which spans the period of six months, the insider accesses over 6000 documents. Half of the original Testing set was held out for validation experiments, so the final ratio of Training to Testing datasets was 2:1. Table 1 shows the content and volume of the resulting Training and Testing datasets.

Both sets come from the same domains of Iraq/Biological weapons (ON-topic) and North Korea/Missile (OFF-topic). While the project context dictated an overlap between the ON-topic sets for Training and Testing, the OFF-topic sets are disjoint, so the classifier was tested on the unseen OFF-topic documents.

Training

	ON-topic (Iraq/Bio)	OFF-topic (NK/Missile)
Documents	6382	165
Queries	461	
Total Class	*6843*	*165*
Total Set	**7008**	

Testing

	ON-topic (Iraq/Bio)	OFF-topic (NK/Missile)
Documents	3194	183
Queries	222	135
Total Class	*3416*	*318*
Total Set	**3734**	

Table 1. Training and Testing datasets

Classification experiments

We chose to run classification experiments using a Support Vector Machine (SVM) classifier, because it has been empirically shown that SVM outperforms kNN, Naïve Bayes, and some other classifiers on the Reuters Collection (Yang and Liu 1999). Joachims (2002) provides detailed theoretical explanation of why SVMs are appropriate for text categorization.

Categorization experiments were run in LibSVM, an SVM-based machine learning application (Hsu, Chang, and Lin). The basic LibSVM algorithm is a simplification of SMO (Platt 1999) and SVMLight (Joachims 2002) algorithms. We modified LibSVM to handle file names in the feature vectors, and to compute a confusion matrix for evaluation.

When producing a document representation, the goal is to choose the features that allow document vectors belonging to different categories to occupy compact and disjoint regions in the feature space (Jain, Duin, and Mao 1999). We ran experiments using different types of information that we extracted from documents for representation in the SVM classifier:

1. Bag-of-words representation (BOW): each unique word in the training corpus is used as a term in the feature vector.
2. Categorized entities (CAT): only the words that are identified as entities or named entities (proper names) from the training corpus are used for representation.
3. TOI/AOI extractions (TOI/AOI): only the words that are extracted as TOI/AOI indicators are used for representation.
4. TOI/AOI extractions + domain-important categories (TOI/AOI_cat): feature vectors include TOI/AOI indicators plus all entities and named entities categorized as person, geographic or domain-related categories.

We applied stemming, a stop-word filter, and lower case conversion to all of the above representations. The associated value for each term in the document representation is the frequency of the term in that document.

All experiments were run with the RBF kernel SVM, default gamma, with the cost of five applied to negative ('off-topic') examples (i.e. the classifier was penalized for misclassifying 'off-topic' documents).

The classifier performance was assessed using standard metrics of precision, recall, and F-score (van Rijsbergen 1979), calculated for the ON-topic class (Figure 2) where *TrueON* are documents correctly classified as being ON-topic; *FalseON* are OFF-topic documents assigned to the ON-topic class; *TrueOFF* are correctly detected OFF-topic documents, and *FalseOFF* are 'false alarms', or ON-topic documents mistakenly assigned to the OFF-topic class.

Precision (ON) = TrueON / (TrueON + FalseON)

Recall (ON) = TrueON / (TrueON + FalseOFF)

F-score = 2*Precision*Recall/(Precision + Recall)

Figure 2. Formulas for precision, recall, and F-score.

Table 2 shows the results of the experiments. While the performance of the LibSVM classifier is quite impressive on all features sets, using more sophisticated features to represent documents does improve the classifier's precision on the ON-topic class, compared to the bag-of-words representation. Interestingly, using all extracted entities and named entities for representation achieves rather modest results compared to both BOW and AOI/TOI representations. But combining AOI/TOI extractions with entities and named entities representing people, geographic and domain-related categories (such as 'WMD', 'missile', 'terrorism', etc.) shows the best results while still using three times fewer features than the baseline (BOW).

	Features	Prec.	Recall	F-score
BOW	19774	94.36	100	97.10
CAT	10682	93.74	100	96.77
AOI/TOI	403	95.79	100	97.85
AOI/TOI_cat	6635	95.93	100	97.92

Table 2. Experimental results for the ON-topic class.

Since our project set a rather uncommon task for the classifier, namely, to seek greater recall of 'negative' (OFF-topic) examples, it seems worthy to compare the classifier performance on the OFF-topic class as well (Table 3). This task was complicated by the small number of OFF-topic examples (2.3% in the Training set; 8.5% in the Testing set), which, though a realistic share in the context of our project, poses a known problem for generalizability of a classifier as it tends to over fit to 'positive' examples. The results on the OFF-topic class are in accordance with the ones described for the ON-topic: using AOI/TOI features alone or in combination with the domain-important categories boosts recall of OFF-topic documents by over 3-5% while using many fewer features. Furthermore, the use of AOI/TOI features retains the high precision of the OFF-topic class. This observation is in accordance with our hypothesis that the use of NLP based features for document representation improves the classification performance by producing a low dimensional but more linearly separable feature space that models the problem domain more accurately.

	Precision	Recall	F-score
BOW	100	35.85	52.78
CAT	100	28.30	44.12
AOI/TOI	100	52.83	69.14
AOI/TOI_cat	100	54.40	70.47

Table 3. Experimental results for the OFF-topic class.

Overall, the results show that the use of NLP-extracted features and NLP-based inferencing helps to improve performance in categorization.

Classification experiments: Domain change

Since it is hard to predict what kind of malicious activity an analyst will engage in, we would like our system to be robust enough to identify 'off-topic' documents from different subject domains without any pre-knowledge of what these are. To explore how robust the trained classifier is to a change in the subject domain, we ran the experiments with the same Training set, but modified the Testing set by drawing the OFF-topic documents from the domain of 'China/Nuclear weapons' (Table 4):

	Testing	
	ON-topic (Iraq/Bio)	OFF-topic (China/Nuclear)
Documents	13194	181
Queries	222	129
Total Class	*3416*	*310*
Total Set	**3726**	

Table 4. Testing dataset with the OFF-topic documents drawn from the 'China/Nuclear Weapons' domain

The results reported below (Table 5) were produced with the same LibSVM parameters as were used in the prior experiments: RBF kernel, default gamma, and cost of five for misclassifying 'off-topic' examples. While the performance of the classifier on the 'ON-topic' class degrades somewhat, it is still satisfactory. The results support the trend observed in the prior experiments: using linguistic features helps improve the classifier's performance. In particular, combining AOI/TOI extractions with important categories gave the best classifier's precision and the F-score on this dataset. Although such an improvement may look slight, one should bear in mind that it was achieved with almost three times fewer features than the BOW representation.

	Features	Prec.	Recall	F-score
BOW	19774	91.80	100	95.72
CAT	10682	91.98	100	95.82
AOI/TOI	403	91.73	100	95.69
AOI/TOI_cat	6635	92.0	100	95.83

Table 5. Experimental results for the ON-topic class: OFF-examples for the Testing set come from the 'China/Nuclear Weapons' domain

The recall of the 'off-topic' documents degraded for the China/Nuclear Weapons experiment. This was an expected consequence of the domain change for the OFF-topic documents used for training and testing. Still, running the classifier on the NLP-enhanced document representations tends to outperform the baseline (BOW). It seems appropriate to note that, since the size of the OFF-topic class is quite small (318 documents), even the slight change in the absolute numbers produces a noticeable effect on the percentage results.

Summing up the results of the experiments, we conclude that the use of NLP-enhanced features for categorization provides the following important advantages.

1. Improvements in effectiveness – categorization using NLP-enhanced features outperforms "bag-of-words" categorization and produces a feature space where documents are more linearly separable.
2. Improvements in efficiency – use of NLP-extracted features helps reduce the feature space by retaining features that are more discriminating in the problem domain.
3. Enables incorporation of external knowledge (available in such resources as ontologies, gazetteers) for generation of categorization models.

Conclusion and directions for future research

Our experimental results demonstrated that document representation using sophisticated NLP-extracted features improved text categorization effectiveness and efficiency with the SVM classifier. While the number of available training documents is limited in homeland security and intelligence environments, richer document representations can lead to categorization models that generalize from such limited training sets. In future research we will focus on how different combinations of linguistic features, extractions from text and concepts inferred from external knowledge bases help to improve document representation for text categorization. Another prospective research direction will be to investigate the effectiveness of the one-class approach, when the classifier is trained on 'positive' examples only, and tested on both 'positive' and 'negative' examples. This should provide us with the solution that will better fit the context of our project, where the subject domain of 'off-topic' examples is not known beforehand. Our preliminary experiments with the one-class LibSVM showed promising results.

Overall, the research reported herein holds potential for providing the IC with the analytic tools to assist humans in detecting early indicators of malicious insider activity; as well as, when applied in the broader context, in categorizing vast document collections.

Acknowledgements

This work was supported by the Advanced Research and Development Activity (ARDA).

References

Center for Natural Language Processing (CNLP). www.cnlp.org.

Center for Nonproliferation Studies (CNS). http://cns.miis.edu/.

LibSVM. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Anderson, J. 1980. Computer Security Threat Monitoring and Surveillance. James P. Anderson Co.

Research and Development Initiatives Focused on Preventing, Detecting, and Responding to Insider Misuse of Critical Defense Information Systems: Results of a Three-Day Workshop. http://www.rand.org/publications/CF/CF151/CF151.pdf.

Bengel, J., Gauch, S., Mittur, E., and Vijayaraghavan, R. 2004. ChatTrack: Chat Room Topic Detection Using Classification. *In Proceedings of the Second NSF/NIJ Symposium on Intelligence and Security Informatics (ISI2004)*.

Burgoon, J., Blair, J., Qin, T., and Nunamaker, J., Jr. 2003. Detecting Deception Through Linguistic Analysis. *In Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics*, Tucson, Arizona.

DelZoppo, R., Brown, E., Downey, M., Liddy, E. D., Symonenko, S., Park, J. S., Ho, S. M., D'Eredita, M., and Natarajan, A. 2004. A Multi-Disciplinary Approach for Countering Insider Threats. *In Proceedings of the Workshop on Secure Knowledge Management (SKM)*, Amherst, NY.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. 1998. Inductive Learning Algorithms and Representations for Text Categorization. *In Proceedings of the Seventh International Conference on Information and Knowledge Management*, Bethesda, Maryland, United States.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Jain, A. K., Duin, R. P. W., and Mao, J. 1999. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(1): 4-37.

Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines*: Kluwer Academic Publishers.

Lawrence, R. H. and Bauer, R. K. 2000. AINT misbehaving: A taxonomy of anti-intrusion techniques. http://www.sans.org/resources/idfaq/aint.php.

Liddy, E. D. 2003. Natural Language Processing. *Encyclopedia of Library and Information Science*. New York: Marcel Decker, Inc.

Platt, J. C. 1999. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. *Advances in Neural Information Processing Systems* 11: 557-563.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1): 1-47. citeseer.ist.psu.edu/article/sebastiani99machine.html.

Sreenath, D. V., Grosky, W. I., and Fotouhi, F. 2003. Emergent Semantics from Users' Browsing Paths. *In Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics*, Tucson, AZ, USA: Springer-Verlag.

Stolfo, S., Hershkop, S., Wang, K., Nimeskern, O., and Hu, C. 2003. Behavior Profiling of Email. *In Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics*, Tucson, AZ, USA.

Sundheim, B. and Irie, R. 2003. Gazetteer Exploitation for Question Answering: Project Summary.

Symonenko, S., Liddy, E. D., Yilmazel, O., DelZoppo, R., Brown, E., and Downey, M. 2004. Semantic Analysis for Monitoring Insider Threats. *In Proceedings of the Second NSF/NIJ Symposium on Intelligence and Security Informatics.* Tucson, AZ.

Twitchell, D. P., Nunamaker Jr., J. F., and Burgoon, J. K. 2004. Using Speech Act Profiling for Deception Detection. *In Proceedings of the Second NSF/NIJ Symposium on Intelligence and Security Informatics*, Tucson, AZ.

Upadhyaya, S., Chinchani, R., and Kwiat, K. 2001. An Analytical Framework for Reasoning About Intrusions. *In Proceedings of the 20th IEEE Symposium on Reliable Distributed Systems*.

van Rijsbergen, C. J. 1979. *Information Retrieval*. London: Butterworth.

Yang, Y. and Liu, X. 1999. A Re-Examination of Text Categorization Methods. *In Proceedings of the 22nd Annual International SIGIR*, Berkeley, CA.

Zhou, L., Burgoon, J. K., and Twitchell, D. P. 2003. A Longitudinal Analysis of Language Behavior of Deception in E-mail. *In Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics*, Tucson, AZ.