2005

Improved Document Representation for Classification Tasks for the Intelligence Community

Elizabeth D. Liddy
Syracuse University

Ozgur Yilmazel
Syracuse University

Svetlana Symonenko
Syracuse University

Niranjan Balasubramanian
Syracuse University

Recommended Citation

Improved Document Representation for Classification Tasks for the Intelligence Community

Description/Abstract

This research addresses the question of whether the AI technologies of Natural Language Processing (NLP) and Machine Learning (ML) can be used to improve security within the Intelligence Community (IC).

Keywords

security, AI, Natural Language Processing, NLP, Machine Learning, ML, Intelligence Community, text categorization

Disciplines

Library and Information Science | Linguistics

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

Background and Problem Area

This research addresses the question of whether the AI technologies of Natural Language Processing (NLP) and Machine Learning (ML) can be used to improve security within the Intelligence Community (IC). 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. The application of NLP-driven information extraction and ML-based text categorization is being applied to the problem of monitoring insider activity, with the goal of detecting malicious insiders within an organization (Symonenko et al., 2004). The capability is being implemented and tested as one piece of a tripartite solution in a system prototype within the context of a larger Insider Threat project being conducted under ARDA's Information Assurance for the Intelligence Community Program. That project, *A Context, Role and Semantic (CRS)-based Approach for Countering Malicious Insider Threats,* 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-based context and to distinguish between expected and anomalous user behavior. This Semantic Analysis of content being described here, is coupled in the prototype with Social Network Analysis which monitors and detects anomalies in social behavior, and Composite, Role-based Monitoring which analyzes insider activity based on organizational, application, and operating system roles (DelZoppo et al., 2004).

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 their expertise and experience. The information that is accessed and/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 TOI / AOI, communication patterns, intelligence products and information systems needed, and the intelligence work products created. Within this mission-focused context, our hypothesis is that NLP-based semantic analysis of text, combined with ML-based text categorization of 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.

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 (the Observed) are then categorized as on-topic or off-topic based on the similarity of their semantic content to the 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 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, Platt, Heckerman, & Sahami, 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. Furthermore, we hypothesize that utilizing NLP features to produce a semantic representation of the documents' content makes it possible to utilize both 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 (noun 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[1]. 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)[2] 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 the 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[3]. 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 & Irie, 2003)[4]. 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_superclass'), 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

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

---

[1] www.cnlp.org
[2] www.cns.org
[3] this simulates analyst's utilizing background knowledge or coming up with useful associations
[4] SPAWAR gazetteer, developed under the AQUAINT Program, combines resources of four publicly available gazetteers: NGA (NIMA); USGS; CIA World Factbook; and TIPSTER.
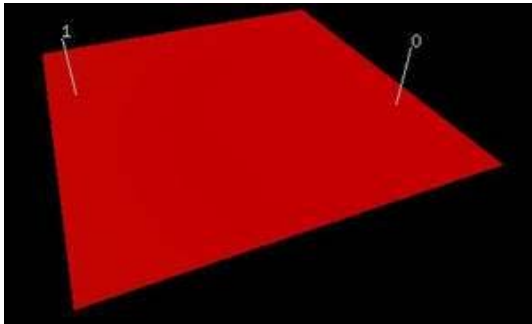
Experimentation
To assess the effectiveness of using NLP-extracted features vs. bag-of-words document representations, we ran clustering and categorization experiments on the Insider Threat dataset, using different sets of features. The dataset contains 172 documents assigned to either Expected or Observed (or New) subsets. The Expected set includes 86 documents: 67 'ON' topic and 19 'OFF' topic. The Observed set has 86 documents: 46 'ON' topic and 40 'OFF' topic.

When choosing 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, & Mao, 2000). We ran experiments using different types of information that we extracted from documents for representation with a Support Vector Machine 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 Proper names and named entities (CAT): Only the tokens that are identified as proper names or named entities from the training corpus are used for representation.
3. TOI/AOI Extractions (TOI/AOI): Only the tokens that are extracted as TOI/AOI indicators are used for representation.

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.

A 2-cluster solution for bag-of-words and TOI/AOI feature sets was generated for the Expected dataset by running repeated bisections algorithm in CLUTO[5]. As Figure 2 shows, clustering on TOI/AOI concepts, as opposed to "bag-of-words", produces a more distinct separation of topics in texts and, contributes to a more accurate conceptual modeling of the entire dataset.
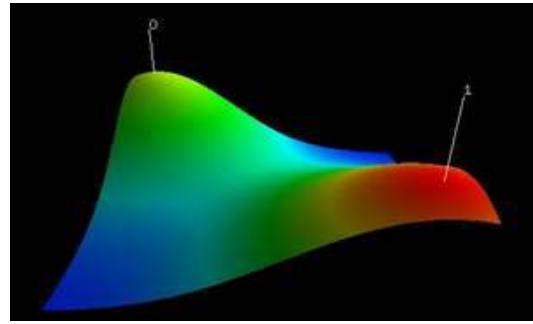
BOW representation   TOI/AOI representation
Figure 2. Two-cluster model on different sets of features

Categorization experiments were run in WEKA[6] using the Expected dataset to train the model and the Observed dataset to test it.

We have chosen SVMs because its been empirically shown that SVM performs better than kNN, Naïve Bayes and some other classifier on Reuters collection (Y. Yang, 1999), also Joachims (2001) explains theoretically why SVMs are appropriate for text categorization

	#Attributes	Accuracy	Precision	Recall	F-Measure
BOW	6538	0.9534	0.953	0.955	0.9555
CAT	3925	0.9534	1	0.9130	0.9545
TOI/AOI	492	0.9888	0.9787	1	0.989

Table 1. Categorization results for the three different feature sets.

---

5 http://www-users.cs.umn.edu/~karypis/cluto/index.html

6 http://www.cs.waikato.ac.nz/ml/weka

Initial experiments we have conducted on our development set show (see Table 1) that using more sophisticated features to represent documents leads to higher precision and recall in categorizing documents as ON or OFF topic, compared to the bag-of-words representation. Experiments using TOI/AOI extractions achieved greater accuracy with twelve times fewer features. Using extracted entities and named entities for representation achieved similar results to BOW representation with half the number of features.

These results show that the use of NLP-extracted features and NLP-based infetencing helps to improve performance in categorization and leads to clusters with higher intra-cluster similarity and lower inter-cluster similarity.

In summary, the use of NLP-extracted features for categorization/clustering provides three main advantages.
1. Improvements in effectiveness – categorization using NLP-extracted 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 for generation of categorizing/clustering models.

Conclusion

Document representation using sophisticated NLP-extracted features improved text categorization effectiveness and efficiency with SVMs. The amount of available training documents is limited in homeland security and counter intelligence environments. Improved document representations can lead to categorization models that generalize from such limited training sets. In future research we will identify how different combinations of linguistic features, extractions from text and concepts inferred from external knowledge bases helps to improve document representation for text categorization.

References

DelZoppo, R., Brown, E., Downey, M., Liddy, E. D., Symonenko, S., Park, J. S., Ho, S. M., D'Eredita, M., Nataraian, A. (2004). *A multi-disciplinary approach for countering insider threats.* Workshop on Secure Knowledge Management (SKM), Amherst,NY.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). *Inductive learning algorithms and representations for text categorization.* 7th ACM International Conference on Information and Knowledge Management, Bethesda,US.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence, 22*(1), 4-37.

Liddy, E. D. (2003). Natural language processing. In *Encyclopedia of library and information science 2nd ed.*New York: Marcel Decker Inc.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1--47.

Sundheim, B., & Irie, R. (2003). Gazetteer exploitation for question answering, project summary, November 2003.

Symonenko, S., Liddy, E. D., Yilmazel, O., Zoppo, R. D., Brown, E., & Downey, M. (2004). *Semantic analysis for monitoring insider threats.* The 2nd NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 2004). Tucson,AZ.

Y. Yang, X. L. (1999). *A re-examination of text categorization methods.* Proceedings of SIGIR'99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkley,US.