Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Exploiting Crowd-Based Labels for Domain Focused Information Retrieval

J. Cory Miniter, Vineet Mehta, Kavitha Chandra
Center for Advanced Computation and Telecommunications
Department of Electrical and Computer Engineering University of Massachusetts Lowell
One University Avenue, Lowell, MA 01854

Abstract

Information search and retrieval from online sources or social forums is often performed with term based boolean queries. Such queries can produce low relevance documents in situations where the user is interested in retrieving information related to a concept, or belonging to a specific domain. In this work an approach for concept-based information retrieval is presented, which exploits word and document distributions derived from topic modeling performed on data from online sources. Documents acquired from the Reddit and Stack Exchange online social forums are used for extracting concepts, and subsequently training and testing a detector that aids in identifying and retrieving documents associated with the concept of interest. The selection of training sets for our concept based detector is aided by pre-partitioning of documents by online users (or crowd) into concept focused sub-forums, such as sub-reddits. Topics derived from a sample of the overall document set are taken to represent concepts. These topics then form the basis for identifying sub-forums that have a strong correspondence with the concept of interest, and documents within are assigned (noisy) binary labels. The applicability of our approach is demonstrated by creating a domain focused detector for Cyber Security content from Reddit data. The cross utility of this detector is demonstrated by successfully retrieving relevant Cyber Security documents from an alternate test online source: Stack Exchange. Document classification results of the proposed approach are compared favorably with classifications performed by human analysts.

Introduction

The growing volume of documents in online media and forums makes them ideal sources for mining historical and emerging information about domains of interest. For example, in just the last two years, the number of Reddit posts increased from 80 to 190 million and the number of comments grew from nearly 69 million to 1.7 billion. The prominence of online media and the availability of large volume of documents provides the potential for high relevance domain specific queries. Keyword or term based searches continue to serve as the dominant technique for information retrieval from such document sets. It has been noted however that the precision and recall performance of this approach is limited (Haav and Lubi 2001). Concept based information retrieval approaches have emerged as alternatives that promise enhanced performance (Haav and Lubi 2001),(Baziz, Boughanem, and Traboulsi 2005),(Lin, Chi, and Hsieh 2012),(Boubekeur and Azzoug 2013). These approaches often rely upon the availability of lexical ontology such as WordNet (Miller 1995) for specification of concepts. The construction of such ontology can require considerable manual effort.

Recently topic modeling has been employed for information retrieval (Wei 2007). Topic modeling can be viewed as a means for inferring concepts embedded in document sets by exploiting co-occurrence of terms that is a consequence of the documents' semantic structure. This approach has the advantage of providing an unsupervised automated means for discovering concepts. However, there can be differences in concepts discovered through topic modeling and those interpreted by human subjects (Boyd-Graber et al. 2009).

Information retrieval based on the use of classification techniques has also been studied (Manning, Raghavan, and Schütze 2008). This approach requires the assignment of labels for training a classifier to filter relevant documents. Though this process may be done manually using a trained expert, approaching the problem this manner is time consuming and resource intensive. The performance of classification techniques hinge on availability of an adequate amount of accurately labeled training data. The issue of label noise in classification problems has been discussed in detail by Frénay and Verleysen (Frénay and Verleysen 2014).

The rise of social media platforms has led to enormous increases in the amount of user generated text data, much of which may be beneficial to classify. Text data generated in the context of social media platforms presents a unique challenge to text classification schemes. Posts can often be short in length, low in content, and often contain multiple different themes. This data is often constrained only by the loosest of stylistic conventions, and the content is generally self-policed. Nevertheless, previous studies have noted that information retrieval tasks can be successfully carried out even with imprecise labels, such as when labeling is performed through crowd sourcing (Snow et al. 2008).

In this paper we consider a hybrid approach for enhanced domain related information retrieval from online forums. We employ topic modeling for learning concepts embedded in documents derived from an online social forum. The *red-*

dit.com social forum provides us access to documents inherently partitioned according to domains by online users (or crowd). A user query in our case takes the form of selecting topics related to the concept of interest after examining the output of the topic model. The concept also serves as the basis for deriving a training data set by extracting documents from sub-Reddits that have high and low concentration of the topics of interest. This training set is used to train a binary classifier for discriminating documents of interest in other online social forums.

The main contributions of the paper are: (i) application of topic modeling in deriving domain specific concepts (a topic or relationships between topics) from sampled data from an online social forum, (ii) use of topic based concepts to reveal and select sub-forums with high concentration of relevant documents for fashioning (imprecisely) labeled training data, (iii) construction and evaluation of a concept related document classifier from imprecisely labeled training data, as well as objective and subjective assessment of its performance.

The rest of this paper is organized as follows. Section II describes the objectives and problem formulation. The data sets analyzed are described in Section III. Section IV discusses the use of topic modeling to derive concepts, as well as means for assessing the concentration of concept related documents in a sub-forum. Section V presents the binary classifier trained on Reddit data. Section VI evaluates the classifier performance for intra-source and cross-source data. This section also compares the classifier results with document classification performed by human analysts. Section VII concludes the paper.

Problem Description

This work focuses on circumstances in which a user requires information associated with a domain or concept that is often referred to by a term or phrase, without necessarily a literal reference to the query term or phrase. The working example we employ in this paper is that of the concept *Cyber Security*. The user is likely interested in documents related to this domain, such as malware, vulnerabilities, and phishing, rather than ones that explicitly contain the term *Cyber*. A search query of the term *Cyber* using popular Web indexes returns documents that explicitly contain the search term(s). The user can of course issue queries using terms related to the concept if the user is familiar with these. This approach can be cumbersome for the user who may need to attempt a variety of multi-term Boolean query expressions to focus the search towards relevant content. The formulation of such queries also assumes the user is knowledgeable about the term vocabulary for the document repository being searched.

In this work we propose an approach that allows the user to perform a concept based information query over a document set, rather than attempting common-place term based Boolean queries. We begin with the premise that users are able to identify the concept or domain aspect of interest from an appropriate grouping of terms when they are presented, while it might be difficult for the user to assemble such a grouping without assistance. A grouping of terms associated with a concept can be inferred from a document set in an unsupervised manner using techniques for topic modeling, such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). The choice of source document set that serves as the oracle for inferring concepts is driven by an interest in retrieving domain relevant documents from online social media and forums. The topics of interest to the user serve as a basis for filtering the document collection to derive a data set for training a concept focused binary classifier. We hypothesize that a classifier trained using documents from a single source oracle can be effective in detecting documents relevant to the concept of interest from other online sources. A functional description of our proposed approach is captured in Fig. 1.

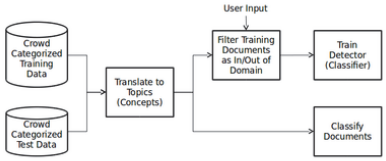

Figure 1: Functional architecture of concept based information retrieval

Description of Data

The training and test data are comprised of user generated text content from two popular online forums: Reddit and Stack Exchange. Content within these forums is divided into sub-forums according to subject matter. As a convenience for our tasks of concept based detector construction and assessment, we have chosen to acquire training and test data from sub-forums that have a connection to the domain of interest, as well as outside the domain of interest, based on subjective examination. Twenty sub-forums were examined from each site (*reddit.com* and *stackexchange.com*), with ten sub-forums from each subset designated a priori as in-domain and out-of-domain. For the purposes of this paper, *in-domain* is defined as any subject related to the computer and network security. *Out-of-domain* encompasses all subjects that we perceive as not directly cyber related. Each document in the data set is composed of a single post or comment. Only documents of length greater than 200 words are considered in this study. The data is processed to remove common "stop-words" and punctuation, as well as to reduce inflections to root-words.

The Reddit data set is comprised of 4626 text documents drawn from the 20 sub-Reddits shown in Table 1. The Stack Exchange data set contains 86887 text documents drawn from the sub-forums shown in Table 2.

Concepts from Topic Modeling

Concepts are viewed as derivations of topics inferred from the Reddit data set using Latent Dirichlet Allocation (LDA), originally proposed by Blei *et al.* (Blei, Ng, and Jordan 2003). LDA treats the documents as *bags-of-words*, where

Table 1: Sub-Reddit to class mapping

in-domain	asknetsec, netsec, blackhat, reverseengineering, crypto, hackbloc, malware, pwned, rootkit, darknetplan
out-of-domain	worldnews, science, movies, technology, music, books, television, sports, fitness, trees

Table 2: Stack Exchange to class mapping

in-domain	cs, dba, crypto, programmers, datascience, codegolf, reverseengineering, codereview, security, networkengineering
out-of-domain	hinduism, cooking, movies, fitness, parenting, christianity, outdoors, sports, gardening, history

language structure does not play a role, apart from the tendency of words in similar documents to co-occur. The problem of extracting topics is formulated using a Bayesian inference framework. A detailed description of the model parameter estimation is provided by Heinrich (Heinrich 2005). In this approach, the algorithm estimates K topic-word distributions and M document-topic distributions, based on a generative model for the joint probability distribution function $p(w, z, \boldsymbol{\theta}_d, \boldsymbol{\phi}_k|\boldsymbol{\alpha}, \boldsymbol{\beta})$ where w and z represent the discrete word and latent topic random variables. The words are drawn from a vocabulary of fixed size. LDA employs the generative assumption where each word in a document is drawn from a topic multinomial distribution with parameter vector $\boldsymbol{\phi}_k$, for the k-th topic. The event of selecting the k-th topic is given by the hidden random variable $z = k$, drawn from a multinomial distribution with parameter vector $\boldsymbol{\theta}_d$. The topics $\boldsymbol{\phi}_k$ and the per-document distribution of topics $\boldsymbol{\theta}_d$ are drawn from Dirichlet distributions with hyperparameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ respectively. The task of estimating the hidden parameters $\{z, \boldsymbol{\theta}_d, \boldsymbol{\phi}_k\}$ is assisted by exploiting the dependence structure implicit in the generative model. The computation of estimates for the hidden parameters is performed using the Gibbs-Sampling technique implemented in the machine learning library Mallet (McCallum, Graham, and Milligan 2012).

Topics generated using text from online forums can support multiple information retrieval tasks of interest to an analyst. Examples of such tasks are: i) identifying topics with top words of highest interest, ii) identifying documents corresponding to high interest topics, and iii) identifying meta data or labels with strong association to topics of interest. The dimensionality reduction ability of topic models can be leveraged to project documents from terms in social forums to low-dimensional feature vectors that are used in creating a classification scheme.

Topic models were derived by considering the number of topics $K = 20, 40$ and 60. By examining the top words that characterized each topic, a subjective classification of topics as belonging to in-domain and out-of-domain was conducted. For $K = 20$, six topics out of twenty are identified to capture the in-domain concepts. A sample of three topics are shown in Table 3.

Table 3: Subjective labeling of selected topics

Topic	Class	Top Words
Topic 20(I)	In-Domain	key encrypt password
Topic 20(II)	Out-of-Domain	movi film realli
Topic 20(III)	In-Domain	server user secur

Table 4: Top words of sample topics for K=40 and K=60

Selected in-domain topics for 40-topic model	
40(I)	key bit random encrypt hash attack algorithm messag generat secur crypto
40(II)	code exploit malwar run program function file window execut tool call
Selected in-domain topics for 60-topic model	
60(I)	bit key algorithm attack ae crypto comput implement secur hash time
60(II)	exploit malwar vulner attack bug tool find ani report hack test

The top words characterizing two in-domain topics for $K = 40$ and 60 topics are shown in Table 4. Comparison of the top words occurring in the in-domain topics for K=20,40, and 60 show strong overlap, although the ordering of the words can vary. This suggests that similar concepts can be inferred robustly over a broad range in the parameter K.

The ability to clearly distinguish topics as belonging to in-domain and out-of-domain based on our understanding of the word groupings provides an initial subjective validation of the performance of the topic model. An excerpt of text from a test document that is assigned the highest weight for the topic 20(I) (The first row shown in Table 3) is shown in Table 5. The K element topic weight vector $\boldsymbol{\theta}_d$ associated with each document d serves to identify high relevance documents associated with a selected topic, as demonstrated in the example above. The topic probability associated with a document can be further analyzed to quantitatively gauge the document composition of sub-Reddits given a topic of interest. The composition of the sub-Reddit relative to a topic k can be measured as the concentration of in-domain documents $\rho_S(k) = \sum_{d \in S} \theta_d(k)/N_S$ where $\theta_d(k)$ is the probability of topic k in document d, and N_S is the number of documents in the sub-Reddit S. The in-domain concentration ρ_S for topic 20(I) is plotted in Fig. 2 in descending order of $\rho_S(k)$ and as a function sub-Reddit label S. The results show the highest concentration of documents related to topic 20(I) is found in the *crypto* sub-Reddit.

Table 5: High ranking test document for topic 20(I)

Title: *Why don't we use provably secure hashes. . .?*	
User 5:	multiple reasons complexity. we want our password hashing scheme to be simple, so implementing it is easy, less error prone, can be effectively done in hardware or small embedded systems, etc. the last thing we want is to have elliptic curves or huge finite field arithmetic. it is not about speed. it is about optimization. hashing passwords will take a preset amount of time, no matter what algorithm you use...

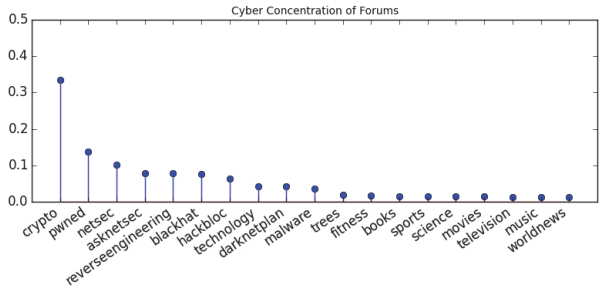

Figure 2: Concentration of documents related to topic 20(I)

Text Classification

In this section we treat the detection problem as a binary classification problem, and assign the labels $\{$*in-domain*, *out-of-domain*$\}$ to two classes. The training data for the classification problem is derived from the Reddit data set. 10 sub-Reddits are assigned to the *in-domain* class and 10 to the *out-of-domain* class. We note that in this case there is nearly exact correspondence between the sub-Reddits selected subjectively as *in-domain*, and those identified as high domain concentration sub-forums by our metric ρ_S. The only outlier is *technology*, which has been included in part because it is expected to host in-domain and out-of-domain documents. The documents in each sub-Reddit are labeled based on the mapping supplied in Table 1. This labeling approach is expected to be imprecise. Nevertheless, the aim here is to develop a detector that can perform coarse grain filtering and significantly reduce the volume of data a human analyst has to sift through to discover domain related entries.

A straightforward means to constructing a domain detector is to use the labeled data to train a classifier. This requires the specification of feature vectors for a training data set. In this work, the feature vectors correspond to document-topic distributions derived from training a topic model using LDA. The top plot in Fig. 3 shows the relative misclassification percentage of a random forest classifier trained on a randomly sampled set of 80% of the 4626 total documents of the 20 sub-Reddits given in Table 1. The bottom plot in Fig. 3 shows the number of test documents in each of the sub-Reddits. The trailing (-i) and (-o) on the sub-Reddit name specifies the class label *in-domain* and *out-of-domain*

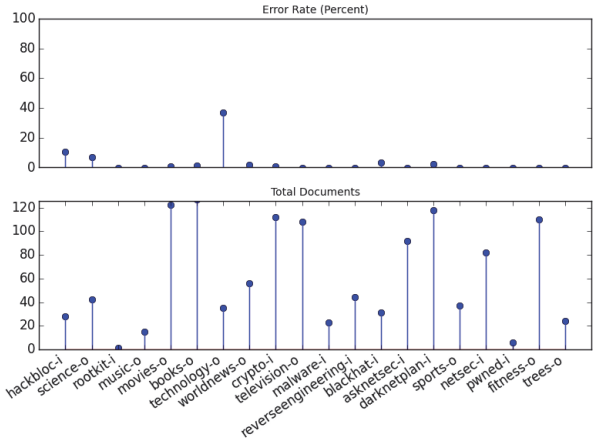

Figure 3: Performance of random forest classifier for 20 sub-Reddit data set, using 20-topic feature vectors

for that sub-Reddit. The highest misclassification rate is found to occur in the *technology* sub-Reddit, at 78%. An examination of misclassified documents in this sub-Reddit reveals many documents where the discussion may be construed as *in-domain* (i.e cyber) related. This sub-Reddit is an example where subjective classification may have a level of uncertainty, which is reflected in the performance of the classifier. The error performance was also examined with increase in the number of topics. The misclassification percentage considering all in-domain and out-of-domain testing data (965 documents) was found to increase with K but minimally, producing 2.5, 2.7, 2.75, 2.8 and 3% error for $K = 10, 20, 40, 60, 80$ respectively. In the rest of this paper, we consider $K = 20$, motivated in part to maintain an equal correspondence with the number of chosen sub-Reddits.

Performance of Text Classifier

In this section, the effect of restricting the source of in-domain training data for the classification process is explored. In the previous section all ten sub-Reddits that were considered as in-domain were included in the training data. In this section, we examine the effect of considering top three sub-Reddits *crypto*, *netsec*, and *pwned* that capture the highest concentration of Topic 20(I) as shown in Fig. 2. This topic is of interest due to its relevance to the cyber-security concept that is the domain of interest in this work. All out-of-domain data sets were however included in the training set.

Fig. 4 represents the error rate considering the three in-domain sub-Reddits and 80% training data. Comparing this result shown in Fig. 3 where all ten in-domain sub-Reddits were included, we see that documents in the technology group show a lower misclassification rate, while some of the in-domain sub-Reddits such as *asknetsec* and *darknetplan* exhibit a small increase in misclassification. The improved error performance for the technology sub-Reddit is not altogether unexpected, as the smaller scope of the training data provides tighter constraints on the documents that are con-

sidered as *in-domain*.

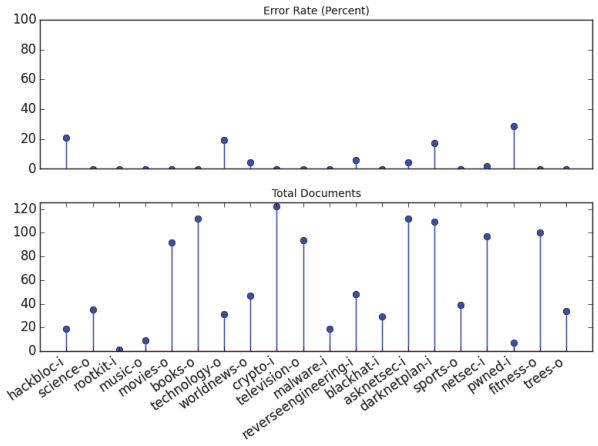

Figure 4: Performance of random forest classifier trained with only 3 in-domain sub-Reddit data

A. Cross-Source Classification: The cross-source classification utilizes 80% of the Reddit data for training the classifier and 100% of the Stack Exchange data for testing. Each Stack Exchange sub-forum is labeled as {*in-domain*, *out-of-domain*} as given in Table 2. The performance of the classifier can be seen in Fig. 5. The bottom plot shows the number of total documents within each sub-forum, and the top plot shows the misclassification percentage. The random forest classifier displays a more uniform performance in the error rate with errors across all of the sub-forums shown to be close to or less than 10%, except for one sub-forum titled *se.parenting* that was found to have documents with nearly 20% error rate in classification. The uniformity in error performance may be the result of the larger number of documents used in this analysis 86,887 compared to 995 Reddit documents used in the intra-source study. The result highlights the fact that the classifier is robust enough to reliably classify documents outside of its training source.

B. Comparison with human enabled classification: Since the approach described in this paper is intended to provide human analysts with a tool for quickly sifting through large volumes of data, the classifier performance is compared to results from human analysts. A group of 8 volunteer analysts, graduate students in electrical engineering, was recruited for this effort. The volunteers had no prior involvement with the project, and no subject matter expertise. They were given brief instructions on the in-domain and out-of-domain tagging scheme, along with several examples of documents from each category. The test data set for this experiment consisted of 374 documents sampled from the Reddit data set. Each sub-Reddit was masked with a label from A to T. Each analyst received randomly selected documents from between two and three sub-Reddits. They categorized each document within their test set as either in-domain or out-of-domain, based on their own interpretation of the content of the document. The test set was also classified using two of the classifiers discussed in previous sections, fully trained

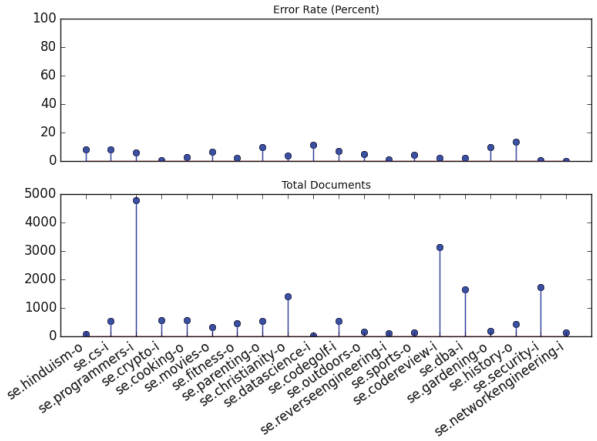

Figure 5: Performance of random forest classifier for Stack Exchange data set, using 20-topic feature vectors

using all 20 sub-Reddits, and a restricted training set using only the top three in-domain sub-Reddits including all ten out-of-domain sub-Reddits. The computer generated classifications were compared to the classifications from the human analysts.

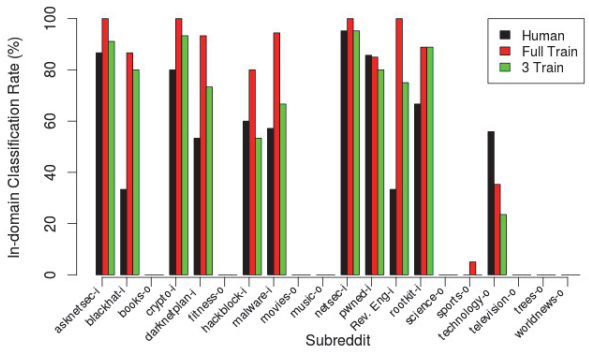

Figure 6: Performance of Classifier as Compared to Human Analysis

Fig. 6 shows the classification rate of documents in each sub-Reddit as belonging to in-domain. The first bar in each group refers to human classification, the second bar is the fully trained model and the third bar is training with three top sub-Reddits. The classification rate of zero for eight of the out-of-domain sub-Reddits shows that all three methods classify these documents correctly. *Technology* results in misclassifications from all three methods, while *science* also shows a small misclassification rate on the part of the fully-trained classifier. In summary, the classifier trained using three in-domain sub-Reddits closely reflected the human analysts' labeling in most cases.

Table 6 shows the correlation between human classification and each of two automated classifiers. This was com-

Table 6: Correlation Between human and automatic classifications

	A-i	B-i	C-o	D-i	E-i	F-o	G-i
Full Train	0.867	0.467	1	0.8	0.6	1	0.8
3-Train	0.867	0.4	1	0.8	0.667	1	0.933
	H-i	I-o	J-o	K-i	L-i	M-i	N-i
Full Train	0.762	1	1	0.952	0.905	0.364	0.667
3-Train	0.81	1	1	0.952	0.905	0.394	0.667
	O-o	P-o	Q-o	R-o	S-o	T-o	
Full Train	0.9	0.95	0.676	0.952	1	1	
3-Train	1	1	0.765	1	1	1	

puted by comparing the classification results from each pair of classifiers (Full-Train, Human, 3-train, Human) for each document in each sub-Reddit and assigning a value of value of one for each document that was classified correctly by both classifiers. A high correlation between human analysis and machine classification is desirable, as we wish the classifier to behave in a predictable manner. The correlation metric $\varrho_{ij}(S) = 1/N_S \sum_{k=1}^{N_S} u(i,j;k)$, where the indicator function $u(i,j;k) = 1$ if the k^{th} document in sub-Reddit S was correctly classified by both classifiers i and j. In-domain sub-Reddits B and M corresponding to *blackhat* and *reverseengineering* show a low correlation between human and random-forest classifiers. This may be attributed to the subjective understanding of our human analysts, who may not have in-depth knowledge these topics. On the other hand, human classifications of documents from sub-Reddits (A:asknetsec, D:crypto, G:hackbloc, H:malware, K:netsec, L:pwned) were found to be well correlated with the automated classifications. This result can be attributed to the words in these documents being more accessible as cyber related to our analysts.

Conclusion

This paper has presented a method for exploiting imprecisely labeled documents in online social forums for the purpose of domain focused information retrieval. We demonstrate the use of topic modeling in learning concepts from the online content source Reddit. The learned topics can be successfully used to identify sub-forums within Reddit that have strong or weak concentration of documents related to the concept of interest. Although the sub-forums selected in this manner are imprecise labels, they can nevertheless be effective in discriminating domain relevant content in other online forums. Though the labeling is necessarily imprecise, the binary classifier shows accuracy given a large enough training set. The classifier is also robust enough to classify data not directly related to its training data, as shown in the cross-source classifier test. This is demonstrated using selected document sets from the StackExchange forum. Further, the classifier shows similar behavior to untrained human analysts, which suggests a good degree of accuracy in individual document classifiation. Future work may expand the classifier beyond the binary case, creating an automatic tagging scheme for multiple concepts of interest.

Acknowledgments

The support of J.C. Miniter as a NSF GK-12 fellow through grant DGE #0841392 is gratefully acknowledged.

References

Baziz, M.; Boughanem, M.; and Traboulsi, S. 2005. A concept-based approach for indexing documents in ir. In *INFORSID*, 489–504.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *J. Machine Learning Research* 3:993–1022.

Boubekeur, F., and Azzoug, W. 2013. Concept-based indexing in text information retrieval. *CoRR* abs/1303.1703.

Boyd-Graber, J.; Chang, J.; Gerrish, S.; Wang, C.; and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Info Processing Systems*, 288–296.

Frénay, B., and Verleysen, M. 2014. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems* 25(5):845–869.

Haav, H., and Lubi, T. 2001. A survey of concept-based information retrieval tools on the web. In *5th East-European Conference, ADBIS 2001*, 29–41.

Heinrich, G. 2005. Parameter estimation for text analysis. Technical report, Technical report.

Lin, H.; Chi, N.; and Hsieh, S. 2012. A concept-based information retrieval approach for engineering domain-specific technical documents. *Advanced Engineering Informatics* 26(2):349–360.

Manning, C.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*, volume 1. Cambridge University Press.

McCallum, A.; Graham, S.; and Milligan, I. 2012. Review of MALLET. *Journal of Digital Humanities* 2(1).

Miller, G. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263.

Wei, X. 2007. *Topic Models in Information Retrieval*. Ph.D. Dissertation, University of Massachusetts Amherst.