Improving Text Classification Using EM with Background Text

Sarah Zelikovitz
Computer Science Department
The College of Staten Island of CUNY
2800 Victory Blvd
Staten Island, NY 10314
zelikovitz@mail.csi.cuny.edu

Haym Hirsh
Computer Science Department
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08855
hirsh@cs.rutgers.edu

Abstract

For many text classification tasks, sets of background text are easily available from the Web and other online sources. We show that such background text can greatly improve text classification performance by treating the background text as unlabeled data and using existing techniques based on EM for iteratively labeling this background text. Although results are most pronounced when the background text falls into categories that mirror those present in the training and test data, we show improved classification accuracy even though the use of background text violates many of the assumptions underlying the original approach, especially in the presence of limited training data.

Introduction

The abundance of digital information that is available has made the organization of that information into a complex and vitally important task. Automated categorization of text documents plays a crucial role in the ability of many applications to sort, direct, classify, and provide the proper documents in a timely and correct manner. With the growing use of digital devices and the fast growth of the number of pages on the World Wide Web, text categorization is a key component in managing information.

Applications of various machine learning techniques that attempt to solve this problem include categorization of Web pages into sub-categories for search engines, and classification of news articles by subject. Supervised machine learning programs often have the limitation that they learn based solely upon previously classified data. It is often both impractical and extremely tedious and expensive to hand-label a sufficient number of training examples to achieve the high accuracy that is needed for a given task. Given the huge proliferation of data on the Web, only a tiny percentage of which can realistically be classified and labeled, these programs are unable to exploit this information to achieve higher accuracy when faced with new unlabeled examples.

Many researchers are exploring the possibilities of incorporating unlabeled examples (Nigam *et al.* 2000; Li & Liu 2003) or test examples (Joachims 1999; 2003; Zelikovitz 2004). The question that we address is as follows: Given a text categorization task, can we possibly find some *other* data that can be incorporated into the learning process that will improve accuracy on test examples while limiting the number of labeled training examples needed? For example, suppose that we wish to classify the names of companies by the industry that it is part of. A company such as *Watson Pharmaceuticals Inc* would be classified with the label *drug*, and the company name *Walmart* would be classified as type *retail*. Although we may not have numerous training examples, and the training examples are very short, we can find other data that is related to this task. Such data could be articles from the business section of an on-line newspaper or information from company home pages. As a result of the explosion of the amount of digital data that is available, it is often the case that text, databases, or other sources of knowledge that are related to a text classification problem are easily accessible. We term this readily available information "background text". Some of this background text can be used in a supervised learning situation to improve accuracy rates, while keeping the hand-labeled number of training examples needed to a minimum.

Nigam and his colleagues have shown (Nigam *et al.* 2000) that text classification can be improved in the presence of a particular kind of background text, unlabeled data. The original training data is first used to associate a probability distribution over the possible labels for each unlabeled example. The training data is then augmented with the now probabilistically labeled data, and the process repeats, relabeling the original unlabeled data. At any point the resulting data — both original training data and newly labeled background data — can be used to label unseen test data. In this paper we show that this approach can also be used with arbitrary background text serving the role of unlabeled data. In essence, we expand upon this previous work to show that it has broader applicability than initially presented, making it possible to improve text classification with other forms of background text.

Using Naive Bayes and EM for Text Classification with Background Text

Naive Bayes and Expectation Maximization

Nigam et al. demonstrated their ideas using the naive Bayes classification method coupled with expectation maximiza-

tion (EM) to estimate the probabilities of class membership for the unlabeled data. Given a set of training documents, $\{x_1, \ldots, x_n\}$, each of which is assigned one of m classes $\{c_1, \ldots, c_m\}$, the probability of any word occurring given class c_j can be estimated from the training data. This is typically the number of times that the word occurs in this class divided by the total number of words in this class. If the document x_k consist of the words $w_{k,1}, \ldots, w_{k,d}$, then assuming that all words are independent of each other, the equation to compute the probability that this document will occur given the class c_j, can be given as follows:

$$P(x_k|c_j) = \prod_{i=1}^{d} P(w_{k,i}|c_j) \quad (1)$$

The probability of a class c_j occurring can also be estimated from the training data. This is typically the number of documents in class c_j divided by the total number of documents.

Using Bayes rule we can then specify the probability that a specific example, x_k is a member of the class, c_j:

$$P(c_j|x_k) = \frac{P(c_j) \times P(x_k|c_j)}{P(x_k)} \quad (2)$$

If we substitute the numerator of Equation 2 with Equation 1 we have:

$$P(c_j|x_k) = \frac{P(c_j) \times \prod_{i=1}^{d} P(w_{k,i}|c_j)}{P(x_k)} \quad (3)$$

During classification time this equation is used on a test document to compute the probabilities of class membership in all classes. In practice, using Equation 3 we are interested in finding $argmax_j P(c_j) \times \prod_{i=1}^{d} P(w_{k,i}|c_j)$. This is the class with the highest probability and is returned as the final classification result.

When unlabeled examples are available in addition to labeled examples, we can view this as a problem of missing data and can apply Dempster's iterative hill-climbing technique, Expectation Maximization (EM) (Dempster, Larid, & Rubin 1977). The idea is to view the labels of the unlabeled data as missing values (Dempster, Larid, & Rubin 1977) that can be approximated via EM. Nigam's approach (Nigam *et al.* 2000) uses EM in this way with a naive Bayes text classifier proceeds as follows:

- Compute the initial parameters of the classifier by using only the set of labeled examples.
- E step: Compute the probabilities of class membership for each of the unlabeled documents given the current classifier parameters. This is done by using the current version of the naive Bayes classifier.
- M step: Using the probabilities that were computed in the E step, recompute the parameters of the naive Bayes classifier.

  The E step gives the probability that each unlabeled example is classified by each class. To reestimate the probability that a class c_k occurs using both the *training* (labeled) set and the *newly labeled examples* (which were the unlabeled set) we no longer calculate the total number of documents in the class divided by the total number of documents. Rather, we calculate the sum of the *probabilities* that all documents belong in c_k divided by the total number of documents. For a document in the training corpus, this probability is equal to one if the document belongs to the class c_k, and zero otherwise. For documents in the newly labeled set this probability is equivalent to the results of the E step. To reestimate the probability that a word will occur given a specific class it is not enough to compute the number of times that the word occurs in each document that belongs to that class, but rather the number of times that the word occurs in each document multiplied by the probability that the document belongs to that class. If an unlabeled example has a non-zero probability of belonging to a specific class, it is used in the calculations for that class. In this way unlabeled examples are actually used numerous times in the recalculation of the model parameters.

The E and M steps are repeated iteratively. Our version of the algorithm iterates for a fixed number of times (seven) that was found to be useful in text classification.[1] We used the rainbow package (http://www.cs.cmu.edu/~mccallum/bow/rainbow/) (McCallum 1996) to preprocess and tokenize the data and to run naive Bayes and EM.

Unlabeled Examples vs. Background Text

At first glance it would seem that although EM might be a useful technique for aiding the classification task via unlabeled examples, the same technique would be useless when dealing with the much broader problem of using background text. This is because the assumption that the naive Bayes classifier makes is that examples (both labeled and unlabeled) have been generated by a mixture model that has a one-to-one correspondence with classes. Even if this assumption is true for the labeled data and the test data, by its very nature, background text should not fit this assumption at all. Background text often comes from a source that differs from that of the training and test data and is of a different form and different size than the training and test data.

Consider, for instance, the text categorization problem of placing advertisements into the correct area in the classified section of a newspaper. If we have a very large number of previously classified advertisements, this might be a task that is not very difficult for an automated machine learning program. However, if the labeled data is scarce, this becomes a much more difficult problem. For example, a piece of test data might be (taken from http://www.courierpost):

> toyota '99 tacoma 4x4 x cab load
> must sell 21 000 nego call joe

and belong to the class *truck*. If the set of training data is small, and the term "toyota" is not part of the training set vocabulary, this advertisement might be misclassified. If we have a set of unlabeled examples of advertisements then perhaps naive Bayes and EM could correctly approach the classification problem. However, suppose that our background

[1] We chose the number 7 based on discussions with Nigam (personal communication).

text consists of sections of advertisements from some other newspaper, where each *section* is a piece of background knowledge. One piece of background text consists of all advertisements under a specific categorization in the second newspaper. Moreover, the grouping in the second newspaper is very different than the first. For example, the second newspaper has one category called *transportation* that combines three categories of the first newspaper – *cars*, *trucks* and *boats*. This piece of background text *should* be helpful, but it clearly violates all assumptions about the generative model, and it does not fit into the classification problem that we wish to learn.

On the other hand, there are many examples where, although the form of the background text is different than the training and test data, the background text may still follow the same classification scheme as the training and test data. Consider the problem of classifying the titles of technical papers in physics by sub-fields. For example, a title (xxx.lanl.gov):

The Nature of Galaxy Bias and Clustering

would be placed in the category *astro physics*. Suppose, also, that for background text we have numerous abstracts of technical papers available. Although it is the case that these pieces of background text are not short title strings, we can still look at them as possibly falling into one of the categories for classification. Since it is the case that in text categorization all data is represented in the same way, as vectors of terms, in that sense we can still look at the background abstracts as examples with missing class information. Therefore, perhaps naive Bayes and EM would help in a case such as this. The interesting observation that we make is that to gain leverage out of unlabeled examples, the unlabeled data that we have need not be specifically and accurately unlabeled examples. As long as the vocabulary and classification structure closely resembles the training/test data, background text can improve classification accuracy in textual data using the EM algorithm.

For generative modeling of classifiers, if the structure of the classifier that is automatically learned is identical to that of the generator of the training, test and unlabeled documents then it has been shown that unlabeled documents will most definitely be helpful (Zhang & Oles 2000). However, this assumption is often unprovable or untrue, even when dealing with unlabeled examples that are extremely "similar" to the labeled data. Certainly with background text that comes from a different source than the training/test data we cannot rely on this theoretical result. Empirically we show in the next section that background text can aid classification.

Experiments and Results

We have tested our system on six distinct text-categorization tasks that we have taken from the World Wide Web. For each of these problems, the source of our background text varies, sometimes originating at the same site from which we obtained the labeled data, and sometimes from unrelated sites also found on the Web. Some of the problems have background text that is similar to the training and test sets, while the background text of some problems are not clearly classifiable at all.

Data Sets

Technical papers One common text categorization task is assigning discipline or sub-discipline names to technical papers. We created a data-set from the physics papers archive (http://xxx.lanl.gov), where we downloaded the titles for all technical papers in the first three areas in physics (astrophysics, condensed matter, and general relativity and quantum cosmology) for the month of March 1999. As background text we downloaded the abstracts of all papers in these same areas from the two previous months – January and February 1999. These background text abstracts were downloaded without their labels (i.e., without knowledge of what sub-discipline they were from) so that our learning program had no access to them. We present results on a two class problem (without quantum cosmology) and the three class problem.

Web page titles We have taken two data sets from previous work on text classification (Cohen & Hirsh 1998; Zelikovitz & Hirsh 2000). The first, NetVet (http://www.netvet.wustle.edu), included the Web page headings for its pages concerning cows, horses, cats, dogs, rodents, birds and primates. For example, a training example in the class birds might have been: "Wild Bird Center of Walnut Creek". Each of these titles had a URL that linked the title to its associated Web page. For the labeled corpus, we chose half of these titles with their labels, in total 1789 examples. We discarded the other half of the titles, with their labels, and simply kept the URL to the associated Web page. We used these URLs to download the first 100 words from each of these pages, to be placed into a corpus for background text. In total there were 1158 entries in the background text database.

Companies The second of these data sets consisted of a training set of company names, 2472 in all, taken from the Hoover Web site (http://www.hoovers.com) labeled with one of 124 industry names. We created background text from an entirely different Web site – http://biz.yahoo.com. We downloaded the Web pages under each business category in the Yahoo! business hierarchy to create 101 pieces of background text. The Yahoo! hierarchy had a different number of classes and different way of dividing the companies, but this was irrelevant to our purposes since we treated it solely as a source of unlabeled background text. Each piece of background text consisted of the combination of Web pages that were stored under a sub-topic in the Yahoo! hierarchy. Each instance in the table of background text was thus a much longer text string than the training or test examples.

Advertisements We created a data set of short classified advertisements off the World Wide Web. For the labeled set of examples, we downloaded the classified advertisements from one day in January 2001 from the Courier Post at http://www.south-jerseyclassifieds.com. The Courier Post online advertisements are divided into 9 main categories. For testing, we simply downloaded adver-

tisements from the same paper, from one day a month later, taking approximately 1000 (25%) of the examples for our test set. The background text from the problem came from another online newspaper – The Daily Record (http://classifieds.dailyrecord.com). The Daily Record advertisements online are divided into 8 categories. We treated the union of the articles from each one of these categories as a separate piece of background text. In this case, each piece of background knowledge is substantially longer than the training and test cases.

ASRS The Aviation Safety Reporting System (http://asrs.arc.nasa-.gov/) is a combined effort of the Federal Aviation Administration (FAA) and the National Aeronautics and Space Administration (NASA). We obtained data from http://nasdac.faa.gov/asp/ and our database contains the incident reports from January 1990 through March 1999. A feature that is associated with each incident is the consequence of the incident that the analyst adds to the report, with six possible values. Training and test sets consist of the *synopsis* part of each incident. The test set consists of data from the year 1999; the training set consists of all data from 1997 and 1998. For the background text, we chose all *narratives* which are much longer descriptions of the incident from the years 1990-1996. For this data set the training and test examples are shorter than the background pieces of knowledge, and the background pieces do not all fit into the categories of the text classification problem.

Thesaurus Roget's thesaurus places all words in the English language into one of six major categories: space, matter, abstract relations, intellect, volition, and affection. From http://www.thesaurus-.com, we created a labeled training/test set of 1000 words. Each word was labeled with its category. For example, *forgiveness* is classified as belonging to the category *affection* whereas *judgment* is classified as *intellect*. We obtained our background text via http://-www.thesaurus.com as well, by downloading the dictionary definitions of all 1000 words in the labeled set. The dictionary definitions explain the words by providing synonyms or example sentences, but do not include the category. Each of these dictionary definitions became an entry in our background text database.

Results

We ran naive Bayes and EM with background text on all the data sets, using the full number of training examples as well as subsets of the training examples. Each result reported for the Physics titles, NetVet, News, Business,and Thesaurus data sets represents an average of five cross-validated runs. For each cross-validated run, four-fifths of the data was used as the training set and one-fifth was used as the test set. Holding each test set steady, the number of examples in the training sets were varied. Each of the five data sets was tested with naive Bayes and EM using 20, 40, 60, 80, and 100 percent of the training data.

For data sets that had a test set that was separate from the training set (Advertisements, ASRS) we created (when enough data was available) up to 10 random training sets for each training set size.

In almost all of the data sets the inclusion of background text helps boost accuracy on the unseen test examples. This improvement is especially noticeable when there are fewer training examples. In general, as the number of training examples increases, background knowledge gives much less leverage. This is consistent with the analysis of other researchers (Nigam 2001; Cozman & Cohen 2001), who show that additional unlabeled examples are most helpful when the training examples are few. Our interesting observation is that these improvements hold even though the background text is sometimes of a very different form than the training and test example. For the business name data (Figure 3) and the advertisement data (Figure 4) the classes of the background text are known to be different than the training/test data, yet classification accuracy still improves. The generative model that EM finds need not model the domain properly, as long as the probabilities that it finds are correlated with accuracy.

Is there a significant difference between the accuracy obtained with and without background text? Each x value that is plotted in Figures 1–6 represents a different data set or size of data set on which naive Bayes and EM was run. To see if EM with background text obtains higher accuracy than naive Bayes, we ran a paired t-test, treating each data set as a separate trial, with an accuracy associated with it for naive Bayes, and one for EM. The paired t-test deals with the difference between the numbers of each pair of data and the p value gives the probability that the mean difference is consistent with zero. In this case the resulting p value was less than .01 so we were able to conclude that there is a significant difference in accuracies with and without background text.

It has been shown that although EM with unlabeled examples can sometimes help accuracy, it can sometimes hurt it as well (Nigam 2001; Cozman & Cohen 2001). Our point to note is not that EM *always* helps, but rather that it can help even when broad background text is used instead of unlabeled examples. In particular, the physics paper title problem in Figure 1 is really helped by the addition of background text. We expected this because the background text follows the exact form and classes of the training and test data. However, it was a greater surprise when the thesaurus data set in Figure 6 performed quite credibly as well.

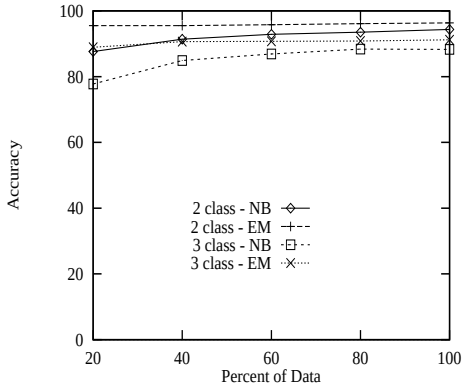

Figure 1: Naive Bayes and EM for the physics title problem

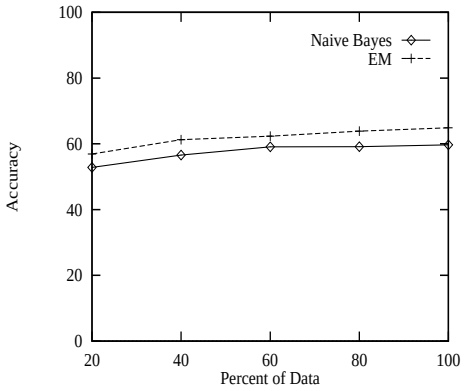

Figure 2: Naive Bayes and EM for the NetVet data

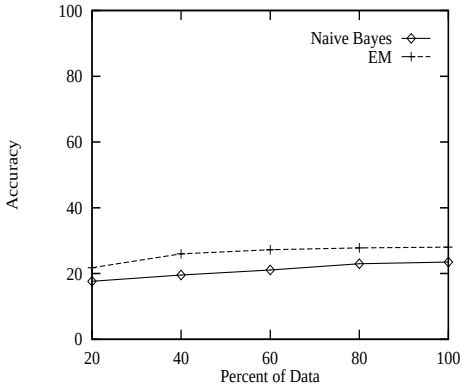

Figure 3: Naive Bayes and EM for the business name data

Intuitively, the use of EM is most appropriate when pieces of background data fit into the classes of the training and test data. This can be seen from some of the results that were graphed above. Firstly, for those domains whose background text most closely fit the data, each EM iteration usually caused accuracy to improve. This was not the case for the advertisement and business domain, where the background text is of a different form than the data. In these cases, the first and second iteration of EM had highest accuracy. Secondly, for those domains that more closely fit the background knowledge, EM helped more when there were less training examples, which is what is expected. We can see this from the physics data where accuracy rose from 87.6% to 95.5% with 20% of the data but from 94.3% to 96.3% with 100% of the data. However, in the netvet domain and business domain, for example, the improvements were the same for smaller and larger data sets, which shows the limitation of this approach.

The Nature of Background Text

Irrelevant Background Text

We explored the use of background text further for four of the domains described above: the 2-class physics problem, NetVet problem, the business problem and the thesaurus problem. We ran each of these data sets without background

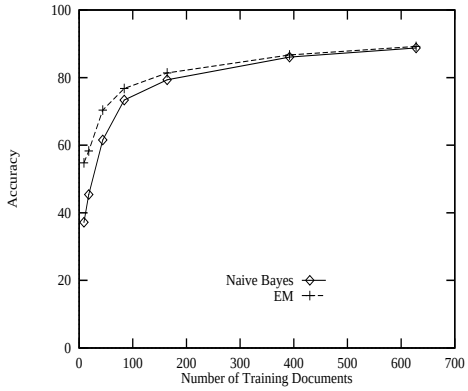

Figure 4: Naive Bayes and EM - advertisements

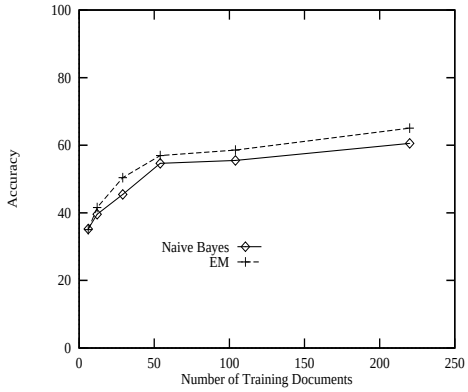

Figure 5: Naive Bayes and EM for the ASRS data

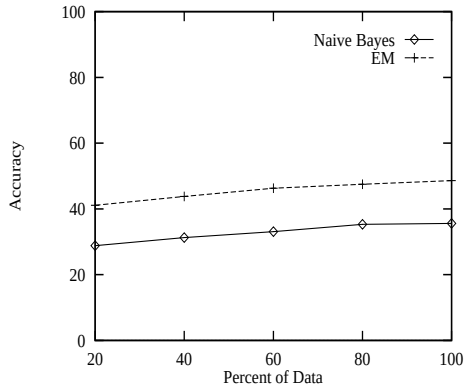

Figure 6: Naive Bayes and EM for the thesaurus problem

Table 1: Comparison of set of Background Text

Data Set	Without	Correct	Mixed	Wrong
Physics	94.3	96.3	95.7	94.6
NetVet	59.7	64.9	61.7	56.2
Business	23.5	28.2	25.2	25.9
Thesaurus	35.6	48.6	50.0	30.8

text, with the correct related set of background text, with a mixed set of background text that contained both the correct background knowledge and additional unrelated background text and with only the unrelated background text. For the unrelated background knowledge we use the background set from the NetVet data for the other three tasks, and the physics abstracts for the NetVet task. The mixed background set consists of all documents in the related background set plus all documents in the unrelated set of background text for each task. Table 1 shows the accuracy results on the full set of training data. What is interesting is that in all four cases, the mixed set of background text does not cause accuracy to be worse than Naive Bayes. In the physics data and thesaurus data, EM with mixed background performs as well as EM with the correct set of background knowledge. Even with the wrong set of background text, EM does not perform more poorly than Naive Bayes on the business names and physics data. If the iterations of EM do not classify the background knowledge as belonging with high probability to any class, it will minimize the effects that this background text will have on the final model parameters. In the NetVet and thesaurus data sets, EM with the wrong background text does perform worse than Naive Bayes. However, our version of EM is the straight forward and simple one. Nigam et al. (Nigam *et al.* 2000) present two extensions to EM that might minimize the effect of wrong background text. Specifically, if the weights of the unlabeled examples in terms of their contribution to the model parameters is reduced, misleading background text would probably have less of an effect on accuracy.

Summary and Future Work

We have substituted the use of background text for unlabeled examples in an expectation maximization algorithm, and have used numerous data sets to test the usefulness of background text. Although at first glance this might seem to be counter-intuitive, we have shown empirically that even background text that is not of the same form as the training data can provide information that allows the learner to improve accuracy on the test set. We are currently looking at methods of measuring the similarity of the background corpus to a set of training and test examples so that we can say a priori whether it would be useful to apply this or other (Zelikovitz & Hirsh 2002) methods.

Acknowledgments

We would like to thank Kamal Nigam for helpful comments and discussions.

References

Cohen, W., and Hirsh, H. 1998. Joins that generalize: Text categorization using WHIRL. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 169–173.

Cozman, F. G., and Cohen, I. 2001. Unlabeled data can degrade classification performance. *Technical Report HPL-2002-234*.

Dempster, A. P.; Larid, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 200–209.

Joachims, T. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 290–297.

Li, X., and Liu, B. 2003. Learning to classify text using positive and unlabeled data. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 587–594.

McCallum, A. K. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

Nigam, K.; Mccallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2/3):103–134.

Nigam, K. 2001. *Using Unlabeled Data to Improve Text Classification*. Ph.D. Dissertation, Carnegie Mellon University.

Zelikovitz, S., and Hirsh, H. 2000. Improving short text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 1183–1190.

Zelikovitz, S., and Hirsh, H. 2002. Integrating background knowledge into nearest-Neighbor text classification. In *Advances in Case-Based Reasoning, ECCBR Proceedings*, 1–5.

Zelikovitz, S. 2004. Transductive LSI for short text classification problems. In *Proceedings of the Seventeenth International FLAIRS Conference*, 67–72.

Zhang, T., and Oles, F. J. 2000. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 1191–1198. Morgan Kaufmann, San Francisco, CA.