A Bayesian Approach to Filtering Junk E-Mail

Mehran Sahami* **Susan Dumais**† **David Heckerman**† **Eric Horvitz**†

*Gates Building 1A
Computer Science Department
Stanford University
Stanford, CA 94305-9010
`sahami@cs.stanford.edu`

†Microsoft Research
Redmond, WA 98052-6399
`{sdumais, heckerma, horvitz}@microsoft.com`

Abstract

In addressing the growing problem of junk E-mail on the Internet, we examine methods for the automated construction of filters to eliminate such unwanted messages from a user's mail stream. By casting this problem in a decision theoretic framework, we are able to make use of probabilistic learning methods in conjunction with a notion of differential misclassification cost to produce filters which are especially appropriate for the nuances of this task. While this may appear, at first, to be a straight-forward text classification problem, we show that by considering domain-specific features of this problem in addition to the raw text of E-mail messages, we can produce much more accurate filters. Finally, we show the efficacy of such filters in a real world usage scenario, arguing that this technology is mature enough for deployment.

Introduction

As the number of users connected to the Internet continues to skyrocket, electronic mail (E-mail) is quickly becoming one of the fastest and most economical forms of communication available. Since E-mail is extremely cheap and easy to send, it has gained enormous popularity not simply as a means for letting friends and colleagues exchange messages, but also as a medium for conducting electronic commerce. Unfortunately, the same virtues that have made E-mail popular among casual users have also enticed direct marketers to bombard unsuspecting E-mailboxes with unsolicited messages regarding everything from items for sale and get-rich-quick schemes to information about accessing pornographic Web sites.

With the proliferation of direct marketers on the Internet and the increased availability of enormous E-mail address mailing lists, the volume of *junk* mail (often referred to colloquially as "spam") has grown tremendously in the past few years. As a result, many readers of E-mail must now spend a non-trivial portion of their time on-line wading through such unwanted messages. Moreover, since some of these messages can contain offensive material (such as graphic pornography), there is often a higher cost to users of actually viewing this mail than simply the time to sort out the junk. Lastly, junk mail not only wastes user time, but can also quickly fill-up file server storage space, especially at large sites with thousands of users who may all be getting duplicate copies of the same junk mail.

As a result of this growing problem, automated methods for filtering such junk from *legitimate* E-mail are becoming necessary. Indeed, many commercial products are now available which allow users to hand-craft a set of logical rules to filter junk mail. This solution, however, is problematic at best. First, systems that require users to hand-build a rule set to detect junk assume that their users are savvy enough to be able to construct robust rules. Moreover, as the nature of junk mail changes over time, these rule sets must be constantly tuned and refined by the user. This is a time-consuming and often tedious process which can be notoriously error-prone.

The problems with the manual construction of rule sets to detect junk point out the need for adaptive methods for dealing with this problem. A junk mail filtering system should be able to automatically adapt to the changes in the characteristics of junk mail over time. Moreover, by having a system that can learn directly from data in a user's mail repository, such a junk filter can be personalized to the particular characteristics of a user's legitimate (and junk) mail. This, in turn, can lead to the construction of much more accurate junk filters for each user.

Along these lines, methods have recently been suggested for automatically learning rules to classify E-mail (Cohen 1996). While such approaches have shown some success for general classification tasks based on the text of messages, they have not been employed specifically with the task of filtering junk mail in mind. As a result, such systems have not focused on the specific features which distinguish junk from legitimate E-mail. The more domain specific work along these lines has focused on detecting "flame" (e.g., hostile) messages (Spertus 1997). This research has looked specifically at particular features that are indicative of "flames", which in general are quite different than those used for junk mail filtering. Moreover, this work only makes use of domain-specific features and does not consider the full text content of messages when trying to identify a "flame".

More generally, however, we find that a rule-based approach is of limited utility in junk mail filtering. This is due to the fact that such logical rule sets usually make rigid binary decisions as to whether to classify a given message as junk. These rules generally provide no sense of a continuous *degree of confidence* with which the classification is made. Such a confidence score is crucial if we are to consider the notion of differential *loss* in misclassifying E-mail. Since the cost of misclassifying a legitimate message as junk is usually much higher than the cost of classifying a piece of junk mail as legitimate, a notion of *utility* modeling is imperative. To this end, we require, first, a classification scheme that provides a probability for its classification decision and, second, some quantification of the difference in cost between the two types of errors in this task. Given these, it becomes possible to classify junk E-mail within a Decision Theoretic framework.

There has recently been a good deal of work in automatically generating probabilistic text classification models such as the Naive Bayesian classifier (Lewis & Ringuette 1994) (Mitchell 1997) (McCallum *et al.* 1998) as well as more expressive Bayesian classifiers (Koller & Sahami 1997). Continuing in this vein, we seek to employ such Bayesian classification techniques to the problem of junk E-mail filtering. By making use of the extensible framework of Bayesian modeling, we can not only employ traditional document classification techniques based on the text of messages, but we can also easily incorporate domain knowledge about the particular task at hand through the introduction of additional features in our Bayesian classifier. Finally, by using such a classifier in combination with a loss model, we can make "optimal" decisions from the standpoint of decision theory with respect to the classification of a message as junk or not.

In the remainder of this paper, we first consider methods for learning Bayesian classifiers from textual data. We then turn our attention to the specific features of junk mail filtering (beyond just the text of each message) that can be incorporated into the probabilistic models being learned. To validate our work, we provide a number of comparative experimental results and finally conclude with a few general observations and directions for future work.

Probabilistic Classification

In order to build probabilistic classifiers to detect junk E-mail, we employ the formalism of Bayesian networks. A Bayesian network is a directed, acyclic graph that compactly represents a probability distribution (Pearl 1988). In such a graph, each random variable X_i is denoted by a node. A directed edge between two nodes indicates a probabilistic influence (dependency) from the variable denoted by the parent node to that of the child. Consequently, the structure of the network denotes the assumption that each node X_i in the network is conditionally independent of its non-descendants given its parents. To describe a probability distribution satisfying these assumptions, each node X_i in the network is associated with a *conditional probability table*, which specifies the distribution over X_i given any possible assignment of values to its parents.

A Bayesian classifier is simply a Bayesian network applied to a classification task. It contains a node C representing the class variable and a node X_i for each of the features. Given a specific instance $\mathbf{x}$ (an assignment of values $x_1, x_2, ..., x_n$ to the feature variables), the Bayesian network allows us to compute the probability $P(C = c_k \mid \mathbf{X} = \mathbf{x})$ for each possible class c_k. This is done via Bayes theorem, giving us

$$P(C = c_k \mid \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_k)P(C = c_k)}{P(\mathbf{X} = \mathbf{x})}. \tag{1}$$

The critical quantity in Equation 1 is $P(\mathbf{X} = \mathbf{x} \mid C = c_k)$, which is often impractical to compute without imposing independence assumptions. The oldest and most restrictive form of such assumptions is embodied in the Naive Bayesian classifier (Good 1965) which assumes that each feature X_i is conditionally independent of every other feature, given the class variable C. Formally, this yields

$$P(\mathbf{X} = \mathbf{x} \mid C = c_k) = \prod_i P(X_i = x_i \mid C = c_k). \tag{2}$$

More recently, there has been a great deal of work on learning much more expressive Bayesian networks from data (Cooper & Herskovits 1992) (Heckerman, Geiger, & Chickering 1995) as well as methods for learning networks specifically for classification tasks (Friedman, Geiger, & Goldszmidt 1997) (Sahami 1996). These later approaches allow for a limited form of dependence between feature variables, so as to relax the restrictive assumptions of the Naive Bayesian classifier. Figure 1 contrasts the structure of the Naive Bayesian classifier with that of the more expressive classifiers. In this paper, we focus on using the Naive Bayesian classifier,

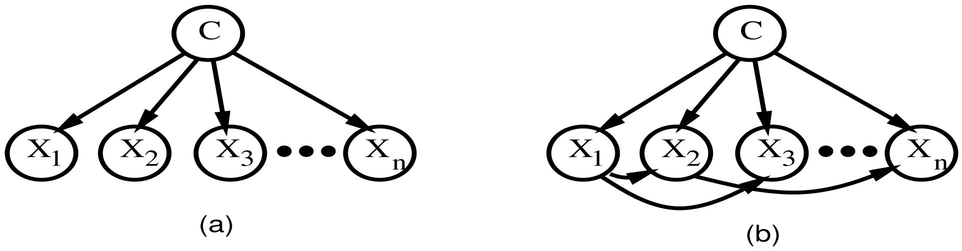

Figure 1: Bayesian networks corresponding to (a) a Naive Bayesian classifier; (b) A more complex Bayesian classifier allowing limited dependencies between the features.

but simply point out here that methods for learning richer probabilisitic classification models exist that can be harnessed as needed in future work.

In the context of text classification, specifically junk E-mail filtering, it becomes necessary to represent mail messages as feature vectors so as to make such Bayesian classification methods directly applicable. To this end, we use the *Vector Space* model (Salton & McGill 1983) in which we define each dimension of this space as corresponding to a given word in the entire corpus of messages seen. Each individual message can then be represented as a binary vector denoting which words are present and absent in the message. With this representation, it becomes straight-forward to learn a probabilistic classifier to detect junk mail given a pre-classified set of training messages.

Domain Specific Properties

In considering the specific problem of junk E-mail filtering, however, it is important to note that there are many particular features of E-mail beside just the individual words in the text of a message that provide evidence as to whether a message is junk or not. For example, particular phrases, such as "Free Money", or over-emphasized punctuation, such as "!!!!", are indicative of junk E-mail. Moreover, E-mail contains many non-textual features, such as the domain type of the message sender (e.g., `.edu` or `.com`), which provide a great deal of information as to whether a message is junk or not.

It is straight-forward to incorporate such additional problem-specific features for junk mail classification into the Bayesian classifiers described above by simply adding additional variables denoting the presence or absence of these features into the vector for each message. In this way, various types of evidence about messages can be uniformly incorporated into the classification models and the learning algorithms employed need not be modified.

To this end, we consider adding several different forms of problem-specific information as features to be used in classification. The first of these involves examining the message text for the appearance of specific phrases, such as "FREE!", "only $" (as in "only $4.95") and "be over 21". Approximately 35 such hand-crafted phrases that seemed particularly germane to this problem were included. We omit an exhaustive list of these phrases for brevity. Note that many of these features were based on manually constructed phrases used in an existing rule set for filtering junk that was readily outperformed by the probabilistic filtering scheme described here.

In addition to phrasal features, we also considered domain-specific non-textual features, such as the domain type of the sender (mentioned previously). For example, junk mail is virtually never sent from `.edu` domains. Moreover, many programs for reading E-mail will *resolve* familiar E-mail address (i.e. replace `sdumais@microsoft.com` with `Susan Dumais`). By detecting such resolutions, which often happen with messages sent by users familiar to the recipient, we can also provide additional evidence that a message is not junk. Yet another good non-textual indicator for distinguishing if a message is junk is found in examining if the recipient of a message was the individual user or if the message was sent via a mailing list.

A number of other simple distinctions, such as whether a message has *attached* documents (most junk E-mail does not have them), or when a given message was received (most junk E-mail is sent at night), are also powerful distinguishers between junk and legitimate E-mail. Furthermore, we considered a number of other useful distinctions which work quite well in a probabilistic classifier but would be problematic to use in a rule-based system. Such features included the percentage of non-alphanumeric characters in the subject of a mail message (junk E-mail, for example, often has subject descriptions such as "$$$$ BIG MONEY $$$$" which contain a high percentage of non-alphanumeric characters). As shown in Figure 2, there are clear differences in the distributions of non-alphanumeric characters in the subjects of legitimate versus junk mes-

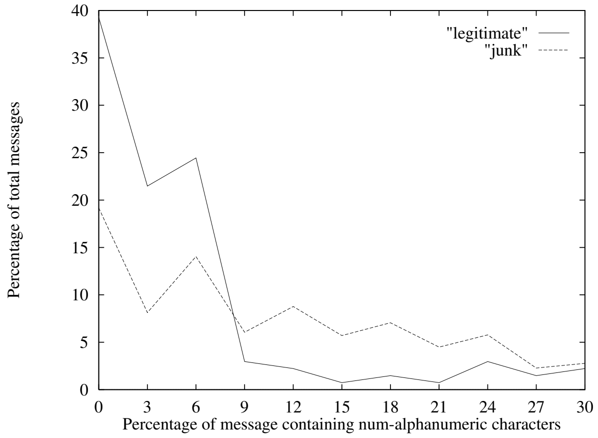

Figure 2: Percentages of legitimate and junk E-mail with subjects comprised of varying degrees of non-alphanumeric characters

sages. But this feature alone (or a discretized variant of it that checks if a message subject contains more than, say, 5% non-alphanumeric characters) could not be used to make a simple yes/no distinction for junk reliably. This is likewise true for many of the other domain-specific features we consider as well. Rather, we can use such features as *evidence* in a probabilistic classifier to increase its confidence in a message being classified as junk or not.

In total, we included approximately 20 non-phrasal hand-crafted, domain-specific features into our junk E-mail filter. These features required very little person-effort to create as most of them were generated during a short brainstorming meeting about this particular task.

Results

To validate our approach, we conducted a number of experiments in junk E-mail detection. Our goal here is both to measure the performance of various enhancements to the simple baseline classification based on the raw text of the messages, as well as looking at the efficacy of learning such a junk filter in an "operational" setting.

The feature space for text will tend to be very large (generally on the order of several thousand dimensions). Consequently, we employ feature selection for several reasons. First, such dimensionality reduction helps provide an explicit control on the model variance resulting from estimating many parameters. Moreover, feature selection also helps to attenuate the degree to which the independence assumption is violated by the Naive Bayesian classifier.

We first employ a Zipf's Law-based analysis (Zipf 1949) of the corpus of E-mail messages to eliminate words that appear fewer than three times as having little resolving power between messages. Next, we compute the mutual information $MI(X_i; C)$ between each feature X_i and the class C (Cover & Thomas 1991), given by

$$MI(X_i; C) = \sum_{X_i = x_i, C = c} P(X_i, C) \log \frac{P(X_i, C)}{P(X_i)P(C)}. \tag{3}$$

We select the 500 features for which this value is greatest as the feature set from which to build a classifier. While we did not conduct a rigorous suite of experiments to arrive at 500 as the optimal number of features to use, initial experiments showed that this value provided reliable results.

Note that the initial feature set that we select from can include both word-based as well as hand-crafted phrasal and other domain-specific features. Previous work in feature selection (Koller & Sahami 1996) (Yang & Pedersen 1997) has indicated that such information theoretic approaches are quite effective for text classification problems.

Using Domain-Specific Features

In our first set of experiments, we seek to determine the efficacy of using features that are hand-crafted specifically for the problem of junk E-mail detection. Here, we use a corpus of 1789 actual E-mail messages of which 1578 messages are pre-classified as "junk" and 211 messages are pre-classified as "legitimate." Note that the proportion of junk to legitimate mail in this corpus makes it more likely that legitimate mail will be classified as junk. Since such an error is far worse than marking a piece of junk mail as being legitimate, we believe that this class disparity creates a more challenging classification problem. This data is then split temporally (all the testing messages arrived after the training messages) into a training set of 1538 messages and a testing set of 251 messages.

We first consider using just the word-based tokens in the subject and body of each E-mail message as the feature set. We then augment these features with approximately 35 hand-crafted phrasal features constructed for this task. Finally, we further enhance the feature set with 20 non-textual domain-specific features for junk E-mail detection (several of which are explicitly described above). Using the training data in conjunction with each such feature set, we perform feature selection and then build a Naive Bayesian classifier that is then used to classify the testing data as junk or legitimate.

Recalling that the cost for misclassifying a legitimate E-mail as junk far outweighs the cost of marking

Feature Regime	Junk		Legitimate	
	Precision	Recall	Precision	Recall
Words only	97.1%	94.3%	87.7%	93.4%
Words + Phrases	97.6%	94.3%	87.8%	94.7%
Words + Phrases + Domain-Specific	100.0%	98.3%	96.2%	100.0%

Table 1: Classification results using various feature sets.

a piece of junk as legitimate, we appeal to the decision theoretic notion of *cost sensitive* classification. To this end, a message is only classified as junk if the probability that it would be placed in the junk class is greater than 99.9%. Although we do not believe that the Naive Bayesian classifier (due to its independence assumption) provides a very accurate probability estimate for classification, a close examination of the values it gives reveal that the 99.9% threshold is still reasonable for this task.

The precision and recall for both junk and legitimate E-mail for each feature regime is given in Table 1. More specifically, *junk precision* is the percentage of messages in the test data classified as junk which truly are. Likewise, *legitimate precision* denotes the percentage of messages in the test data classified as legitimate which truly are. *Junk recall* denotes the proportion of actual junk messages in the test set that are categorized as junk by the classifier, and *legitimate recall* denotes the proportion of actual legitimate messages in the test set that are categorized as legitimate. Clearly, junk precision is of greatest concern to most users (as they would not want their legitimate mail discarded as junk) and this is reflected in the asymmetric notion of cost used for classification. As can be seen in Table 1, while phrasal information does improve performance slightly, the incorporation of even a little domain knowledge for this task greatly improves the resulting classifications.

Figure 3 gives the junk mail Precision/Recall curves using the various feature sets. The figure focuses on the range from 0.85 to 1.0 to more clearly show the greatest variation in these curves. We clearly find that the incorporation of additonal features, especially non-textual domain-specific information, gives consistently superior results to just considering the words in the messages. We believe that this provides evidence that for some targeted text classification problems there is a good deal of room for improvement by considering simple salient features of the domain in addition to the raw text which is available. Examples of such features for more general text categorization problems can include information relating to document authors, author affiliations, publishers, etc.

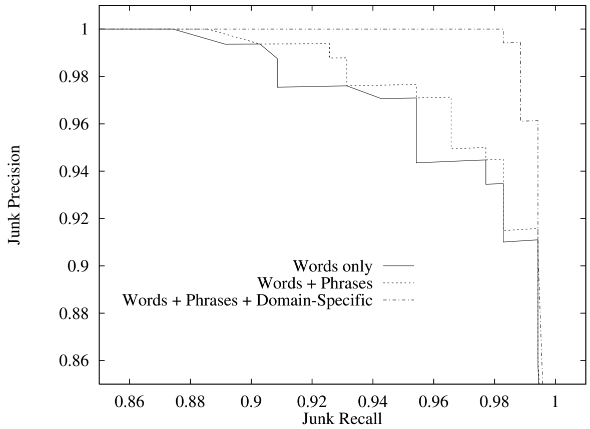

Figure 3: Precision/Recall curves for junk mail using various feature sets.

Sub-classes of Junk E-Mail

In considering the types of E-mail commonly considered junk, there seem to be two dominant groupings. The first is messages related to pornographic Web sites. The second concerns mostly "get-rich-quick" money making opportunities. Since these two groups are somewhat disparate, we consider the possibility of creating a junk E-mail filter by casting the junk filtering problem as a three category learning task. Here, the three categories of E-mail are defined as *legitimate*, *pornographic-junk*, and *other-junk*. By distinguishing between the two sub-groups of junk E-mail, our goal is to better capture the characteristics of such junk by allowing for more degrees of freedom in the learned classifier.

For this experiment, we consider a collection of 1183 E-mail messages of which 972 are junk and 211 are legitimate. This collection is split temporally, as before, into a training set of 916 messages and a testing set of 267 messages. To measure the efficacy of identifying sub-groupings of junk E-mail, we label this data in two different ways. In the first trial, each message is simply given one of the two labels *legitimate* or *junk*. In the second trial, each junk message is relabeled as either *pornographic-junk* or *other-junk*, thus creating a three-way classification problem.

Categories	Junk		Legitimate	
	Precision	Recall	Precision	Recall
Legitimate and Junk	98.9%	94.2%	87.1%	97.4%
Legitimate, Porn-Junk and Other-Junk	95.5%	77.0%	61.1%	90.8%

Table 2: Classification results considering sub-groups of junk E-mail.

Considering the results of our previous experiments on domain-specific features, we include both phrasal and domain-specific features in the feature sets for the present experiments. As before, we apply feature selection to the initial feature set to produce 500 features which are then used to learn a Naive Bayesian classifier. We again use the 99.9% certainty threshold for classifying test messages as junk to reflect the asymmetric cost of errors in this task.

Note that since our true goal is only to filter junk from legitimate E-mail, and not really to identify subgroups of junk E-mail, we consider any test messages classified as either *pornographic-junk* or *other-junk* to be "junk" E-mail. Thus any "junk" messages given either of these labels in the three-category task is considered correctly classified. We realize that this gives an advantage in terms of evaluation to the three-category task over the two-category task, since, in the three-category task, misclassifications between the two subcategories of junk mail (i.e., pornographic-junk messages being classified as other-junk or vice versa) are not penalized. Nevertheless, this advantage turns out not to help as seen below.

The results of the experiments on sub-groups of junk E-mail are given in Table 2. Here we find, rather surprisingly, that modeling the sub-categories of junk E-mail not only does not improve the results, but actually makes them much worse. This result is also clearly echoed in the the junk mail Precision/Recall curves for this experiment (shown in the range from 0.75 to 1.0) given in Figure 4. The curve of the two-category task dominates that of the three-category task over the entire range of Precision/Recall values. We believe there are two main reasons for these results. The first is that while some features may be very clearly indicative of junk versus legitimate E-mail in the two-category task, these features may not be as powerful (i.e., probabilistically skewed) in the three-category task since they do not distinguish well between the sub-classes of junk. The second, and more compelling, reason is the increase in classification variance that accompanies a model with more degrees of freedom. Since the classifier in the three-category task must fit many more parameters from the data than the classifier in the two-category task, the variance in the estimated

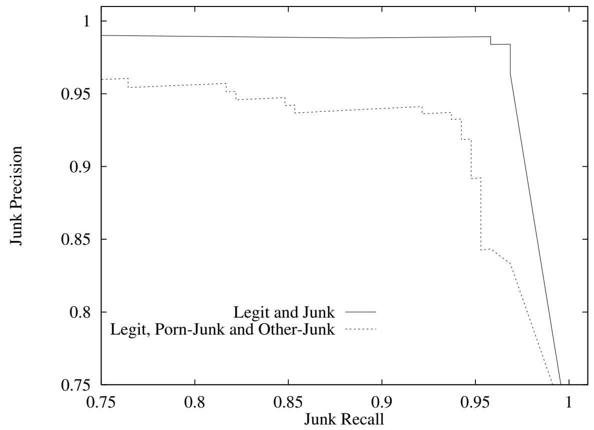

Figure 4: Precision/Recall curves considering subgroups of junk mail.

parameters leads to an overall decrease in the performance of the former classifier. This is especially true given that the parameters for each of the sub-classes of junk are estimated from less data (since the data is sub-divided) than in the two-category task. Such behavior has been seen in other contexts, such as decision tree induction, and is known as the *data fragmentation* problem (Pagallo & Haussler 1990).

Real Usage Scenario

The two test E-mail collections described so far were obtained by classifying existing E-mail folders. The users from which these collections were gathered had already viewed and deleted many legitimate messages by the time the data was sampled. For actual deployment of a junk filter, however, it is important to make sure that the user's *entire* mail stream is classified with high accuracy. Thus, we cannot simply evaluate such a filter using a testing set of legitimate messages that includes only those messages that a user would read *and choose to store* in his or her mail repository. Rather, a junk mail filter must also be able to accurately discern true junk from mail which a user would want to read once and then discard, as the latter should be considered legitimate mail even though it is not permanently stored.

To measure the efficacy of our junk mail filters in

	Classified Junk	Classified Legitimate	Total
Actually Junk	36 (92.0% precision)	9	45
Actually Legitimate	3	174 (95.0% precision)	177
Total	39	183	222

Table 3: Confusion matrix for real usage scenario.

such a real usage scenario, we consider a user's real mail repository of 2593 messages from the previous year which have been classified as either *junk* or *legitimate* as the training set for our filter. As the testing data we use *all* 222 messages that are sent to this user during the week following the period from which the training data was collected. To show the growing magnitude of the junk E-mail problem, these 222 messages contained 45 messages (over 20% of the incoming mail) which were later deemed to be junk by the user.

As before, in this experiment we consider phrasal and domain-specific features of the E-mail as well as the text of the messages when learning a junk filter. Again, we employ a Naive Bayesian classifier with a 99.9% confidence threshold for classifying a message as junk.

The confusion matrix for the results of this experiment is given in Table 3. While the precision results seem promising in this experiment, there is still concern that the three messages classified as junk by the filter which are actually deemed legitimate by the user might be quite important. If this is the case, then such a filter might still not be considered suitable for real world usage. A "post mortem" analysis of these misclassifications, however, reveals that the filter is in fact working quite well. Of the three legitimate messages classified as junk by the filter, one is a message which is actually a junk mail message forwarded to the user in our study. This message begins with the sentence "Check out this spam..." and then contains the full text of a junk E-mail message. The other two misclassified legitimate messages are simply news stories from a E-mail news service that the user subscribes to. These messages happen to be talking about "hype" in the Web search engine industry and are not very important to the user. Hence, there would be no loss of significant information if these messages were classified as junk by the filter. Moreover, we find that the filter is in fact quite successful at eliminating 80% of the incoming junk E-mail from the user's mail stream. For completeness, we also provide the Precision/Recall curve for this task in Figure 5. Based on these results, we believe that such as system would be practical for usage in commercial E-mail applications.

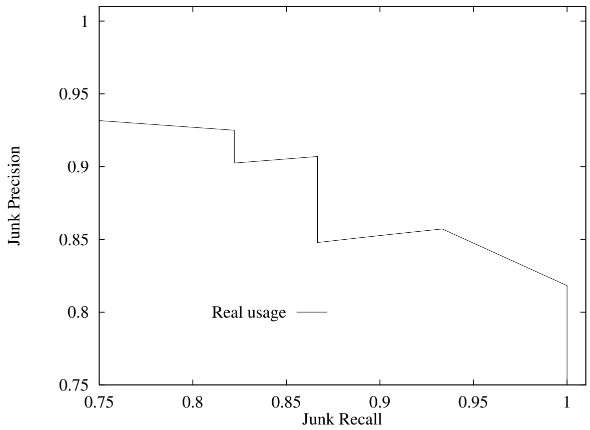

Figure 5: Precision/Recall curve for junk mail in a real usage scenario.

Conclusions

In examining the growing problem of dealing with junk E-mail, we have found that it is possible to automatically learn effective filters to eliminate a large portion of such junk from a user's mail stream. The efficacy of such filters can also be greatly enhanced by considering not only the full text of the E-mail messages to be filtered, but also a set of hand-crafted features which are specific for the task at hand. We believe that the improvement seen from the use of domain-specific features for this particular problem provides strong evidence for the incorporation of more domain knowledge in other text categorization problems. Moreover, by using an extensible classification formalism such as Bayesian networks, it becomes possible to easily and uniformly integrate such domain knowledge into the learning task.

Our experiments also show the need for methods aimed at controlling the variance in parameter estimates for text categorization problems. This result is further corroborated by more extensive experiments showing the efficacy of Support Vector Machines (SVMs) in text domains (Joachims 1997). SVMs are known to provide explicit controls on parameter variance during learning (Vapnik 1995) and hence they seem particularly well suited for text categorization. Thus, we believe that using SVMs in a decision theoretic framework that incorporates asymmetric misclassification costs is a fruitful venue for further research.

In future work, we also seek to consider using Bayesian classifiers that are less restrictive than Naive Bayes. In this way we hope to obtain better classification probability estimates and thus make more accurate costs sensitive classifications. Finally, we are also interested in extending this work to automatically classify messages into a user's hierarchical mail folder structure using the Pachinko Machine classifier (Koller & Sahami 1997). In this way we hope to provide not just a junk mail filter, but an entire message organization system to users.

References

Cohen, W. W. 1996. Learning rules that classify e-mail. In *Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access.*

Cooper, G. F., and Herskovits, E. 1992. A bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9:309–347.

Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory.* Wiley.

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian network classifiers. *Machine Learning* 29:131–163.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press.

Heckerman, D.; Geiger, D.; and Chickering, D. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20:197–243.

Joachims, T. 1997. Text categorization with support vector machines: Learning with many relevant features. Technical Report LS8-Report, Universitaet Dortmund.

Koller, D., and Sahami, M. 1996. Toward optimal feature selection. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 284–292. Morgan Kaufmann.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In *Machine Learning: Proceedings of the Fourteenth International Conference*, 170–178. Morgan Kaufmann.

Lewis, D. D., and Ringuette, M. 1994. Comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR*, 81–93.

McCallum, A.; Rosenfeld, R.; Mitchell, T.; and Ng, A. 1998. Improving text classification by shrinkage in a hierarchy of classes. In *Machine Learning: Proceedings of the Fifteenth International Conference.*

Mitchell, T. M. 1997. *Machine Learning.* McGraw-Hill.

Pagallo, G., and Haussler, D. 1990. Boolean feature discovery in empirical learning. *Machine Learning* 5:71–99.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan-Kaufmann.

Sahami, M. 1996. Learning limited dependence bayesian classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 335–338.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company.

Spertus, E. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, 1058–1065.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* Springer Verlag.

Yang, Y., and Pedersen, J. 1997. Feature selection in statistical learning of text categorization. In *Machine Learning: Proceedings of the Fourteenth International Conference*, 412–420. Morgan Kaufmann.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort.* Addison-Wesley.