Genre-based Decomposition of Email Class Noise

Aleksander Kołcz
Microsoft Live Labs
One Microsoft Way
Redmond, WA
USA

Gordon V. Cormack
Cheriton School of School of Computer Science
University of Waterloo
2502 Davis Centre
Waterloo, Ontario N2L 3G1
Canada

ABSTRACT

Corruption of data by class-label noise is an important practical concern impacting many classification problems. Studies of data cleaning techniques often assume a uniform label noise model, however, which is seldom realized in practice. Relatively little is understood, as to how the natural label noise distribution can be measured or simulated. Using email spam-filtering data, we demonstrate that class noise can have substantial content specific bias. We also demonstrate that noise detection techniques based on classifier confidence tend to identify instances that human assessors are likely to label in error. We show that genre modeling can be very informative in identifying potential areas of mislabeling. Moreover, we are able to show that genre decomposition can also be used to substantially improve spam filtering accuracy, with our results outperforming the best published figures for the trec05-p1 and ceas-2008 benchmark collections.

1. INTRODUCTION

Contamination of training/test data by class-label noise is an important practical concern that impacts the applicability of machine learning techniques to classification problems. With significant levels of such noise during model creation, learners that try to fit the data too closely are in danger of overfitting, which would result in poor test-time performance. When class noise is present at test time, there is a potential for wrongly interpreting the results and misestimating the classification accuracy.

The presence of class noise is particularly problematic when high accuracy is required. In the email spam filtering domain it was confirmed that the presence of class label noise significantly decreases the expected performance of spam filters, particularly those that perform the best on noise free data [18]. Moreover, it was found that naturally occurring class noise is non-uniform, and the bias in the noise distribution poses a more difficult problem for the classifiers than noise distributed uniformly. When labeled data come from a large number of users, another type of a noise-related issue may crop up. While the personal spam/ham assignment by each user may be correct, the same kind of a message may be labeled as ham by some users while spam by others, which is also known as the "graymail effect" [21]. More generally, various labeling inconsistencies may plague the evaluation data, which makes it difficult to arrive at a gold standard.

While it is generally accepted that class-label noise makes the creation of accurate models more difficult, it is not fully understood why natural distribution of such noise tends to pose a more difficult challenge than a distribution that is random. In particular, in the spam filtering context, a recent spam filtering competition[1] revealed that a realistic distribution of class noise may reduce some solutions to random guessing. In this work, we seek to quantify why natural distribution of class noise is more problematic and gain insight into how classifier design can utilize such information in order to maintain acceptable performance. The paper is organized as follows: In Section 2 the problem of class noise in the spam filtering domain is presented in more detail. Section 3 discusses related work. In Section 4 we outline the class-noise measurement framework and contrast the methodologies relying on human assessors vs. those relying on automatic techniques. Section 5 proposes genre based decomposition as a vehicle for better understanding of class noise distributions, also with application to improving the quality of filtering by using genre-based features. Section 6 presents the experimental framework, with the results discussed in Section 7. The paper is concluded in Section 8.

2. CLASS NOISE AND LABEL AMBIGUITY IN EMAIL DATA

Email spam has become a major nuisance for Internet users and a great amount of effort has been spent on trying to eradicate it. Many different techniques have been applied to the filtering problem and although some of them are quite effective, the majority of SMTP email traffic remains spam and individual users still continue to receive spam in their inboxes. The spam filtering problem can be addressed from the machine learning and data mining perspectives. Email messages represent semi-structured text documents, which in addition to content, also contain information about the sender and the recipient as well as some routing data about the path that the message took on its way to the end user. Spam email is defined as an Unsolicited Commercial Mes-

[1] http://www.ceas.cc/2008/challenge/

sage (UCE) and so for each message there is underlying truth signifying whether or not the message was solicited and commercial. Spam filtering can therefore be addressed as a supervised learning problem [17], where labeled data is used to extract patterns and learn classification models that are then applied to unlabeled messages in order to separate spam from legitimate emails, which are known as "ham". What makes the spam filtering problem particularly challenging is the dynamic nature of email, where content tends to naturally change over time, often quite unpredictably, and where learning is inherently adversarial [10], with spammers continuously working on bypassing existing filtering defences. Recent years have seen a growing interest in the spam filtering problem from the research community and the body of literature devoted to this problem has grown to be quite large (see [7] for a comprehensive review).

In the spam filtering domain, class noise poses different challenges depending on whether one treats the problem from a personal filtering or community filtering perspective. In the personal spam filtering context, the labels of the training data are provided by the target user, who ultimately should always be able to determine if a piece of mail is spam or not. Nevertheless, people make mistakes, which may be due to a variety of reasons, e.g., lack of attention, confusing user interface, ambiguity of content (e.g., in the case of emails from a commercial entity, a user may treat some of them as spam and some not). Also, the definition of spam depends on the "solicitation" by the target recipient. However, due to the nature of business relationships a user may in fact solicit mail from various entities without being aware of it. Consequently, many user-provided labels reflect the wanted/unwanted nature of each email, rather than the desired solicited/unsolicited.

When filtering for a community, the problem is compounded by the fact that spam definitions are personal. Many commercial emails are sent out in campaigns, where highly similar content is sent to multiple recipients. These messages may have been solicited by only a fraction of the recipients or else may be wanted by only a fraction of them. In either case, one cannot perform a campaign-wide decision that would satisfy all recipients, which makes the filtering problem ambiguous.

A variant of the community effect occurs when a group of assessors is asked to label messages that were in fact sent to different recipients. In such situations (common in dataset preparation), it is sometimes very difficult to tell if a message may have been solicited or wanted by the recipient, and in effect the assessors' labels may be different from one that the target user would have assigned to the same message.

3. RELATED WORK

Research in the machine learning community has been focusing on detecting noisy training cases and either removing them from the training set or trying to assign them their correct label [5][14]. The relationship between the amount of label noise present and the deterioration of classifier performance has also been investigated.

The problem of the impact of data noise on the effectiveness of machine learning and data mining procedures has been studied from the class noise [5] and attribute noise perspectives. Attribute noise often arises from pure measurement errors and data corruption, which may for example result in missing attributes, misspelling of text, etc. The problem of class noise tends to be more challenging, since mislabelings are hard to distinguish from naturally occurring outliers, which represent legitimate rare manifestations of the target class. Detection of mislabeled instances depends on the assumption that such instances tend to be classified with lower confidence. Thus when many different classifiers are built [5][20], some of them are likely to disagree about the label of a mislabeled instance. The diversity of opinion in the ensemble can be increased when each classifier sees only a portion of the data, which is quite natural in distributed data environments [22]. In the case of a single model on the other hand, a mislabeled instance is likely to be classified as belonging to a class different from the one stated, or even if it is classified to the class stated, the classification confidence (e.g., measured by the instance's distance from the decision boundary) is likely to be low.

Given that noisy instance detection is by itself imperfect, a question arises as to the optimal use of the information returned by such a process. Approaches studied in the literature include instance removal [5], correction [14][18] and weighting [16]. The results reported indicate that removing the potentially mislabeled instances tends to be more beneficial than trying to correct them. Also, more recently it has been demonstrated that weighting instances according to their mislabeling confidence tends to outperform instance removal [16].

One of the problems associated with studying the class noise problem is the difficulty of obtaining the ground truth, especially for large datasets. Typically, researchers assume that the original data is noise free (which is not necessarily the case) and apply an artificial noise model, e.g., by changing class labels for a fraction of instances uniformly at random. Such a process, however, is unlikely reflect the reality where human assessors make labeling mistakes for instances that are particularly confusing for humans. This has been recognized, for example, in [5] where for multilabel datasets the authors identify pairs of classes that are prone to be confused during labeling. This represents a process closer to capturing naturally occurring class noise and results indicate that rule based noise poses indeed a tougher problem when compared to the uniform model. The procedure used in [5] requires domain knowledge and is not applicable to two-class problems. It also assumes that for pairs of classes where confusion is possible, all instances are equally likely to be confused.

4. MEASURING THE CLASS NOISE LEVEL: ASSESSORS VS. CLASSIFIERS

Some of the most effective ways of detecting class noise in a data set are based on classifier ensembles [5][20]. The original training data are used to induce a number of different classifiers, which are then run on the data to be cleansed and, for each instance, the amount of classifier disagreement is measured. Instances where this disagreement is high are considered as potential labeling errors. There are some challenges in measuring classifier disagreement this way, since the base accuracy of individual classifiers may be differ-

ent and thus equally weighting their contributions is not always appropriate. Alternatively, if only a single, but well-calibrated, classifier is present, instances which the classifier determines to have a different class label with high confidence can be considered as potentially mislabeled. In both approaches there is a risk that instances that are not mislabeled but hard to classify might be identified as noise.

For a dataset $D = \{(x_i, y_i)\}_{i=1}^{m} \subset \mathcal{X} \times \mathcal{Y}$, existing methods rely on assigning to each input pair (x_i, y_i) a reliability measure r_i that quantifies (not necessarily in a calibrated way) the probability estimate that the true label x_i of is indeed y_i i.e.,

$$r_i \sim p\left(l_i = y_i | x_i; D\right)$$

where $l_i \in \mathcal{Y}$ is the true label of x_i. When the objective is to clean the training data, instances for which $r_i < T$ are removed from the dataset, are "corrected" to their estimated true label or are weighted according to some function of r_i. There are reasons to believe that at least some instances identified in this way as label noise are in fact outliers or instances that are correctly labeled but ambiguous. This is because, "correcting" the label to the one predicted by a classifier ensemble (or a single classifier) tends to lead to lower overall accuracy than simply removing the problematic instances from the training pool [18]. We are more interested in estimating the noise level in D, which can be expressed as

$$noise\left(D\right) = \frac{\sum_{i=1}^{m} |r_i < T|}{|D|}$$

where T is a user-specified threshold that affects the accuracy of the estimate.

It should be noted that committee or confidence-based class noise detection reflects the protocols employed in data preparation with human labelers. For example, for datasets used by the Text Retrieval Conference (TREC[2]) under the patronage of NIST, oftentimes each data instance is evaluated by several assessors and instances for which there is disagreement may be adjudicated by another expert. Also, in many labeling interfaces found in practice, one has the option of associating some form of confidence with the labeling decision so as to allow the target system to identify potential problem cases.

People can make labeling mistakes for a variety of reasons, not necessarily related to the difficulty of the task at hand (e.g., lack of concentration, carelessness, etc.). It is therefore unclear if instances identified as "noisy" by humans would overlap with instances deemed as noisy according to an automatic data cleaning technique. In this work we attempt to quantity the agreement between these two methodologies using email data.

5. QUANTIFYING THE NATURAL CLASS NOISE: GENRES AS LABEL RELIABILITY INDICATORS

While assessing the overall (or per class) noise level is useful, it inherently assumes that mislabeling is more or less uniform. I.e., if we were to simulate the same amount of noise for a noise free dataset, in the absence of further information we would resort to altering the labels for a fraction of randomly chosen instances. This type of uniformity does not have to be the case, however, and is likely to be far from a realistic scenario. It is reasonable to expect that labeling errors would be concentrated near the optimum boundary separating the classes in the input space. Thus, if this boundary were known (and of course it is not in practice), one might generate artificial class noise by varying the probability of label errors inversely proportionally to the distance between an instance and the optimum decision boundary. One can argue, however, that such a methodology, while general, does not shed much light on why the class noise occurs in a particular domain and its application is thwarted by the uncertainty of where the optimum decision boundary is located, which after all is the problem we are trying to solve in the first place.

Table 1: Email genres: note that many can contain both ham and spam.

category	description
advertising	commercial offers
listserv	discussion forums
newsletter	periodic news/bulletin updates
personal	person to person communications
scam	phishing, 409/Nigerian scams
transactional	order confirmations, notifications, alerts

In the email domain one can expect that certain types of messages (e.g., commercial advertising and phishing) will be harder to label than others (e.g., personal correspondence), especially if label assignment is performed by somebody other than the original recipient. More generally, we can expect that for some domains the input data can be subdivided into a number of subregions sharing common characteristics (e.g., similarity of content). We hypothesize that different regions may exhibit different amounts of class-noise, which provides a domain specific quantification of the labeling noise bias. A key question is how to choose the regions in which the amount of class noise is measured. For text-classification problems one possible way of accomplishing this is to project the original instances onto a fixed content-based taxonomy (e.g., DMOZ[3]). One has to be careful when choosing a taxonomy, however, since with a large number of potential categories some may receive too few instances for effective estimation to be possible. Alternatively, the regions could be identified via clusters naturally occurring in the data, although the results of clustering are not always easy to interpret. In this work, we propose to define the regions via a number of email specific content categories, shown in Table 1, which are believed to capture the most common usage patters. Similar categories could be defined for other problem domains. It is important to note that genres can span the class boundary. For example, both spam and ham can contain messages advertising products. This makes the genre decomposition different from past efforts, where each class was subdivided into categories specific to that class, so as to facilitate cost-sensitive learning [13].

Once the distribution of class noise is known it is possible

[2]http://trec.nist.gov

[3]http://www.dmoz.org

to simulate the impact of natural mislabeling errors over a noise free dataset. An obvious question, however, is whether the distribution of class noise estimated using one particular dataset reflects global mislabeling tendencies or if it is in fact specific to users associated with these data. In this work we consider this question by using data from two different data sources.

5.1 Incorporating genre-based reliably indicators into model induction

There is certain similarity between characterizing class-noise by region and using content based reliability indicators when fusing the outputs of several classifiers [2] . In the first case, the knowledge of a region membership for an instance provides a level of information to the learner about the extent to which the label assignment can be trusted. In the latter case, the knowledge of a region membership allows one to rank the classifiers according to their expected reliability (e.g., to pick the most promising one) or even to use the reliability information directly in merging the classifier outputs. Given that classifiers tend to find noisy data quite challenging, it is interesting to consider to what an extent region-membership information might provide a useful feature to a learning algorithm, which would allow it to distinguish between labels that are potentially noisy and the ones that are likely to be solid. We therefore consider a transformation of the original training set into

$$D = \{(x_i, y_i, R_i)\}_{i=1}^{m} \subset \mathcal{X} \times \mathcal{Y} \times \mathcal{R}$$

where $\mathcal{R}$ captures the information associated with the region membership. This could consist simply of the region label (e.g., one of the symbolic names in the first column of Table 1). However, $\mathcal{R}$ could also contain the estimate class-noise estimate for each label and/or the confidence associated with assigning x_i to R_i. Note that in the extreme case, each instance can be considered to represent a separate region and if then $R_i = r_i$ (i.e., R_i is the instance-level label reliability indicator) one arrives at the instance-based weighting approach to class noise mitigation proposed in [16].

6. EXPERIMENTAL SETUP

Here we analyze two datasets exhibiting different aspects of community based noise. The trec05p-1 dataset was used in the 2005 TREC Spam Filtering Competition. It contains messages from the Enron public corpus, as well as additional ham and spam emails collected from private users. Where possible the messages were assigned labels by the original recipients and were otherwise carefully adjudicated by two expert assessors to provide reference labels. While the resulting gold-standard label set is not class-noise free, it is believed that the noise level is low (i.e., around 0.5%). In the trec05p-1 dataset, in addition to labels provided by target users each message was labeled by a variable number of volunteer human assessors who were not the target recipients of that message. This large scale labeling effort (known as the SpamOrHam project) provides a unique view into human judgement disagreements pertaining to email spam. In particular, it is possible to measure inter-assessor disagreement, as well as the disagreement between the assessors and the gold standard. The number of assessor labels per message is quite varied, with some messages receiving no judgements and some receiving multiple ones, with the median being around 4.

The ceas-08 dataset corresponds to a sample of messages collected from a large community of users of a major ISP (the pool of users was larger and more diverse than in the case of trec05p-1), where each labeling decision was performed by the target recipient. The data was used as the private corpus in the Spam Filtering Competition[4] held in conjunction with the 2008 Conference on Email and Anti-Spam (CEAS-08[5]). Collectively, this dataset captures the noise related to emails of similar content being considered to be spam or ham by different users. The class noise level for this dataset is believed to be high and, in fact, we expect that poor performance of some the classifier entries in the CEAS 2008 competition, which used this dataset, was due the significant presence of class noise.

For the trec05p-1 dataset, it is possible to study the effects of class noise by randomly flipping the labels provided by the gold standard (this has been done in [18]). It is also possible to compare the effect of uniform flipping to that of using natural noise introduced by human assessors. This type of study was performed in [18], determining that natural noise tends to be more difficult than a random one, but it was not quantified what is the nature of this natural noise. One could conjecture that it should correspond to messages that are rather ambiguous (i.e., closer to the decision boundary), but one could also expect that certain types of content are more ambiguous than others. One possible way of assessing such a relationship is to project each email message onto a fixed taxonomy. Given the knowledge of such a projection for all messages in a corpus, together with the gold standard and assessors-provided labels, one can then measure the prevalence of class-noise on a per-category basis. A question arises if the natural noise represents the most difficult type of noise possible (from the standpoint of a classifier), or if in fact a more difficult setup could be accomplished by altering the distribution of class noise among different categories.

Another question is whether the natural distribution of class noise is consistent with the one that can be derived from the classifier disagreement achieved when training a number of classifiers over the gold standard. I.e., to the extent that classifier disagreement detects outliers as well as class noise, to what a degree outliers detected by an automated method reflect human disagreements.

6.1 Online Filtering Methodology

In all experiments we followed the on-line procedure, whereby the messages in each datasets were ordered according to their arrival time, and when evaluating any particular message only earlier messages could be used to provide model information. This type of an experimental setup is most realistic for email filtering, since it acknowledges the strong dependence of both ham and spam on the time axis [9].

6.2 Hypotheses and Methods

Our overall objective is to model the distribution of labeling errors, for the purpose of achieving better understanding

[4]http://www.ceas.cc/2008/challenge/
[5]http://www.ceas.cc/2008/

their nature, better estimates of filter effectiveness and, ultimately, better spam filters. We consider the effect of three factors of labeling error: self-reported classifier confidence, agreement within a committee of classifiers, and the genre of the labeled message. We determine the influence of these factors through their ability to predict disagreement among human adjudicators and through their overall contribution to filter effectiveness. To this end, we conducted to test four specific hypotheses.

E1: Does classifier agreement predict label noise? To test this hypothesis, we used a tribunal of classifiers to predict agreement within a tribunal of human assessors. For each tribunal unanimous agreement was considered the positive result, regardless of whether the tribunal consensus was spam or ham (i.e. non-spam). Better than chance concordance between the classifier and human tribunals may be interpreted as supporting the hypothesis.

E2: Does classifier confidence predict label noise? We measured the ability of separate individual classifiers to predict human tribunal agreement. Classifiers were selected which render their result by computing a score s and comparing it to a threshold t. Our hypothesis predicts that $|s - t|$ is a positive predictor of tribunal agreement. The predictive ability is measured using the area under the receiver operating characteristic curve (AUC) [3].

E3: Does message genre predict label noise? To test this hypothesis, we use active learning to construct a classifier to partition messages into the following genres: personal, advertising, scam, news, mailing list, e-transaction, and other (see Table 1 for a detailed description). Within each genre, the label noise level is computed from the rate of unanimity within the human tribunal. Our hypothesis predicts that noise will differ substantially among genres.

E4: Does partitioning by genre improve classifier performance? If the noise levels among genres are different, one might expect a classifier trained using examples exclusively from a low-noise genre to perform better, at least for messages belonging to the genre. Therefore, deferring to this classifier for such messages may be expected the overall performance of a classifier trained on all messages. Absent prior knowledge of the noise in particular genres, one may predict that an ensemble of classifiers – one specific to each genre and one for all messages – would improve upon a the single classifier. We tested this hypothesis by using our classifier from E3 to partition the messages, inducing separate instances of the same classifier for each the genres as well as the set of all genres, and combining the results using logistic regression. The noisy labels were used to train the ensemble members, and also for evaluation. The labels played no role in partitioning by genre.

E5: Is population-specific genre adjudication necessary to improve classifier performance? To test this hypothesis, we build a genre classifier and the spam filter from entirely separate populations of messages. The genre classifier is induced using active learning on population A; this classifier is used to route the messages of B to a particular member of an ensemble spam filter. The hypothesis predicts that the ensemble filter will better classify the messages of B.

6.3 Data and filter selection

Messages from the TREC 2005 Public Spam Corpus trec05p-1[6] were used for our experiments E1 through E4. The gold standard labels associated with the corpus were used only to evaluate the results of E4; they played no role in training the classifiers or in estimating noise. For these purposes we used labels rendered by participants in the SpamOrHam Internet labeling effort [12]. The TREC corpus contains 92,189 messages, 39,399 labeled ham and 52,790 labeled spam. SpamOrHam comprises 342,771 for messages selected at random from the corpus, with replacement, an average of 3.7 per message in trec05p-1. The messages form a chronological sequence that is presented to the spam filter for online classification, following the TREC methodology [8].

From trec05p-1, we selected only those messages for which there was at least one SpamOrHam label, and from those labels, selected one at random for use as a training label. A total of For estimating noise in E1, E2 and E3, we selected those messages for which there were three or more SpamOrHam labels, and from those labels, we selected three at random (not necessarily including the training label) to form the human tribunal. Because the SpamOrHam acquired labels for messages selected at random, the training messages form an independent identically distributed (i.i.d.) sample of the TREC messages, and the estimation labels are furthermore an i.i.d. sample of the training messages. For evaluating overall spam filtering effectiveness in E4 we use the TREC gold standard labels. To construct the genre classifier in E3, the authors adjudicated a total $3,239$ messages for membership in each of the six genres.

The CEAS 2008 [1] corpus was used exclusively for E5. No statistics, labels, or messages from the corpus were used to tune the genre classifier or the ensemble members. The CEAS corpus was collected from messages delivered to clients of a large service provider. The labels were rendered by the clients themselves, in response to adjudication requests presented by the user interface for randomly selected messages.

The corpus CEAS contains 198,574 messages, of which 89,451 are labeled ham, and 109,123 spam. There is exactly one label per message, so it is not possible to measure label noise directly. Filter evaluation results suggest that the overall noise level is comparable to that of the SpamOrHam labels – about 6%.

The classifier tribunal consisted of three spam filters known to exhibit state-of-the art performance: DMC [4], ROSVM [19] and Bogofilter [15]. logistic our implementation [6] of on-line gradient descent logistic regression [11], itself a state-of-the-art spam filtering method, was used for the genre categorization, for the genre-specific ensemble members, as well as for the overall meta-classifier. All classifiers performed either a very simple extraction of word token features from

[6] trec.nist.gov/data/spam.html

each message, or relied on character n-grams as an even simpler document representation.

7. RESULTS AND DISCUSSION

E1: Agreement between classifier and human tribunals. DMC, ROSVM and Bogofilter were run on-line and the results captured using the TREC spam filter evaluation toolkit. The first 1,200 results were discarded to avoid undue influence of the filters' learning curves. The results were joined to create, in effect, a new classifier predicting label agreement rather than ham or spam. Every message which DMC, ROSVM and Bogofilter all deemed spam, or all deemed ham, was classified as positive; the others negative (i.e., a consensus regimen). The true class for each message was computed in the same manner using the three SpamOrHam labels: if they all agreed, the true class was positive, otherwise negative. The effectiveness of the classifier unanimity in predicting human unanimity was measured using the area under the receiver operating characteristic curve using the TREC toolkit: $1 - AUC = 42.0\%$ with 95% confidence limits $(41.3\% - 42.7\%)$. H_0 predicts that $1 - AUC \approx 50\%$, which is well outside this interval, so may be rejected with very high confidence. While classifier consensus can therefore be considered as a better-than-random predictor of inter-assessor agreement, it is not a particularly strong one. One possible explanation is that the diversity of the classifiers is not as high as the diversity of people's opinions.

E2: Predicting human agreement from classifier confidence. Logistic was run on-line as described above. The score s returned by logistic for each message m is an estimate of the logarithm of the odds that the message is spam; i.e.

$$s = log\left(\frac{Prob(m \in spam)}{Prob(m \in ham)}\right)$$

A threshold value of $t = 0$ yields the maximum likelihood classifier; $|s|$ yields the probability that the classifier is correct in this instance. We replaced s by $|s|$ and evaluated the outcome using the TREC tools. The resulting AUC scores, along with those for the other three classifiers, are shown in Table 2. We note that all the filters' self-confidence estimates were good predictors of human agreement; for logistic especially so. This is further illustrated in Figure 1, where the density of disagreements as well as the overall messages density are shown as functions of the absolute values of the logistic regression score. Since score of 0 denotes the decision boundary it is clear that messages lying in the natural ambiguity regions for the classifier are also the ones that human assessors tend to disagree with most.

E3: Predicting noise from genre. Under the assumption that three assessors independently have the same error rate e, the probability of unanimity in a tribunal is

$$u = (1 - e)^3 + e^3$$

That is, to agree the assessors must all be right or all be wrong. The "correct" answer is immaterial. Solving

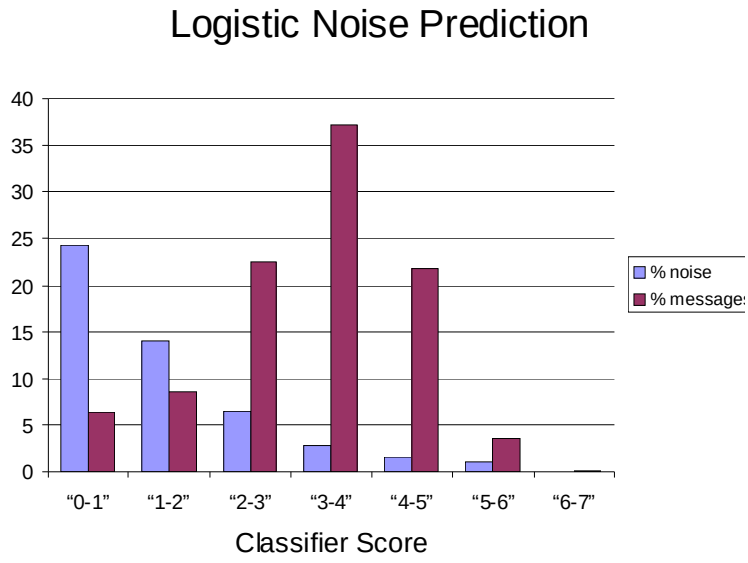

Figure 1: Distribution density of trec05-p1 messages as a function of the absolute value of the score assigned to them by logistic regression. A corresponding distribution of class-noise (measured by human assessor disagreement) is shown for comparison.

Table 2: Individual spam filter effectiveness at predicting unanimity among a tribunal of human assessors.

Classifier	(1-AUC)(%)
	Agreement
logistic	23.0 (22.4 - 23.5)
Bogofilter	36.9 (36.3 - 37.7)
DMC	37.6 (37.1 - 38.2)
ROSVM	29.9 (29.4 - 30.1)

Table 3: Email genres: note that many can contain both ham and spam.

genre	n	$(1-u)(\%)$	$e(\%)$
adver	20,894	8.5	2.9
list	1,163	46.9	19.4
newsl	2,163	46.6	19.2
none	14,180	13.6	4.8
perso	17,939	15.1	5.3
scam	4,989	17.0	6.0
trans	2,652	27.0	10.0
total	63,980	14.9	5.2

Table 4: Number of messages (n), human tribunal disagreement ($1-u$), and noise (e) according to genre. Ham messages only.

genre	n (ham)	$(1-u)(\%)$	$e(\%)$
adver	576	56.1	24.9
list	959	43.2	17.4
newsl	1769	42.5	17.1
none	2960	30.5	11.5
perso	17883	15.0	5.3
scam	69	33.3	12.7
trans	2609	26.7	9.9
total	26825	21.6	7.8

Table 5: Number of messages (n), human tribunal disagreement ($1-u$), and noise (e) according to genre. Spam messages only.

genre	n (spam)	$(1-u)(\%)$	$e(\%)$
adver	20,318	7.2	2.5
list	204	64.7	31.5
newsl	394	65.0	31.7
none	1,1220	9.1	3.1
perso	56	46.4	19.1
scam	4,920	16.7	5.9
trans	43	46.5	19.2
total	37,155	10.1	3.5

for e we have

$$e = 0.5 - \frac{\sqrt{-3+12u}}{6}$$

Given a set of labels, u is easily estimated. Over the set of 63,980 messages used for evaluation, 54,451 have unanimous labels ($u = \frac{54451}{63980} = 85.1\%$), while 9,529 do not ($1-u = 14.9\%$). It follows that $e = 5.2\%$. That is, the SpamOrHam labels have an overall noise level of 5.2%.

We used logistic regression with uncertainty sampling to fetch and label examples of each of our six genres. These examples were used to train six logistic classifiers, each classifying every message as in the genre or not. Each message was labeled with the genre of the most confident classifier, and none if no classifier yielded a positive result. The number of messages in each genre is reported, along with the values of $1-u$ and e, in table 3. We observe that e is strongly predicted by genre, varying by a factor of 6 with advertising having the lowest at 2.9% and mailing lists and newsletters each approaching 20%. Tables 4 and 5 further stratify these results by the classification of each message as ham or spam, according to the TREC label. It is interesting to note that messages of the same genre are characterized by lower labeling noise in the class in which they are more prevalent. In a way, this is to be expected since by the very fact that a genre is more common in one class simplifies the labeler's decision.

E4: Improving classifier performance. First the messages were categorized into genres using logistic regression. We then ran nine instances of the logistic classifier eight times on the TREC message sequence. One run classified all the messages, using the SpamOrHam labels for training. Then seven runs formed a committee of experts (the 7th expert corresponds to the category of none, when no specific genre can be reliably selected), with the particular expert for each message determined by the genre categorization. Finally, the results were combined using logistic regression as a meta-classifier. More specifically, each expert was trained using only those messages that corresponded to its genre. During evaluation, first the genre of the message was selected, then the message was classified by the corresponding expert and then the expert's score was combined with the regular score of a genre-blind classifier. The results were then evaluated using the TREC labels as a gold standard (recall that the TREC labels were never used as input to the classifiers). The TREC summary measures, AUC and LAM (logistic average misclassification rate) are given in table 6. LAM is defined as

$$LAM = \text{logit}^{-1}\left(\frac{\text{logit}\,(fpr) + \text{logit}\,(fnr)}{2}\right)$$

with $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$, where fpr and fnr are the false positve and false negative rates of a classifier, respectively. We see that while the performance of the committee by itself is poor, when combined with the overall classifier a substantial improvement is seen in both scores (this is in fact the best performance reported for this corpus). The 95% confidence intervals do not overlap, so H_0 is rejected with high confidence. Genre information thus proves to be remarkably useful in terms improving the filtering performance. We also noted similar results for other forms of integrating the genre-specific classifiers with the overall classifier. We hypothesize that final meta classifier is able to learn the reliability of stated training-set labels vis-a-vis the genres the messages appear to belong to.

Table 6: TREC corpus AUC and LAM scores for genre-specific committee of experts, single overall classifier, and metaclassifier combining the two.

training genre	$(1-AUC)(\%)$	$LAM(\%)$
7 experts	3.7 (3.6 - 3.9)	7.6 (7.4 - 7.8)
overall	0.14 (0.13 - 0.15)	2.1 (2.0 - 2.2)
meta	0.097 (0.09 - 0.11)	1.7 (1.6 - 1.7)

Table 7: CEAS corpus AUC and LAM scores for genre-specific committee of experts, single overall classifier, and metaclassifier combining the two.

training genre	$(1-AUC)(\%)$	$LAM(\%)$
overall	5.5 (5.4 - 5.6)	6.2 (6.1 - 6.3)
meta	5.3 (5.2 - 5.4)	5.5 (5.4 - 5.7)

E5: Transfer learning to another corpus. We repeated E4 on the CEAS messages, using the TREC-trained genre

Table 8: Effect of CEAS-specific genre training.

training genre	$(1 - AUC)(\%)$	$LAM(\%)$
overall	5.5 (5.4 - 5.6)	6.2 (6.1 - 6.3)
meta	5.3 (5.2 - 5.4)	5.4 (5.3 - 5.5)

categorizer. No CEAS messages or corpus statistics were involved in its construction or tuning. The logistic classifier was used exactly as in E4. Since there are no gold standard labels for CEAS, we used the same noisy labels for evaluation. Table 7 presents the results from the overall logistic classifier and the meta classifier (results for the committee are not available for presentation). The metaclassifier improves on the single classifier, supporting the hypothesis. The confidence intervals do not overlap so H_0 is rejected. These results are superior to any previously reported for this corpus.

After conducting E5 we conducted a sequel (E5a) to determine whether or not training the genre classifier on population-specific examples would improve performance. About 100 CEAS messages per genre were adjudicated using the same active learning technique in E3. These messages were added to the training examples used to induce the genre categorizer. The result, detailed in table 8 showed no measurable improvement. This would suggest that email genre definitions are rather stable and it is encouraging from the perspective of applying this method in practice, since measurable improvements in performance can be gained without re-launching an expensive corpus-specific labeling effort.

8. CONCLUSIONS

We investigated the nature of class noise in the spam filtering domain using two of the largest available datasets. Our results indicate that label noise has a clear content-based bias, with certain genres of email being much more likely to confuse than others. We hypothesize that similar patterns could be found in other classification problems involving text. Experiments demonstrate that email messages that automatic classifiers, such as logistic regression, find confusing correspond quite closely to messages that human assessors are likely to find confusing. Thus classifier confidence based data-cleaninig methods can be thought of as a good substitute for the more expensive approach of having each message reviewed by a number of human judges.

We proposed a method of quantifying the content bias in the distribution of class noise based on genres. While the particular genres used in this study are email specific, they could easily be redefined for any particular domain. Our results show than for genres spanning the class boundary, label noise is more of a problem for the class in which the genre is less likely to be present. Interestingly, incorporating genre membership indicators into the classifier learning process also leads to significant performance improvements, with the results obtained in this work outperforming previously reported results for the two datasets in question. Moreover, the improvements appear to be stable even if the genre definition is transferred across collections. Further research into the effective use of genre information in classifier design will be the subject of future work.

9. REFERENCES

[1] The CEAS 2008 live spam challenge. http://www.ceas.cc/2008/challenge/challenge.html, 2007.

[2] P. N. Bennett, S. T. Dumais, and E. Horvitz. Probabilistic combination of text classifiers using reliability indicators: models and results. In SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 207–214. ACM Press, 2002.

[3] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7):1145–1159, 1997.

[4] A. Bratko, G. V. Cormack, B. Filipič, T. R. Lynam, and B. Zupan. Spam filtering using statistical data compression models. Journal of Machine Learning Research, 7(Dec):2673–2698, 2006.

[5] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. JAIR, 11:131–167, 1999.

[6] G. V. Cormack. University of waterloo participation in the trec 2007 spam track. In Sixteenth Text REtrieval Conference (TREC-2007), Gaithersburg, MD, 2007. NIST.

[7] G. V. Cormack. Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval, 1(4):335–455, 2008.

[8] G. V. Cormack and T. R. Lynam. TREC 2005 Spam Track overview. http://plg.uwaterloo.ca/~gvcormac/trecspamtrack05, 2005.

[9] G. V. Cormack and T. R. Lynam. On-line supervised spam filter evaluation. ACM Transactions on Information Systems, 25(3), 2007.

[10] N. N. Dalvi, P. Domingos, Mausam, S. K. Sanghai, and D. Verma. Adversarial classification. In Proceedings of the Tenhth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), pages 99–108, 2004.

[11] J. Goodman and W. tau Yih. Online discriminative spam filter training. In The Third Conference on Email and Anti-Spam, Mountain View, CA, 2006.

[12] J. Graham-Cumming. SpamOrHam. Virus Bulletin, 2006-06-01.

[13] A. Kołcz and J. Alspector. SVM-based filtering of e-mail spam with content-specific misclassification costs. In Proceedings of the Workshop on Text Mining (TextDM'2001), 2001.

[14] S. Lallich, F. Muhlenbach, and D. A. Zighed. Improving classification by removing or relabeling mislabeled instances. In ISMIS '02: Proceedings of the 13th International Symposium on Foundations of Intelligent Systems, pages 5–15, London, UK, 2002. Springer-Verlag.

[15] E. S. Raymond, D. Relson, M. Andree, and G. Louis. Bogofilter. http://bogofilter.sourceforge.net/, 2004.

[16] U. Rebbapragada and C. E. Brodley. Class noise mitigation through instance weighting. In Proceedings of the 18th European Conference on Machine

Learning, 2007.

[17] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian Approach to Filtering Junk E-Mail. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, 1998.

[18] D. Sculley and G. V. Cormack. Filtering spam in the presence of noisy user feedback. In Proceedings of the 5th Conference on Email and Anti-Spam (CEAS 2008), 2008.

[19] D. Sculley and G. M. Wachman. Relaxed online support vector machines for spam filtering. In 30th ACM SIGIR Conference on Research and Development on Information Retrieval, Amsterdam, 2007.

[20] S. Verbaeten and A. V. Assche. Ensemble methods for noise elimination in classification problems. In Multiple Classifier Systems 2003, pages 317–325. Springer-Verlag, 2003.

[21] W. Yih, R. McCann, and A. Kołcz. Improving spam filtering by detecting gray mail. In Proceedings of the 4th Conference on Email and Anti-Spam (CEAS 2007), 2007.

[22] X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In Proceedings of the Twentieth International Conference on Machine Learning, pages 920–927, 2003.