Soft Labeling for Multi-Pass Document Review

Jianlin CHENG[*], Amanda JONES[*], Caroline PRIVAULT[¬], Jean-Michel RENDERS[¬]

Xerox Research Centre Europe 6 Chemin de Maupertuis, 38240 Meylan, FRANCE	Xerox Litigation Services 485 Lexington Ave New York, NY 10017

* {Firstname.Lastname}@xls.xerox.com
¬ {Firstname.Lastname}@xrce.xerox.com

Abstract

In this paper we examine the use of machine learning classifiers utilized in technology-assisted review (TAR) and, more specifically, the multi-pass manual coding process that supports the training and testing of these classifiers. Manual document coding is known to be subject to error, misinterpretation, and disagreement in reviews conducted for litigation matters. It is also known that the accuracy and consistency of such coding has significant impacts on the performance and evaluation of resulting classifiers, since the classifiers utilize this coding as a basis for "learning by example." Correspondingly, the value of rigorous quality control (QC) for the training and testing documents used for classifier development is also well established. In a traditional approach, coding decisions made during QC review are assumed to be accurate and are used going forward without reference to previous coding. We describe a method for integrating multiple coding judgments in the construction of a document classifier, based on multi-pass review efforts, and explain the benefit of such an approach.

1. Introduction

Text classification is an increasingly popular machine learning application in e-discovery, used primarily to rank order documents for review. Text classifiers are built by having statistical algorithms "learn" from patterns in a set of pre-classified (labeled) training documents provided by a subject matter expert. The statistical model obtained from this learning phase can then be used to generate scores that predict the likely categories of new (unlabeled) documents.

The standard machine learning approach assumes that each training document is either completely relevant to a class or completely irrelevant. In practice, though, this type of strict binary categorization may fail to capture the true nuance of the situation. Further, with a binary labeling scheme, disagreement among reviewers, misinterpretation of relevance, and simple human error can lead to discrepancies in the training that will significantly impair the quality of the classifiers' scoring of unlabeled data. Inter-reviewer labeling conflicts are known issues for text classification in the context of document review for litigation [Grossman and Cormack, 2011a]. As Grossman and Cormack point out, "It is well established that human assessors will disagree in a substantial number of cases as to whether a document is relevant, regardless of the information need or the assessors' expertise and diligence."

In another 2011 paper (published at DESI IV conference), the same authors [Grossman and Cormack, 2011b] provide a qualitative analysis of the cases of disagreement in responsiveness judgments rendered during the construction of "gold standard" labeling for documents utilized in the TREC 2009 Legal Track Interactive task [Hedin et al., 2010]. In this study the authors reexamined each document for which there was a disagreement and made their own determination of whether the document was "clearly responsive," "clearly non-responsive," or "arguable." Arguable was taken to mean that the document could reasonably be interpreted as either responsive or not responsive, given the production request and review guidelines. Grossman and Cormack conclude in their study that “the vast majority of cases of disagreement are a product of human error rather than documents that fall in some ‘gray area’ of responsiveness.” While this specific view is not shared by all practitioners, there is widespread consensus that binary tagging by human reviewers is susceptible to inconsistency and that this inconsistency can lead to inaccurate and unreliable training material for statistical classifiers.

Exploring a related theme, Eli Nelson [Nelson, 2011] discusses what he calls “the false dichotomy of relevance.” According to Nelson, even in straightforward legal matters there are few documents that can be considered 100% responsive or 100% non-responsive to a production request. Therefore, he suggests that the standard binary labeling employed in responsiveness review may not be appropriate, describing it as an “enforced binary classification of relevance.”

When reviewers have at their disposal only Yes/No coding values to express their judgments, the following effects are observed:

- “Safety net” strategies may emerge, whereby reviewers err on the side of positive judgments to avoid missing responsive documents. Consequently, marginally responsive documents are assigned labels that are equivalent to those assigned to strongly responsive documents.
- Inter-reviewer disagreement may be exacerbated. Having no means of conveying ambiguity, uncertainty, or nuance in their judgments, reviewers may develop their own rules and strategies for arriving at binary decisions, and these individual rule systems will inevitably differ from one reviewer to the next, creating systematic conflicts in labeling.

These coding strategies are quite understandable and may even be necessary in the context of a traditional linear review where, ultimately, all documents must simply be produced or not produced according to the label they have been assigned. However, in the context of TAR, these strategies, along with the types of labeling errors and discrepancies discussed earlier, may lead to at least two very negative consequences:

- The classifier may be poorly trained, since it will interpret and treat the text of marginally responsive documents, and even some mislabeled non-responsive documents, as equal in responsiveness to the text of strongly responsive documents. Thus, the classifier will not learn to differentiate accurately between relevant and irrelevant text.
- Evaluation of the classifier’s performance may be skewed in unpredictable ways, and calibration of the system may be impaired, because marginally relevant and mislabeled documents will be included in the gold standard against which the classifier is tested and tuned.

The issues raised by a strict Yes/No coding system pose challenges and create inefficiencies even within a fully manual review. As Nelson [Nelson, 2011] observes, “when forcing documents to live in a yes/no world, a marginal yes will be considered the same as an obvious, smoking-gun yes for all follow-on evaluations.” This means that all yes-coded documents will require the same downstream resource investments, regardless of their actual value to the case or the accuracy of their labeling. For machine-learning approaches to technology-assisted review, where performance measurement plays a central role in model refinement and ultimately determines the course for downstream review, these issues can have an even more significant impact. A model informed by erroneous or conflicting coding will propagate that confusion across an entire review population, resulting in unreliable classifications.

This suggests that a rigorous QC process is vital for ensuring accurate labeling of documents participating in the training and testing of statistical models for TAR. Barnett and co-authors [Barnett et al., 2009] illustrated that even a modest number of corrections of erroneous coding can, in some instances, lead to non-trivial improvements in TAR model performance. Re-reviewing documents for the purpose of QC can be a time-consuming endeavor, however, and it is sometimes perceived by review teams to be little more than a costly burden. Thus, it would be beneficial to have a means of leveraging QC efforts to the greatest possible advantage.

In this paper, we explore the possibility of utilizing multiple coding inputs, obtained via first- and second-pass QC reviews, to realize non-binary relevance statuses for documents. The goal of this approach would be to enhance TAR performance by increasing sensitivity to the more gradient nature of relevance, thereby maximizing the value and potential of the QC process. The paper is organized as follows: in Section 2 we propose a "soft labeling" strategy and discuss the benefits observed when this strategy is applied to several mock Enron matters from previous years' TREC Legal Track exercises. Section 3 details those experiments, and Section 4 presents related research. Finally, Section 5 discusses conclusions.

2. Soft Labeling

We propose utilization of a non-binary labeling scheme for documents that we call "soft labeling." A soft-labeled document will be associated with weights for one or more categories. These weights are based on the multiple coding inputs that may be obtained for each document as a result of typical first-pass manual review being followed by a second-pass QC review.

We focus on the category of responsiveness, but the approach could be applied to any other binary classification scheme.

2.1 Labeling documents with non-binary codes

This section briefly reviews two basic machine learning methods, namely logistic regression and Probabilistic Latent Semantic Analysis (PLSA).

If we denote a training set as a collection of n pairs $\left\langle y^i, d^i \right\rangle$, where y^i = 1 represents "*belonging to class R*" versus y^i = 0 represents "*not belonging to class R*". A logistic regression model formulation (see [Agresti, 2007] and [Balaji et al. 2005]) expresses the probability for d^i to belong to class R as: $p(y=1|d^i) = e^{\alpha_0 + \sum_{j=1}^{M} \alpha j * d_j^i} / (1 + e^{\alpha_0 + \sum_{j=1}^{M} \alpha j * d_j^i})$, where $d^i = (d_1^i, d_2^i, ..., d_M^i)$ is the feature vector encoding a document d^i on M features. The regression coefficients $\alpha = (\alpha_0, \alpha_1, ..., \alpha_M)$ are computed by maximizing the regularized log - likelihood of the observed data expressed as: $ll = \sum_{i=1}^{n} [y^i * (\alpha_0 + \sum_{j=1}^{M} \alpha_j d_j^i) - \log[1 + e^{\left(\alpha_0 + \sum_{j=1}^{M} \alpha_j d_j^i\right)}]]$ + $R(\alpha_1, \ldots, \alpha_M)$, where the last term is the regularization term (typically the L1-norm or the L2-norm of the regression coefficient vector). The y^i values here are not limited to binary values and can, therefore, naturally serve the purpose of expressing nuance in a document's class membership. In other words, one can still use most of the standard logistic regression training algorithms (e.g., the Iteratively Reweighted Least Squares or the Bound Optimization algorithms [Balaji et al., 2005]) without any change.

Similarly, PLSA (see [Hofmann, 1999]) offers a natural way to build a text classifier using soft-labeled samples. PLSA is essentially designed to be a soft clustering/soft classification framework. It attempts to identify the $P(w \mid C)$ values – i.e., word language models for categories or clusters (C) -, such that $P(w \mid \mathrm{d}^i) \cong \sum_c P(w|\mathrm{C}) * P(\mathrm{C}| d^i)$, where $P(w \mid \mathrm{d}^i)$ is the observed frequency of word w in document d^i. In the categorization task, assuming that there are two classes corresponding to y=0 and y=1, the word occurrences in training documents [$P\big(d_j^i \mid d^i\big)$] as well as the class memberships of the training documents' [$P\big(y = 1 \mid d^i\big)$] values are provided as inputs; the class membership can be expressed as the soft label weight associated with each document. Solving the equation typically involves the use of the EM-algorithm in a way that is similar to the folding-in algorithm [Hofmann, 1999]. Note that, in the hard label case ($P\big(y = 1 \mid d^i\big)$ is either 0 or 1), solving this equation is much simpler as it amounts to computing word counts by category, followed by a simple normalization.

This review of two basic machine learning techniques is intended to illustrate that both approaches naturally support soft-labeled inputs.

The more challenging task is determining how soft labeling can be incorporated at the manual coding stage of the modeling process with minimal disruption to review workflow. A very direct approach might be to allow reviewers to freely express the level of uncertainty or ambiguity they wish to associate with a document using a continuous numeric scale ranging from 0-1. For example, reviewers could assign a value of 0.85 to documents they believe to be responsive but with some degree of uncertainty. A value of 1.0 could be assigned to clearly responsive documents, a value of 0.20 could be applied non-responsive documents with a hint of responsiveness, and a value of 0 could be assigned to completely irrelevant documents.

In this way, reviewers would be able to convey much more information regarding their assessment of a document's contents and its value to the review. Unfortunately, defining a responsiveness scale that could be applied consistently and intuitively by human reviewers in a way that would be maximally beneficial from a statistical modeling standpoint would be extremely challenging, if not impossible.

If a responsiveness scale cannot be defined such that it is simple and intuitive for human reviewers to use consistently, then adopting a graded approach to coding would simply reintroduce and exacerbate the problems stemming from the type of inconsistent binary coding discussed above. Thus, instead of approaching the non-binary responsiveness scoring directly, we put forth, in the next section, more practical options for indirectly obtaining graded responsiveness assessments at the quality control stage of a review.

2.2 Soft labels for modeling QC changes

In most technology-assisted document reviews, a certain number of documents are selected for QC, wherein the documents' coding is re-examined and corrected as needed before being included as part of the finalized training sample for the machine learning classifier. Documents that undergo QC can have their current labeling either confirmed or overturned by reviewers. Typically, the document coding assigned during the QC review prevails and is used in the final modeling effort. As an alternative, we propose deriving adjusted labels, based on the results of the QC examination, to reflect any disagreements between the documents' original and final coding. This information can then be incorporated into the model in the form of document soft labels.

The approach can be formulated generally as follows:

- let y_1^i be the 0/1 labeling of document d^i assigned during a first-pass review for class C
- let y_2^i be the 0/1 labeling of sample d^i assigned during a second-pass review for class C.

The new label of document d^i is computed as, $y_3^i = \min\{(\alpha\, y_1^i + \beta\, y_2^i), 1\}$ with $\alpha,\ \beta \in [0; 1]$ and $\alpha + \beta \geq 1$. When $\alpha = 0$, all emphasis is given to the 2nd review judgment and mitigated through the parameter β. When $\beta = 0$, all emphasis is given to the 1st review judgment and mitigated through the parameter α. Note that the "min" operator is only active when both reviewers agree on the positive label. In this case, the new label is 1; there is no real truncation effect.

Documents from the training set that are simply confirmed by the QC team as being correctly coded as responsive or non-responsive will retain their existing binary y_1^i values; documents whose coding is changed by the QC team will receive a y_3^i soft label.

In the cases that we address in the experiment section, the coding history revealed an interesting asymmetry in the reviewers' disagreements. Specifically, we observed that disagreements were much more frequent for documents originally coded as responsive.

For example, consider the two-pass adjudicated TREC 2009 Legal Track Interactive Task review described by Grossman and Cormack [Grossman and Cormack, 2011a]. From the figures those authors collected for each topic, we observe the following:

- 1552 documents of the 5386 documents coded as responsive in the first-pass review were re-coded as non-responsive (28.8%).
- 1100 documents of the 41,913 documents coded as non-responsive in the first-pass review were re-coded as responsive (2.6%).

Overall if we use d(A)=0/1 to denote the judgment of a 1st reviewer A on a document d, and d(B)=0/1 to denote the judgment of a reviewer B on a document d in second-pass review, with 1 meaning responsive and 0 meaning non-responsive, we observe from the same figures that the probability that d(A)≠d(B) is around 5.6% overall, while the probability that d(A)≠ d(B) rises to ~29% when d(A)=1. That is, there is a much higher likelihood of documents originally coded as responsive to be changed to non-responsive than vice versa.

Consequently, we judged that disagreements between reviewers' judgments should be treated somewhat differently depending on whether the QC disagreement involved documents originally labeled responsive as opposed to non-responsive.

In a symmetrical model, we would introduce soft labels upon reviewers' disagreement through a single parameter, regardless of the original coding, as follows: $y_3^i = \min\{(\alpha\, y_1^i + (1-\alpha)\, y_2^i), 1\} = (\alpha\, y_1^i + (1-\alpha)\, y_2^i)$, with $\alpha,\ \beta \in [0;1]$. If α= 1, the soft label model is equal to the model trained from the y_1^i binary codes after first-pass review. With α= 0, the soft label model is equal to the model trained from the y_2^i binary codes from second-pass review. This symmetrical soft label model should be compared to an asymmetrical soft label model where two independent parameters, α and β, are used for tuning the soft label effect differently according to the original coding: $y_3^i = \min\{(\alpha\, y_1^i + \beta\, y_2^i), 1\}$. This general function could be interpreted as follows:

- if the 1st judgment is responsive and the 2nd is non-responsive, then assign a weighted (soft) label of α to the document;
- if the 1st judgment is non-responsive and 2nd is responsive, then assign a weighted (soft) label of β to the document

Experimental evidence (see next section) indicated that choosing α = 0.75 and β = 0.9 resulted in superior performance, regardless of the topic. While this was universally true for all the data sets and topics we examined, we would not assert that using these values as fixed settings would necessarily lead to optimal performance in all circumstances. Instead, customized tuning of the α and β values to fit the given data set would likely lead to optimal performance most often.

2.3 An alternative strategy: late fusion of models

Our soft labeling approach targets training inputs as a means of reflecting non-binary relevance in document review. However, it would also be possible to operate on classifier outputs to achieve this objective. This alternative would be considered a "late fusion" approach and would involve building separate models based on the pre-QC binary coding and the post-QC binary coding and combining the output scores of these models to arrive at final classifications. In other words, this alternative consists of training a first classifier from first-pass binary codes, a second classifier from second-pass binary codes, and combining their outputs by convex combination. We will compare this strategy with the soft labeling strategy in the following section.

3. Detailed Experiments

This section provides a description of the experimental results obtained from testing the soft labeling approach described above against the TREC 2010 Legal Track Learning Task data.

3.1 Simulation of QC review using 2010 TREC Legal Track data

The TREC 2010 Legal Track Learning Task provides a suitable framework for simulation of a two-stage review (see [Cormack et al., 2010]). Thus, we were able to draw upon this as a source of data for testing our proposed soft labeling approach. In the 2010 Learning Task, participants were required to estimate the probability of responsiveness for each document in a large collection, based on a given seed set of documents that had been manually coded for responsiveness.

The full document collection was made up of a variant of the Enron email corpus comprising 685,592 documents. The Learning Task reused the 7 topics, numbered 201 to 207, from the TREC 2009 Interactive Task (see [Hedin et al,. 2010]), plus one novel topic, number 200. For our experiments, all 8 topics were considered. The table below presents the composition of each topic's seed set with respect to its manual responsiveness coding.

Table 1: Seed set composition by topic (after first-pass manual coding)

Topic	Responsive	Not Responsive	Total
200	230	621	851
201	168	523	691
202	1006	403	1409
203	67	892	959
204	59	1132	1191
205	333	1506	1839
206	19	336	355
207	80	511	591

Participants used this seed set coding to train their classifiers. They then submitted estimated responsiveness probabilities for every document in the collection.

For the purposes of soft label testing, we treated the original seed set's manual coding as our first-pass review result and treated the coding obtained from the final results, following adjudication by the Topic Authority, as our second-pass results (see [Cormack et al., 2010] for more details). The final post-adjudication assessments for the 685,592 Enron emails constituted the 2010 gold standard for the TREC Legal Track Learning Task.

Comparing the original seed set coding for each topic to the final gold standard, we were able to identify all of the manual coding changes that resulted from adjudication, and track the direction of these changes as well as their impact on the rate of responsiveness in the seed set (Table 2). This informed the soft labeling used in our experimental models.

Table 2: Responsiveness rates in seed sets and coding changes per topic

Topic	200	201	202	203	204	205	206	207
R/(R+NR) ratio first-pass	24.8%	24.4%	71.3%	6.9%	5%	18%	5.3%	14.9%
R/(R+NR) ratio second-pass	15%	10.6%	24.8%	4.3%	1.6%	8.6%	0.6%	10.8%
First-pass R re-coded NR	36	78	31	28	19	30	16	18
First-pass NR re-coded R	18	1	0	4	1	0	0	0
#changed	54	79	31	32	20	30	16	18

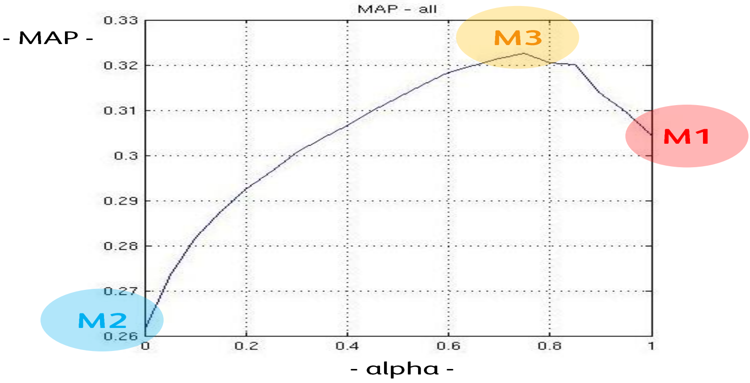
We used the entire 685,592 gold standard population to evaluate our classifiers, but, for each topic, we excluded documents that had been used in the seed set (i.e., as training examples). Thus, the final gold standard test set sizes varied somewhat per topic. We made the simplifying assumption that documents unassessed after the final adjudication could be considered non-responsive; this assumption is partly motivated by the fact that the participants had the opportunity to bring more responsive documents into evidence during the appeal phase.

The classifiers presented in the following sections were built using logistic regression modeling – an approach identical to that used to generate the highest scores achieved in the TREC 2010 Legal Track Learning Task [Cormack et al., 2010]. Similar results were observed for classifiers built using PLSA modeling as well. Thus, these logistic regression classifiers serve as reasonable state-of-the-art baseline models for this experimental context.

3.2 Soft labeling symmetric model compared to binary standard model

In the figure below, the M1 models are trained using the original first-pass seed set coding, and the M2 models are trained using the same documents, but with post-adjudication second-pass coding. The M3 models are symmetric soft-labeled models, tuned using parameter α ranging from 0 to 1. With α=1, the M3 soft label model is equal to the M1 model; with α=0, M3 models are equal to M2 models. The curve in Figure 1 illustrates the MAP (mean average precision) of each type of model, averaged across all 8 topics. It shows that, on average, M3 models outperform both M1 models and M2 models (with α=0.75), and we found that this holds true across all of the individual topics as well, with the exception of topic 205.

Figure 1: Soft labeling compared to binary labeling – symmetric soft labeling

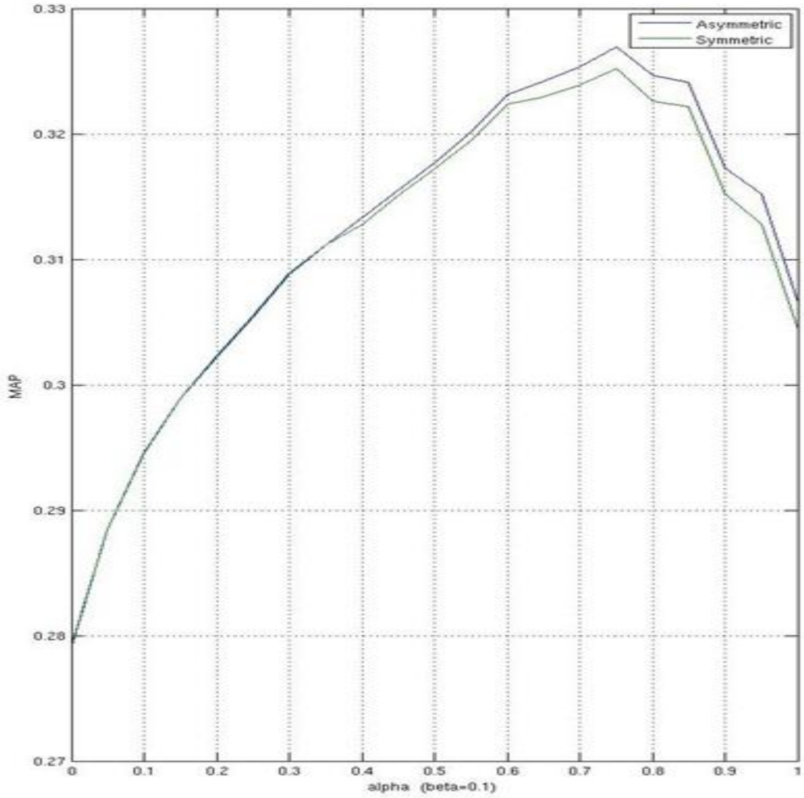


Surprisingly, we observed that M1 models outperform M2 models on average, and for 6 of the 8 topics individually. It is counterintuitive that M2 models, which are coded along the same lines as the gold standard evaluation sets, are unable to discriminate responsive and non-responsive documents more effectively than their M1 counterparts. We hypothesize that this results from the fact that adjudication favored responsive to non-responsive coding changes, which lowered rates of responsiveness. With fewer responsive documents available for training, M2 model performance suffered. In other words, even with standard binary coding, annotating borderline documents as responsive, rather than narrowly considering them non-responsive, provides valuable information to the classifier.

3.3 Soft labeling models: symmetric versus asymmetric

Figure 2 compares symmetric and asymmetric soft labeling models. The green curve provides the averaged MAP of the soft-labeled symmetric models. The blue curve provides the same measures for the asymmetric models with α ranging from 0 to 1 and β set to 0.9. The strongest models are observed at approximately α = 0.75 for both types, but the asymmetric model outperforms the symmetric model at that point. Results show average improvements of 7.2% on MAP (relative) and 4% on the average NDCG (normalized discounted cumulative gain). While the figure reports only the macro-average of MAP, a statistical significance test (one-tailed Wilcoxon T test) was applied on the set of 8 (paired) observations corresponding to the 8 topics, leading to a rejection of the null hypothesis (median difference between the pairs is zero) at a significance level of 0.01.

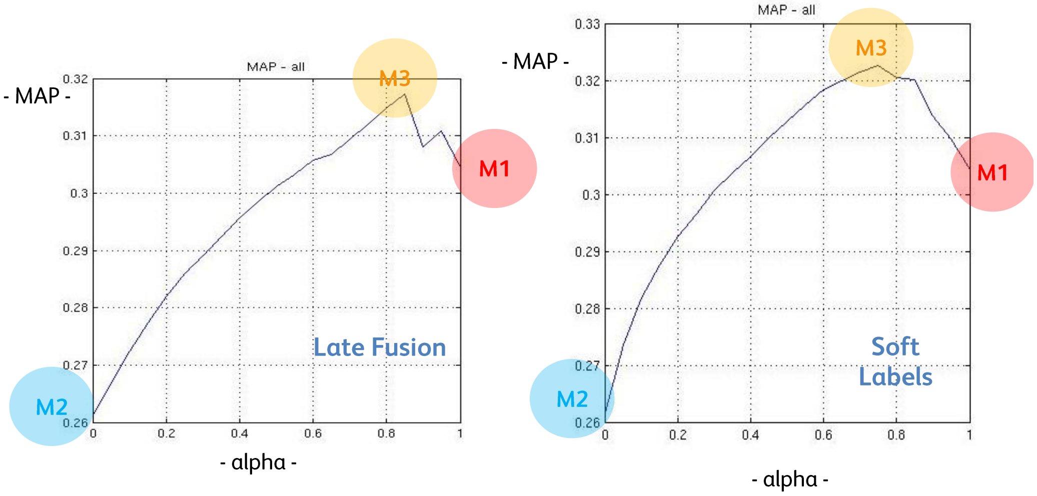
Figure 2: Symmetric vs. asymmetric soft labeling



3.4 Soft labeling versus late fusion

The comparison of classification performance for the symmetric soft labeling and the late fusion strategies is given in Figure 3. We compared these because they use only one parameter (α) for combining both sources of coding. Recall that a late fusion model (denoted as M5 in the Figure) uses convex combination with $\alpha \in [0; 1]$ to combine the classifier scores of the M1 and M2 models. MAP averages across all topics are presented on the left in Figure 3 for the late fusion M5 model and on the right for the symmetric soft labeling M3 model. The best M3 soft label model, occurring at approximately $\alpha = 0.75$, outperforms the late fusion model. The asymmetric soft labeling approach (not represented on the figure) gives a relative improvement of 2.5% with respect to the late fusion strategy, as far as average MAP is concerned.

Figure 3: Late fusion compared to soft labeling



4. Related Work

Related research falls into three main categories: 1) studies on soft or “fuzzy” labeling of training samples, 2) studies on the detection and correction of coding errors, and 3) studies on the use of multiple classifiers and/or training sample weighting.

Outside the context of document review for litigation, soft or fuzzy labeling approaches to text classification are commonly considered when training samples are not coded with “hard” yes/no assessments. This approach is appropriate in situations where labeling experts are given the option of choosing several categories for their assessments or when they are asked to supply an indication of the degree of certainty they would attach to their judgments. Soft or fuzzy modeling is also used in situations where multiple reviewers code the same documents. In the above cases, the fuzzy labels are generally provided directly by the labeling expert or they are obtained by taking the mean of multiple reviewers’ scores and normalizing. [Thiel et al., 2007] present a fuzzy-input fuzzy-output Support Vector Machine approach that both accepts and produces fuzzy labels.

[Ipeirotis et al., 2013] also explore issues involving multiple sources of labeling for training data, but they focus on the distinct challenge of incorporating repeated labeling to improve training for supervised learning classifiers. Online outsourcing systems for labeling, such as Amazon Mechanical Turk, can provide multiple low-cost, non-expert labels for training samples, but very often this results in noisy labeling. [Ipeirotis et al., 2013] propose more complex repeated labeling strategies and illustrate the benefit of carefully selecting the items in the samples to be labeled repeatedly. Beyond majority voting, which is the simplest method for integrating multiple labelers’ judgments, “probabilistic labels,” similar to soft labels, are used to model the uncertainty of labelers’ judgments.

When simple yes/no labeling is being used but manual correction is too expensive to apply to all coded documents, processes are needed to automatically detect and correct labeling errors. In some cases, a distinction can readily be made between higher and lower quality assessments. For instance, there may be a clear segregation of high-quality samples coded by experienced subject matter experts and low-quality samples coded by junior reviewers, as in [Malik and Bhardwaj, 2011]. These authors addressed the problem of noisy training data for news article classification, proposing that samples with high-quality labels from expert journalist reviewers be used to automatically correct lower quality labels. Their method uses classifier predictions, from classifiers built using only high-quality labeled samples, and clustering to extend high-quality labels to noisy lower quality samples.

There are, however, many occasions when no a priori assumptions can reasonably be made regarding the quality of the training samples. [Fukumoto and Suzuki, 2004] present a method for coping with this by first identifying potentially erroneously coded samples using Support Vector Machines (SVM). Those samples are removed from the training set in order to build a classifier of higher quality that is, in turn, used to generate predictions for the eliminated samples. They employ a loss function based on the entropy of the posterior class distribution to either validate the original manual coding for the potentially erroneously coded samples or to assign a corrected label based on the model predictions.

In the context of text classification, [Esuli and Sebastiani, 2009] present three methods for detecting and correcting likely mislabeled data in classifier training samples. These methods involve using the trained models (or a committee of models, in the spirit of a “boosting algorithm” family) together with their associated confidence scores to identify inconsistency between the original and predicted labels. Our work differs from these approaches in that we act upstream in the process, before classifier training, by implementing a new coding scheme that combines labels generated at different phases of the training sample review, independent of any specific classification method.

The most common approach for incorporating multiple labels in training samples is to use binary codes together with weighting strategies. The weighting is generally achieved by duplicating certain documents within the training set or by applying multiplying weights to sample features before or during the classifier learning phase. Weighting techniques can be coupled with multi-stage or cascading classification – i.e., using a first stage classifier whose output is used by a second stage classifier. This technique can be used for "boosting" the learning algorithm [Schapire, 2003]. The boosting strategy involves utilizing several classifiers sequentially, successively weighting the training samples to place the greatest emphasis on items that have been misclassified most often by preceding classifiers. The thinking is that items that have been misclassified before are more difficult to classify. Thus, they are boosted in importance via reweighting for the next classifier in the sequence. This approach could be adopted in the context of classification for e-discovery document review as well, by both reincorporating documents whose coding has been updated and duplicating those documents to highlight the correct treatment. However, these "hard to classify" documents in a responsiveness review may be genuinely ambiguous or arguable, because they contain both responsive and non-responsive language. If so, duplicating those items in the training may simply lead to more confusion for the classifier.

5. Conclusion

In this paper we examined soft labeling for multi-pass document review as a means of addressing the issue of reviewer disagreement and coding conflict in manual document review. The central idea motivating this approach is the recognition that disagreement among reviewers regarding the coding for certain documents need not be viewed simply as noise to be corrected in training and testing data via burdensome QC. On the contrary, it may be possible to take advantage of coding discrepancies to obtain extra information about the text of these documents that can be utilized to enhance classifier performance.

Our experiments using soft labeling models to classify Enron data from the 2012 TREC Legal Track support this idea. They show that allowing the classifier to learn from "partially responsive" soft-labeled documents leads to superior results over training without QC or training with binary codes that have simply been updated categorically following the multi-pass review process. In fact, our experiments indicate that there are occasions when simply using updated binary coding, as given by the final level of review, can result in poorer performance than using the original "incorrect" labels. In practice it is often difficult, if not impossible, to discriminate reliably between pure human coding errors and cases of true ambiguity or arguable document coding. Our method does not attempt to make this distinction, but, regardless of the sources of coding conflicts, our experiments indicate that "multi-coded" documents can serve as effective tools for enhancing the training and results of machine learning classifiers.

6. References

Ipeirotis, P.G., Provost, F. , Sheng, V. and Wang, J. (2013). *"Repeated Labeling Using Multiple Noisy Labelers".* Data Mining and Knowledge Discovery pp. 1-40, March 16, 2013.

Nelson, Eli. (2011). "*The False Dichotomy of Relevance: The Difficulty of Evaluating the Accuracy of Discovery Review Methods Using Binary Notions of Relevance*". In Proc. of DESI IV workshop on Setting Standards for Searching Electronically Stored Information. ICAIL 2011 June 6, Pittsburgh PA.

Barnett, T., Godjevac, S., Renders, J. M., Privault, C., Schneider J. and Wickstrom, R. (2009). *"Machine Learning Classification for Document Review"*, In Proc. of the DESI III workshop on Setting Standards for Searching Electronically Stored Information. ICAIL 2009 Twelfth International Conference on Artificial Intelligence and Law, Barcelona, Spain.

Grossman, M. R. and Cormack, G. V. (2011a). "*Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?"*. In Proc. of DESI IV workshop on Setting Standards for Searching Electronically Stored Information. June 6, 2011. ICAIL 2011, University of Pittsburgh School of Law, PA. (http://www.umiacs.umd.edu/~oard/desi4/).

Grossman, M. R. and Cormack, G. V. (2011b). "*Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*". In Richmond Journal of Law and Technology, Vol. XVII, Issue 3, Article 11 (2011).

Cormack, G.V., Grossman, M.R, Hedin, B. and Oard, D.W. (2010). "*Overview of the TREC-2010 Legal Track*". In Working Notes of the Nineteenth Text Retrieval Conference, pp. 30-38, Gaithersburg, MD, 2010.

Balaji K., Carin, L., Figueiredo, M.A.T. and Hartemink, A.J. (2005) "*Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds*". In IEEE Transactions on Pattern Analysis and Machine Intelligence archive, Vol 27, Issue 6, pp. 957- 68, ISSN:0162-8828.

Agresti, A. (2007). "*Building and applying logistic regression models"*. An Introduction to Categorical Data Analysis. Hoboken, New Jersey: Wiley. pp. 138. ISBN 978-0-471-22618-5.

Hedin, B., Tomlinson, S., Baron, J. R. and Oard, D. W. (2010). "*Overview of the TREC 2009 Legal Track*". In the Eighteenth Text REtrieval Conference TREC 2009 Proceedings. NIST Special Publication: SP 500-278 at: http://trec.nist.gov/pubs/trec18/t18_proceedings.html

Schapire, R. E. (2003). "*The boosting approach to machine learning: An overview*". In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, "*Nonlinear Estimation and Classification"*. Springer, 2003.

Thiel, C., Scherer, S. and Schwenker, F. (2007). "*Fuzzy-Input Fuzzy-Output One-Against-All Support Vector Machines*". In KES 3, Vol. 4694 Springer (2007), pp 156-165.

Hofmann, T. (1999). "*Probabilistic Latent Semantic Analysis*". Proc. 15th Conf. on Uncertainty in Artificial Intelligence. pp. 289-296. Morgan Kaufmann.

Malik, H.H. and Bhardwaj, V.S. (2011)."*Automatic Training Data Cleaning for Text Classification*". In proc. ICDM Workshops 2011, pp. 442-449.

Fukumoto, F. and Suzuki, Y. (2004). "*Correcting category errors in text classification*". In Proceedings of COLING 2004, pp. 868-874, Geneva.

Esuli, A. and Sebastiani, F. (2009). "*Training data cleaning for text classification*". In Proc. of the 2nd Int. Conference on Theory of Information Retrieval: ICTIR '09, pp. 29–41.