Behavioral Factors in Interactive Training of Text Classifiers

Burr Settles
Machine Learning Department
Carnegie Mellon University
Pittsburgh PA 15213, USA
bsettles@cs.cmu.edu

Xiaojin Zhu
Computer Sciences Department
University of Wisconsin
Madison WI 53715, USA
jerryzhu@cs.wisc.edu

Abstract

This paper describes a user study where humans interactively train automatic text classifiers. We attempt to replicate previous results using multiple "average" Internet users instead of a few domain experts as annotators. We also analyze user annotation behaviors to find that certain labeling actions have an impact on classifier accuracy, drawing attention to the important role these behavioral factors play in interactive learning systems.

1 Introduction

There is growing interest in methods that incorporate human domain knowledge in machine learning algorithms, either as priors on model parameters or as constraints in an objective function. Such approaches lend themselves well to natural language tasks, where input features are often discrete variables that carry semantic meaning (e.g., words). A *feature label* is a simple but expressive form of domain knowledge that has received considerable attention recently (Druck et al., 2008; Melville et al., 2009). For example, a single feature (word) can be used to indicate a particular label or set of labels, such as "excellent" ⇒ positive or "terrible" ⇒ negative, which might be useful word-label rules for a sentiment analysis task.

Contemporary work has also focused on making such learning algorithms *active*, by enabling them to pose "queries" in the form of feature-based rules to be labeled by annotators in addition to — and sometimes lieu of — data instances such as documents (Attenberg et al., 2010; Druck et al., 2009). These concepts were recently implemented in a practical system for *interactive* training of text classifiers called DUALIST[1]. Settles (2011) reports that, in user experiments with real annotators, humans were able to train near state of the art classifiers with only a few minutes of effort. However, there were only five subjects, who were all computer science researchers. It is possible that these positive results can be attributed to the subjects' implicit familiarity with machine learning and natural language processing algorithms.

This short paper sheds more light on previous experiments by replicating them with many more human subjects, and of a different type: non-experts recruited through the Amazon Mechanical Turk service[2]. We also analyze the impact of annotator behavior on the resulting classifiers, and suggest relationships to recent work in curriculum learning.

2 DUALIST

Figure 1 shows a screenshot of DUALIST, an interactive machine learning system for quickly building text classifiers. The annotator is allowed to take three kinds of actions: ① label query documents (instances) by clicking class-label buttons in the left panel, ② label query words (features) by selecting them from the class-label columns in the right panel, or ③ "volunteer" domain knowledge by typing labeled words into a text box at the top of each class column. The underlying classifier is a naïve Bayes variant combining informative priors,

[1] http://code.google.com/p/dualist/
[2] http://mturk.com

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 563–567,
Montréal, Canada, June 3-8, 2012.

Figure 1: The DUALIST user interface.

maximum likelihood estimation, and the EM algorithm for fast semi-supervised training. When a user performs action ① or ②, she labels queries that should help minimize the classifier's uncertainty on unlabeled documents (according to active learning heuristics). For action ③, the user is free to volunteer any relevant word, whether or not it appears in a document or word column. For example, the user might volunteer the labeled word "oscar" ⇒ positive in a sentiment analysis task for movie reviews (leveraging her knowledge of domain), even if the word "oscar" does not appear anywhere in the interface. This flexibility goes beyond traditional active learning, which restricts the user to feedback on items queried by the learner (i.e., actions ① and ②). After a few labeling actions, the user submits her feedback and receives the next set of queries in real time. For more details, see Settles (2011).

3 Experimental Setup

We recruited annotators through the crowdsourcing marketplace Mechanical Turk. Subjects were shown a tutorial page with a brief description of the classification task, as well as a cartoon of the interface similar to Figure 1 explaining the various annotation options. When they decided they were ready, users followed a link to a web server running a customized version of DUALIST, which is an open source web-based application. At the end of each trial, subjects were given a confirmation code to receive payment.

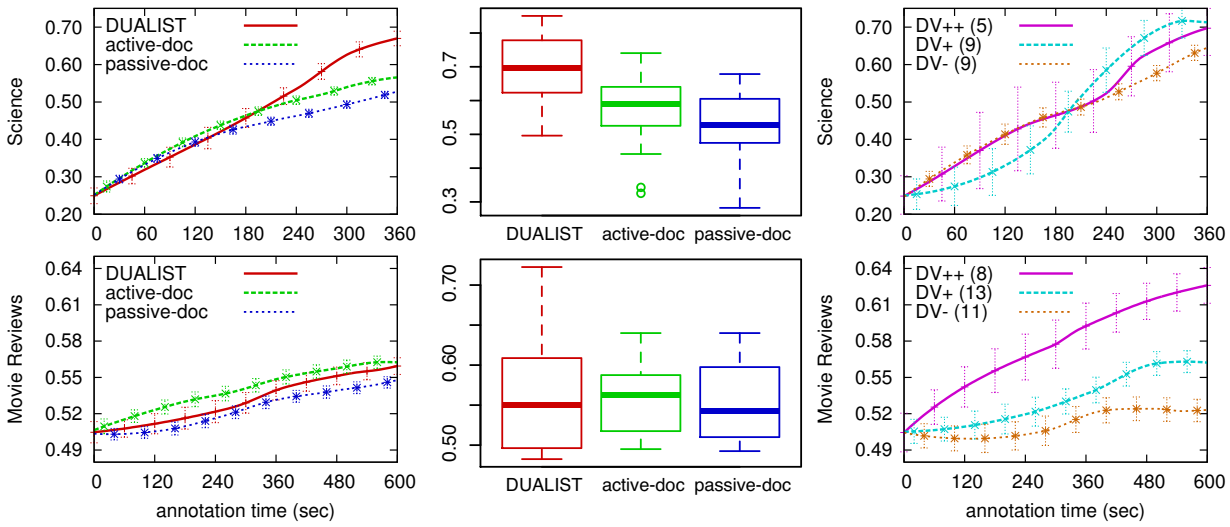
We conducted experiments using two corpora from the original DUALIST study: *Science* (a subset of the 20 Newsgroups benchmark: cryptography, electronics, medicine, and space) and *Movie Reviews* (a sentiment analysis collection). These are not specialized domains, i.e., we could expect average Internet users to be knowledgable enough to perform the annotations. While both are generally accessible, these corpora represent different types of tasks and vary both in number of categories (four vs. two) and difficulty (Movie Reviews is known to be harder for learning algorithms). We replicated the same experimental conditions as previous work: *DUALIST* (the full interface in Figure 1), *active-doc* (the left-hand ① document panel only), and *passive-doc* (the ① document panel only, but with texts selected at random and not queried by active learning).

For each condition, we recruited 25 users for the Science corpus (75 total) and 35 users for Movie Reviews (105 total). We were careful to publish tasks on MTurk in a way that no one user annotated more than one condition. Some users experienced technical difficulties that nullified their work, and four appeared to be spammers[3]. After removing these subjects from the analysis, we were left with 23 users for the Science DUALIST condition, 25 each for the two document-only conditions (73 total), 32 users for the Movie Reviews DUALIST condition, and 33 each for the document-only conditions (98 total). DUALIST automatically logged data about user actions and model accuracies as training progressed, although users could not see these statistics. Trials lasted 6 minutes for the Science corpus and 10 minutes for Movie Reviews. We did advertise a "bonus" for the user who trained the best classifier to encourage correctness, but otherwise offered no guidance on how subjects should prioritize their time.

4 Results

Figure 2(a) shows learning curves aggregated across all users in each experimental condition. Curves are LOESS fits to classifier accuracy over time: locally-weighted polynomial regressions (Cleveland et al., 1992) ± 1 standard error, with the actual user data points omitted for clarity. For the Science task (top), DUALIST users trained significantly better classifiers after about four minutes of annotation time. Document-only active learning also outperformed

[3] A spammer was ruled to be one whose document error rate (vs. the gold standard) was more than double the chance error, and whose feature labels appeared to be arbitrary clicks.



(a) learning curves (b) final classifier accuracies (c) behavioral subgroup curves

Figure 2: (a) Learning curves plotting accuracy vs. actual annotation time for the three conditions. Curves are LOESS fits (± 1 SE) to all classifier accuracies at that point in time. (b) Box plots showing the distribution of final accuracies under each condition. (c) Learning curves for three behavioral subgroups found in the DUALIST condition. The DV++ group volunteered many labeled words (action ③), DV+ volunteered some, and DV- volunteered none.

standard passive learning, which is consistent with previous work. However, for Movie Reviews (bottom), there is little difference among the three settings, and in fact models trained with DUALIST appear to lag behind active learning with documents.

Figure 2(b) shows the distribution of final classifier accuracies in each condition. For Science, the DUALIST users are significantly better than either of the baselines (two-sided KS test, $p < 0.005$). While the differences in DUALIST accuracies are not significantly different, we can see that the top quartile does much better than the two baselines. Clearly some DUALIST users are making better use of the interface and training better classifiers. How?

It is important to note that users in the active-doc and passive-doc conditions can only choose action ① (label documents), whereas those in the DUALIST condition must allocate their time among three kinds of actions. It turns out that the annotators exhibited very non-uniform behavior in this respect. In particular, activity of action ③ (volunteer labeled words) follows a power law, and many subjects volunteered no features at all. By inspecting the distribution of these actions for natural breakpoints, we identified three subgroups of DUALIST users: DV++ (many volunteered words), DV+ (some words), and DV- (none; labeled queries only). Note

	Movie Reviews		Science	
Group	# Words	Users	# Words	Users
DV++	21–62	8	24–42	5
DV+	1–15	13	2–19	9
DV-	0	11	0	9

Table 1: The range of volunteered words and number of users in each behavioral subgroup of DUALIST subjects.

that DV- is *not* functionally equivalent to the active-doc condition, as users in the DV- group could still view and label word queries. The three behavioral subgroups are summarized in Table 1.

Figure 2(c) shows learning curves for these three groups. We can see that the DV++ and DV+ groups ultimately train better classifiers than the DV- group, and DV++ also dominates both the active and passive baselines from Figure 2(a). The DV++ group is particularly effective on the Movie Reviews corpus. This suggests that a user's choice to volunteer more labeled features — by occasionally side-stepping the queries posed by the active learner and directly injecting their domain knowledge — is a good predictor of classifier accuracy on this task.

To tease apart the relative impact of other behaviors, we conducted an ordinary least-squares regression to predict classifier accuracy at the end of a trial. We included the number of user events for each ac-

tion as independent variables, plus two controls: the subject's document error rate in [0,1] with respect to the gold standard, and class entropy in [0, $\log C$] of all labeled words (where C is the number of classes). The entropy variable is meant to capture how "balanced" a user's word-labeling activity was for actions ② and ③, with the intuition that a skewed set of words could confuse the learner, by biasing it away from categories with fewer labeled words.

Table 2 summarizes these results. Surprisingly, query-labeling actions (① and ②) have a relatively small impact on accuracy. The number of volunteered words and entropy among word labels appear to be the only two factors that are somewhat significant: the former is strongest in the Movie Reviews corpus, the latter in Science[4]. Interestingly, there is a strong positive correlation between these two factors in the Movie Reviews corpus (Spearman's $\rho = 0.51$, $p = 0.02$) but not in Science ($\rho = 0.03$). When we consider change in word label entropy over time, the Science DA++ group is balanced early on and becomes steadily more so on average .ıllı, whereas DA+ goes for several minutes before catching up (and briefly overtaking) ..ıll. This may account for DA+'s early dip in accuracy in Figure 2(c). For Movie Reviews, DA++ is more balanced than DA+ throughout the trial. DA++ labeled many words that were also class-balanced, which may explain why it is the best consistently-performing group. As is common in behavior modeling with small samples, the data are noisy and the regressions in Table 2 only explain 33%–46% of the variance in accuracy.

	Movie Reviews		**Science**	
Action	β	SE	β	SE
(intercept)	0.505	0.038 ***	0.473	0.147 **
① label query docs	0.001	0.001	0.005	0.005
② label query words	-0.001	0.001	0.000	0.001
③ volunteer words	0.002	0.001 *	0.000	0.002
human error rate	-0.036	0.109	-0.328	0.230
word label entropy	0.053	0.051	0.201	0.102 .
	$R^2 = 0.4608$ **		$R^2 = 0.3342$	

*** $p < 0.001$ ** $p < 0.01$ * $p < 0.05$. $p < 0.1$

Table 2: Linear regressions estimating the accuracy of a classifier as a function of annotator actions and behaviors.

5 Discussion

We were able to partially replicate the results from Settles (2011). That is, for two of the same data sets, some of the subjects using DUALIST significantly outperformed those using traditional document-only interfaces. However, our results show that the gains come not merely from the interface itself, but from which labeling actions the users chose to perform. As interactive learning systems continue to expand the palette of interactive options (e.g., labeling and/or volunteering features), understanding how these options impact learning becomes more important. In particular, training a good classifier in our experiments appears to be linked to (1) volunteering more labeled words, and (2) maintaining a class balance among them. Users who exhibited both of these behaviors — which are possibly artifacts of their good intuitions — performed the best.

We posit that there is a conceptual connection between these insights and *curriculum learning* (Bengio et al., 2009), the commonsense notion that learners perform better if they begin with clear and unambiguous examples before graduating to more complex training data. A recent study found that some humans use a curriculum strategy when teaching a 1D classification task to a robot (Khan et al., 2012). About half of those subjects alternated between extreme positive and negative instances in a relatively class-balanced way. This behavior was explained by showing that it is optimal under an assumption that, in reality, the learning task has many input features for which only one is relevant to the task.

Text classification exhibits similar properties: there are many features (words), of which only a few are relevant. We argue that labeling features can be seen as a kind of training by curriculum. By volunteering labeled words in a class-balanced way (especially early on), a user provides clear, unambiguous training signals that effectively perform feature selection while biasing the classifier toward the user's hypothesis. Future research on mixed-initiative user interfaces might try to detect and encourage these kinds of annotator behaviors, and potentially improve interactive machine learning outcomes.

[4]Science has four labels and a larger entropy range, which might explain the importance of the entropy factor here. Also, labels are more related to natural clusterings in this corpus (Nigam et al., 2000), so class-balanced priors might be key for DUALIST's semi-supervised EM procedure to work well.

Acknowledgments

This work was funded in part by DARPA, the National Science Foundation (under grants IIS-0953219 and IIS-0968487), and Google.

References

J. Attenberg, P. Melville, and F. Provost. 2010. A unified approach to active dual supervision for labeling features and examples. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 40–55. Springer.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 119–126. Omnipress.

W.S. Cleveland, E. Grosse, and W.M. Shyu. 1992. Local regression models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*. Wadsworth & Brooks/Cole.

G. Druck, G. Mann, and A. McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602. ACM Press.

G. Druck, B. Settles, and A. McCallum. 2009. Active learning by labeling features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 81–90. ACL Press.

F. Khan, X. Zhu, and B. Mutlu. 2012. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 1449–1457. Morgan Kaufmann.

P. Melville, W. Gryc, and R.D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1275–1284. ACM Press.

K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134.

B. Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1467–1478. ACL Press.