Appendix

A Theoretical results for the probabilistic disjunction model

A.1 Proof of Theorem 1

Proof. The problem is clearly in NP. To show hardness, we will use a reduction from 3SAT.

Given a 3SAT instance $\phi(x_1, \ldots, x_q) = C_1 \wedge C_2 \wedge \cdots \wedge C_p$, where each clause C_j is a disjunction of three literals, create the following topic labeling problem:

- There are $2q$ topics: $t_1, \ldots, t_q, t'_1, \ldots, t'_q$. Think of t_i as corresponding to the positive literal x_i and t'_i the negative literal $\overline{x_i}$.
- For each variable x_i, create a document d_i whose topic distribution $\theta^{(d_i)}$ has probability $1/2$ on t_i and on t'_i and zero elsewhere.
- For each clause C_j, create a document d'_j whose topic distribution puts $1/3$ probability on (the t_i or t'_i corresponding to) each of the literals in C_j.
- The data set consists of document-label pairs $(d_i, 0), (d_i, 1), (d'_j, 1)$: a total of $p + 2q$ labeled documents.

Now, suppose there is an assignment $\ell : \{t_1, \ldots, t_q, t'_1, \ldots, t'_q\} \to \{0, 1, ?\}$ with nonzero likelihood. Then for each labeled document (d, y) there is at least one topic t such that $\theta^{(d)}_t > 0$ and $\ell(t) = y$. Now, document d_i appears with label 0 as well as with label 1. Therefore, one of $\ell(t_i), \ell(t'_i)$ must be 0 and one of them must be 1. If $\ell(t_i) = 0, \ell(t'_i) = 1$, we will assign $x_i = 0$. If $\ell(t_i) = 1, \ell(t'_i) = 0$, we will assign $x_i = 1$. To see that this is a satisfying assignment, pick any clause C_j. The corresponding document d'_j has label 1; therefore at least one of the three topics corresponding to its literals must be assigned label 1 under $\ell(\cdot)$. Hence that literal is assigned a value of 1.

Conversely, if ϕ is satisfiable, then the mapping

$$\ell(t_i) = 0, \ell(t'_i) = 1 \text{ if } x_i = 0$$
$$\ell(t_i) = 1, \ell(t'_i) = 0 \text{ if } x_i = 1$$

has nonzero likelihood. □

A.2 Proof of Theorem 2

Proof. First, fix any t, y with $\ell(t) \neq y$. Under Assumption 1, each time topic t is selected, there is less than a $\lambda/2$ probability that the label is y. Conditioned on n_t, the expected value of n_{ty} is therefore at most $\lambda n_t/2$, and by a multiplicative Chernoff bound,

$$\Pr(n_{ty} \geq \lambda n_t) \leq e^{-n_t \lambda/6},$$

which is $\leq \delta/(Tk)$ if $n_t \geq n_o$.

Likewise, for any predictive feature $t \in P$, the expected value of $n_{t,\ell(t)}$ is at least $2\lambda n_t$. Again using a multiplicative Chernoff bound,

$$\Pr(n_{t,\ell(t)} < \lambda n_t) \leq e^{-n_t \lambda/6}.$$

Taking a union bound over all pairs $(t, y) \in [T] \times [k]$, we conclude that with probability at least $1 - \delta$, the following holds whenever $n_t \geq n_o$:

- If $y \neq \ell(t)$ then $n_{ty} < \lambda n_t$.
- If $t \in P$ then $n_{t,\ell(t)} \geq \lambda n_t$.

Therefore, $\widehat{\ell}(t) = \ell(t)$ for $t \in P$ and ? otherwise. □

A.3 Proof of Theorem 3

Proof. Pick any predictive topic $t \in P$, and let $y = \ell(t)$. For a document x chosen at random,

$$\begin{aligned}\Pr_x(\text{topic } t \text{ selected}) &\geq \Pr_x(\text{document label} = y)\Pr_x(\text{topic } t \text{ selected} \mid \text{document label} = y) \\ &\geq \mathbb{E}_x\left[\frac{\sum_{t':\ell(t')=y}\theta_{t'}(x)}{\sum_{t'\in P}\theta_{t'}(x)} \cdot c_o \frac{\theta_t(x)}{\sum_{t':\ell(t')=y}\theta_{t'}(x)}\right] \;=\; c_o\gamma_t.\end{aligned}$$

Therefore, the expected number of documents that need to be seen before n_t reaches n_o is at most $n_o/(c_o\gamma_t)$. □

B Incorporating feature feedback through regularization

B.1 Proof of Theorem 5

Recall that we wish to bound $R_n(\mathcal{F})$. The powerful results of [Kakade et al., 2009] achieve this for a wide range of cases: for any $\mathcal{F} = \{w : \|w\| \leq W\}$, where $\|\cdot\|$ satisfies a strong convexity property. Specifically, they show

$$R_n(\mathcal{F}) \;\leq\; W \cdot \max_{x\in\mathcal{X}} \|x\|_* \cdot \sqrt{\frac{2}{n}}$$

where $\mathcal{X}$ is the input space, and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

We now apply this bound to our setting, where our regularizer norm is $\|\cdot\|_A$ for positive definite A.

Lemma 7. *Pick any positive definite $p \times p$ matrix A and consider the Mahalanobis norm $\|\cdot\|_A$ on $\mathbb{R}^p$.*

1. *The function $\|\cdot\|_A^2$ is 2-strongly convex. In particular, for any $u, v \in \mathbb{R}^p$ and $0 \leq \alpha \leq 1$,*

$$\alpha\|u\|_A^2 + (1-\alpha)\|v\|_A^2 - \|\alpha u + (1-\alpha)v\|_A^2 \;=\; \alpha(1-\alpha)\|u-v\|_A^2.$$

2. *The dual norm of $\|\cdot\|_A$ is $\|\cdot\|_{A^{-1}}$.*

Proof. The first assertion follows directly by expanding the expression. For the second, we note that the dual norm of $\|\cdot\|_A$ is defined by

$$\|x\|_* = \sup_{\|y\|_A \leq 1} x \cdot y.$$

We will show that this is $\|x\|_{A^{-1}}$.

First, take

$$y = \frac{A^{-1}x}{\sqrt{x^T A^{-1} x}}.$$

Then

$$\|y\|_A^2 = y^T A y = \frac{x^T A^{-1} A A^{-1} x}{x^T A^{-1} x} = 1$$

so $\|y\|_A = 1$. Moreover, $x \cdot y = \sqrt{x^T A^{-1} x} = \|x\|_{A^{-1}}$.

Conversely, pick any y with $\|y\|_A \leq 1$. Then

$$x \cdot y \;=\; x^T A^{-1/2} A^{1/2} y \;=\; (A^{-1/2}x)^T (A^{1/2}y) \;\leq\; \|A^{-1/2}x\|_2 \|A^{1/2}y\|_2 \;=\; \|x\|_{A^{-1}}\|y\|_A \;\leq\; \|x\|_{A^{-1}}.$$

□

If w^* is the sparse target classifier, the function class of interest is $\mathcal{F} = \{w : \|w\|_A \leq \|w^*\|_A\}$ and by [Kakade et al., 2009] we have

$$R_n(\mathcal{F}) \;\leq\; \|w^*\|_A \cdot \max_{x\in\mathcal{X}} \|x\|_{A^{-1}} \sqrt{\frac{2}{n}}$$

Let $R = \{i \in [p] : w_i^* \neq 0\}$ denote the relevant features. We can split any x into its relevant and other components, $x = (x_R, x_o)$, and when we downweight the diagonal R-entries of A by a factor of c, we get

$$\|x\|_{A^{-1}}^2 = \|x_o\|_2^2 + c\|x_R\|_2^2$$

whereas

$$\|w^*\|_A^2 = \frac{1}{c}\|w\|_2^2$$

(assuming we have captured all the features on which w^* is non-zero). Thus

$$R_n(\mathcal{F}) \;\leq\; \|w^*\|_2 \cdot \max_{x \in \mathcal{X}} \sqrt{\left(\frac{1}{c}\|x_o\|_2^2 + \|x_R\|_2^2\right)} \sqrt{\frac{2}{n}}.$$

C Proof of Lemma 6

Proof. Consider the optimization problem for computing the support vector classifier using the Mahalanobis regularizer.

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \frac{1}{2}\|w\|_A^2 + C\sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & \xi_i \geq 0,\ y_i(x_i^T w + b) \geq 1 - \xi_i,\ \forall i. \end{aligned} \tag{1}$$

The Lagrangian of (1) is

$$\begin{aligned} L(w, b, \xi, \mu, \alpha) = \frac{1}{2}\|w\|_A^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i}^{N} \mu_i \xi_i \\ - \sum_{i}^{N} \alpha_i [y_i(x_i^T w + b) - (1 - \xi_i)], \end{aligned}$$

where the α_i, μ_i are the Lagrange multipliers. It easy to see that the Lagrange dual function L_D is

$$L_D(\mu, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T A^{-1} x_j.$$

which corresponds to the ℓ_2-regularized SVM with data $(A^{-1/2}x_i, y_i)$. □

D Experiments

D.1 Data sets

20 NewsGroups: The 20-Newsgroups collection is a set of approximately 20,000 newsgroup documents, partitioned evenly across the 20 different newsgroups. The documents are postings about politics, sports, technology, religion, science etc., and contain subject lines, signature files, and quoted portions of other articles. Some of the newsgroups are very closely related to each other (e.g., IBM computer system hardware *vs* Macintosh computer system hardware), while others are unrelated (e.g., misc for sale *vs* social religion and christian). A processed version of the data set was obtained. The original data set can be found on Jason Rennie's website. [1].

Reuters-21578: This is another widely used collection for text categorization research. The documents appeared on the Reuters newswire in 1987 and were manually classified into several topics by personnel from Reuters Ltd. See Lewis et al. [2004] for further details on the data set. Sub-collections **R10** (10 classes with the highest number of topics) and **R90** (at least one positive and one training example) are usually considered for text categorization tasks. As our goal here was to consider single-labeled data, all the documents with less than or with more than one label were eliminated, resulting in **R8** (8 classes) and **R52** (52 classes).

webkb: This data set contains web pages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base project of the CMU text learning group [2].

cade: The documents in this collection correspond to web pages extracted from the CADE Web Directory, which points to Brazilian web pages classified by human experts in 12 classes, including services, education, sciences, sports, culture etc.

ohsumed: This data set includes medical abstracts from the MeSH (Medical Subject Headings) categories of the year 1991 [3] on 23 cardiovascular disease categories. We only considered documents with a single label.

For each data set we only considered tokens that occurred at least 3 times. Figure 3 below provides a summary of the data as they were used in the experiment.

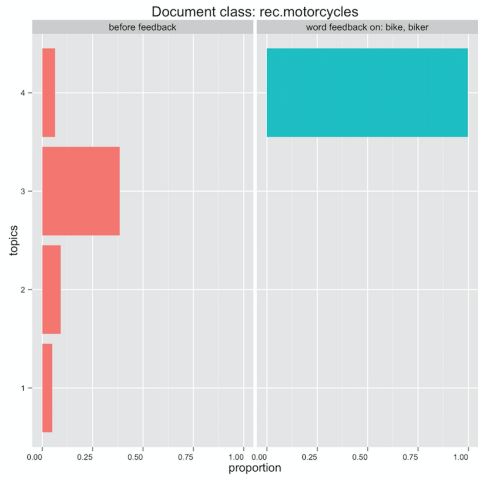
	# tokens	# training docs	# test docs	# topics	# classes
20 NewsGroups (20ng)	33,223	11,293	7,528	200	20
Reuters 8 (R8)	7,744	5,485	2,189	80	8
Reuters 52 (R52)	8,868	6,532	2,568	520	52
cade	68,983	27,322	13,661	120	12
webkb	7,644	2,803	1,396	40	4
ohsumed	13,627	3,357	4,043	230	23

Figure 3: Summary of the datasets and the number of topics used in the experiment

[1] http://qwone.com/ jason/20Newsgroups/
[2] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-4/text-learning/www/index.html
[3] ftp://medir.ohsu.edu/pub/ohsumed

D.2 Results

An example of a PDM from the **20ng** dataset is shown in figure 4. Figures 5 - 10 show our experimental results for each one of the data sets in more detail. Figures 11- 12, show the *amount* of feedback over time.



	Topic 1	Topic 2	Topic 3	Topic 4
1	gener	air	unit	bike
2	process	heat	engin	dod
3	thi	temperatur	cross	ride
4	sinc	water	bnr	motorcycl
5	effect	cold	adjust	bmw
6	anoth	pressur	link	rider
7	requir	hot	pre	helmet
8	real	fan	replac	sun
9	result	effect	nick	drink
10	case	ga	put	biker

Figure 4: Left : Topic representation of a document with the class **rec.motorcycles** before and after feature feedback, on the oracle features **bike** and **biker** Right: Descriptive words of the topics that are present in the document.

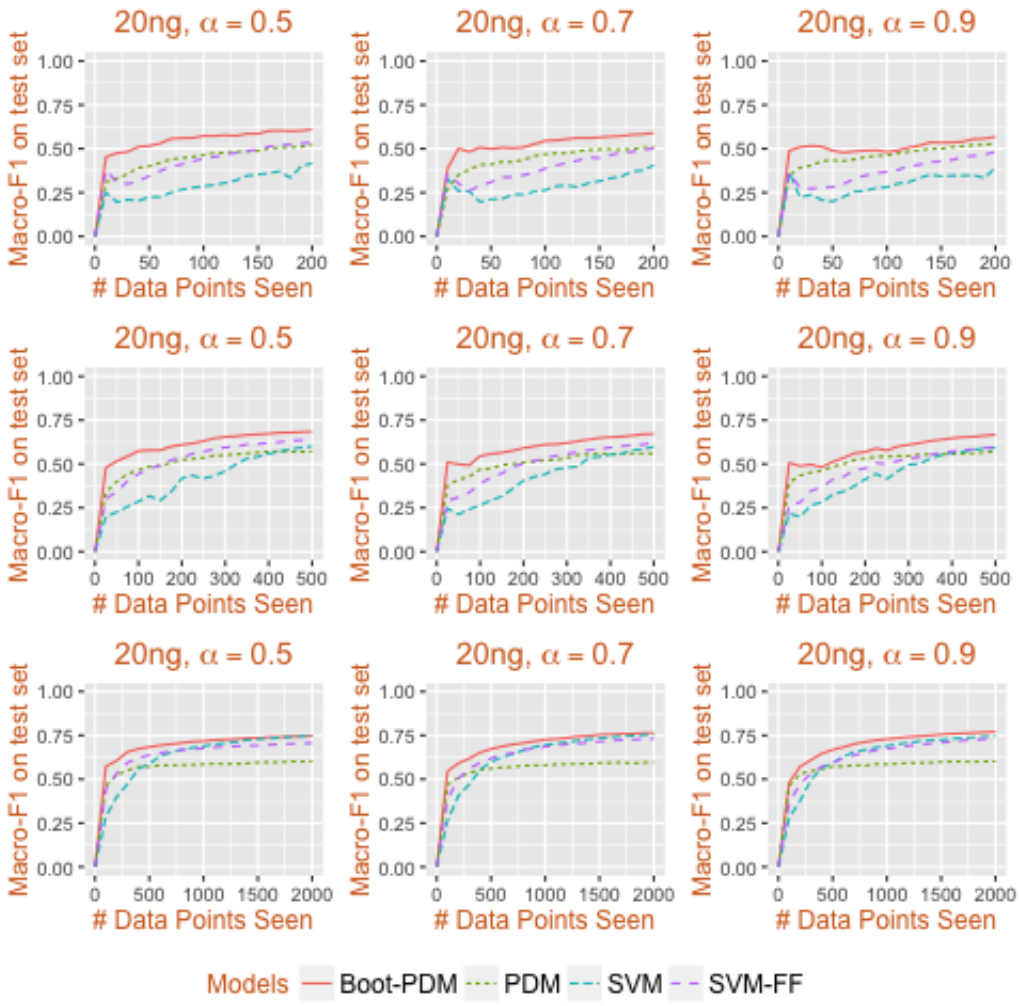


Figure 5: 20ng

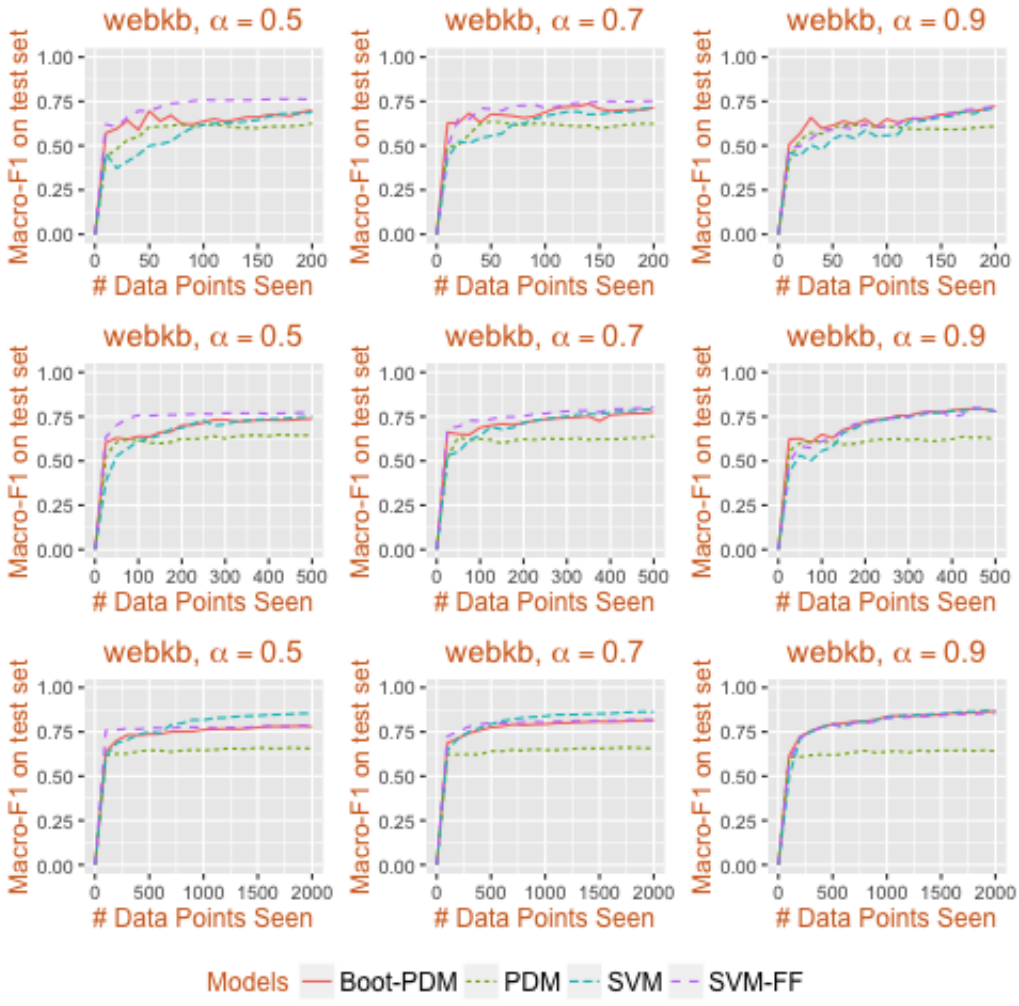


Figure 6: webkb

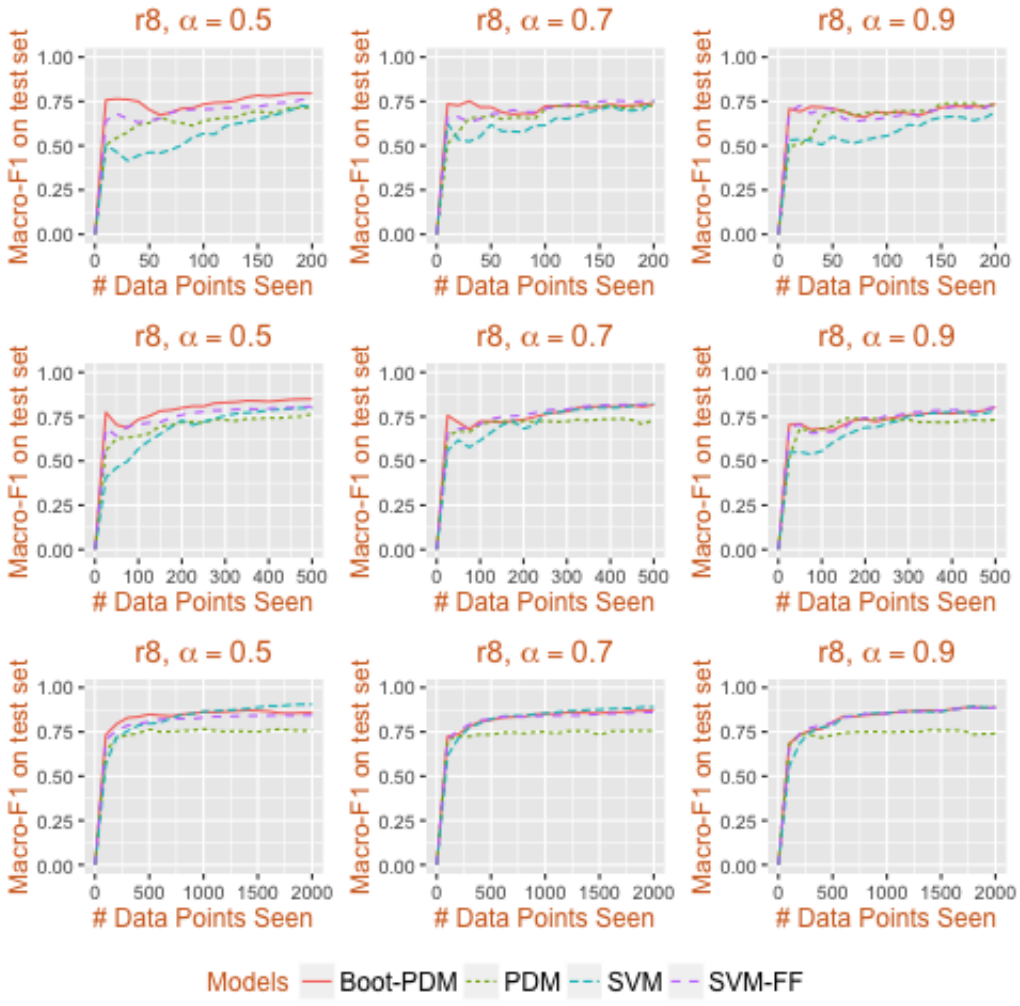

r8, α = 0.5
r8, α = 0.7
r8, α = 0.9
Macro-F1 on test set
Data Points Seen
Models
Boot-PDM
PDM
SVM
SVM-FF

Figure 7: R8

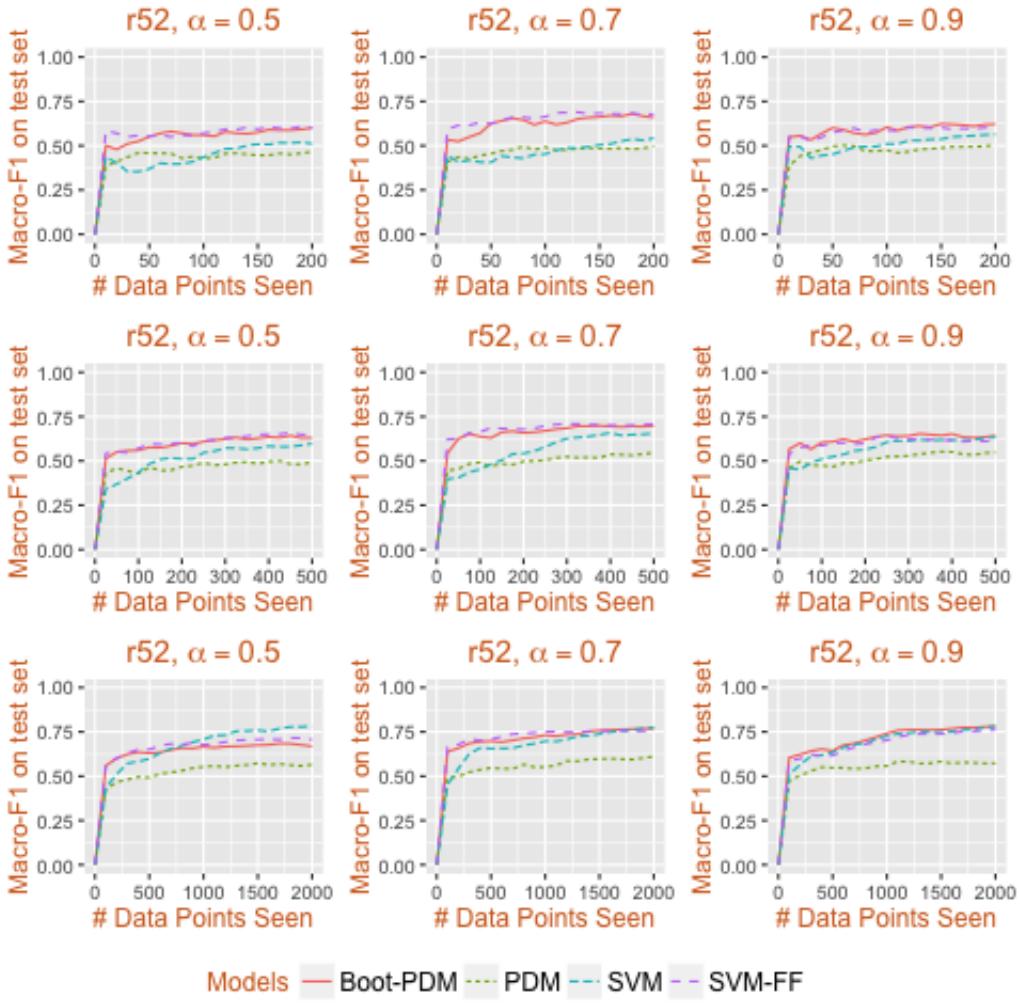

r52, α = 0.5
r52, α = 0.7
r52, α = 0.9
Macro-F1 on test set
Data Points Seen
Models
Boot-PDM
PDM
SVM
SVM-FF

Figure 8: R52

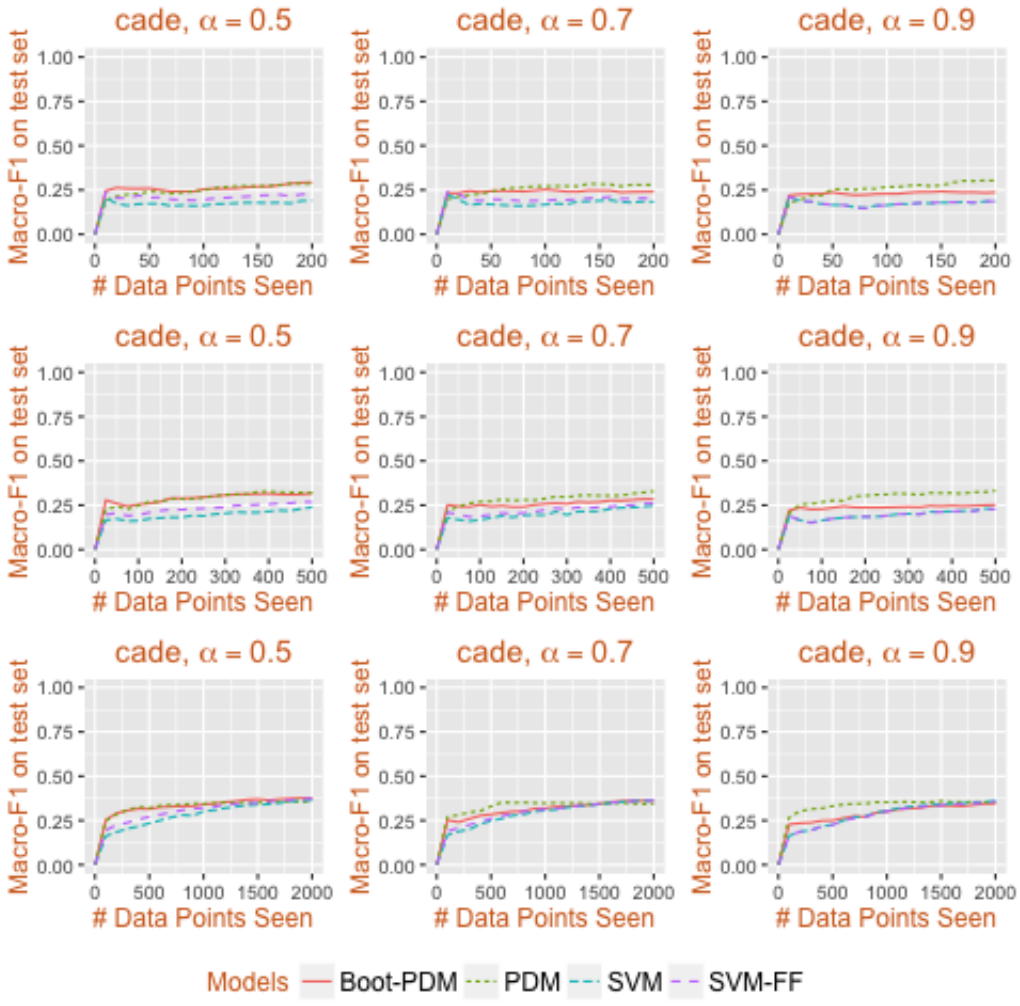


Figure 9: Cade

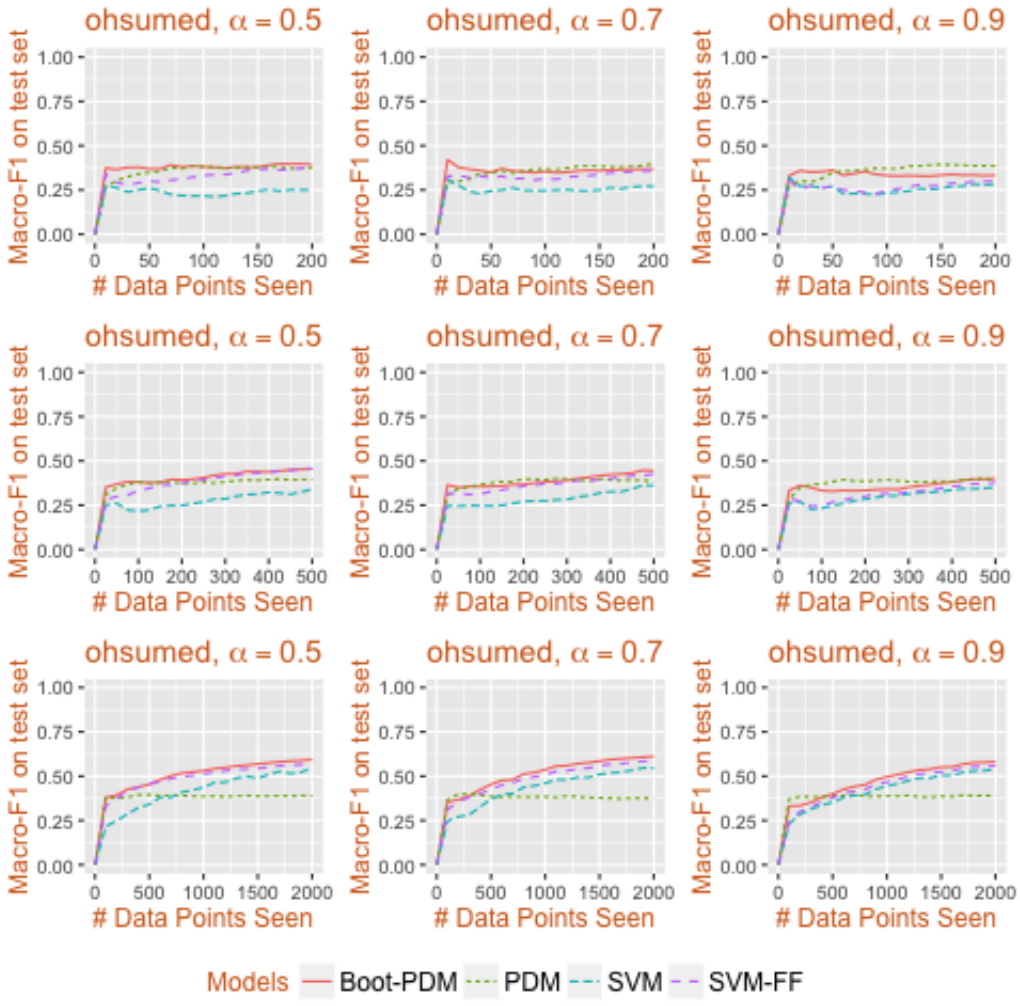


Figure 10: Ohsumed

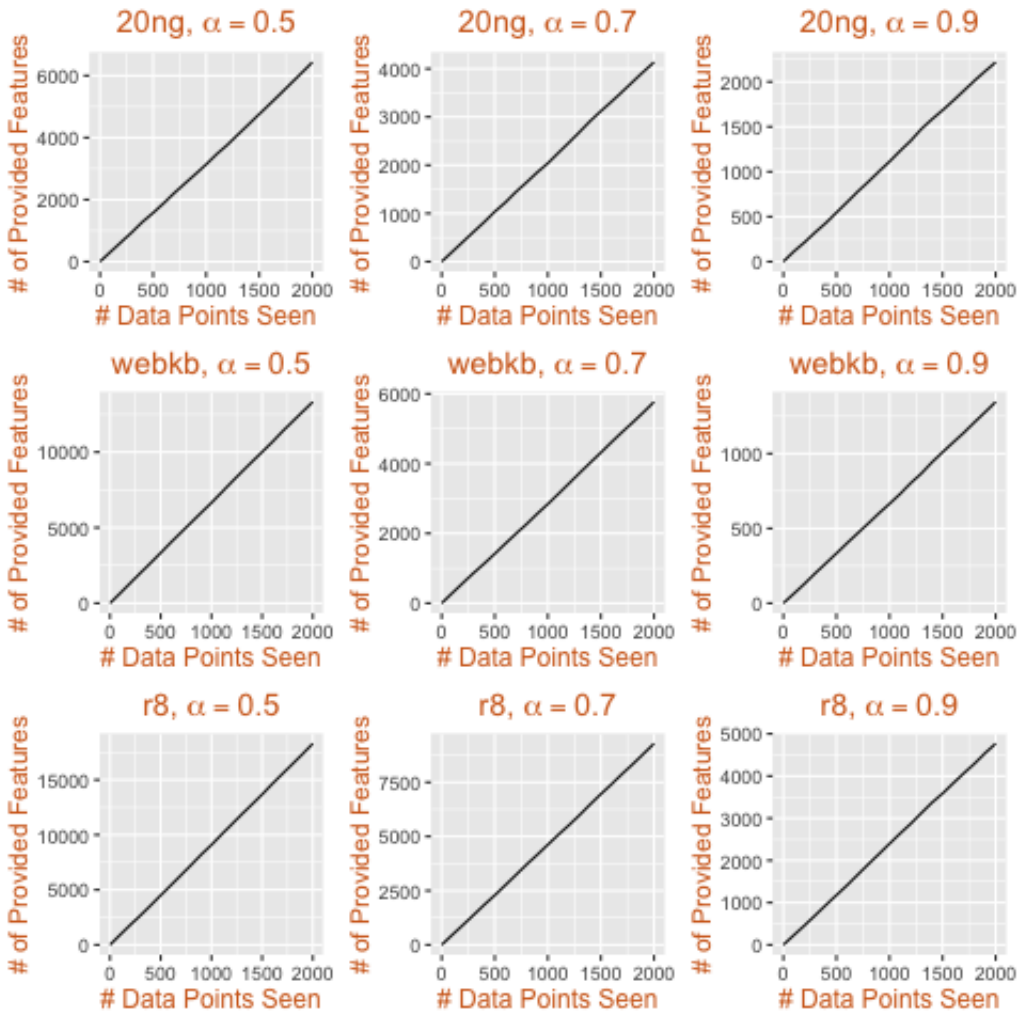


Figure 11: Amount of Feature Feedback

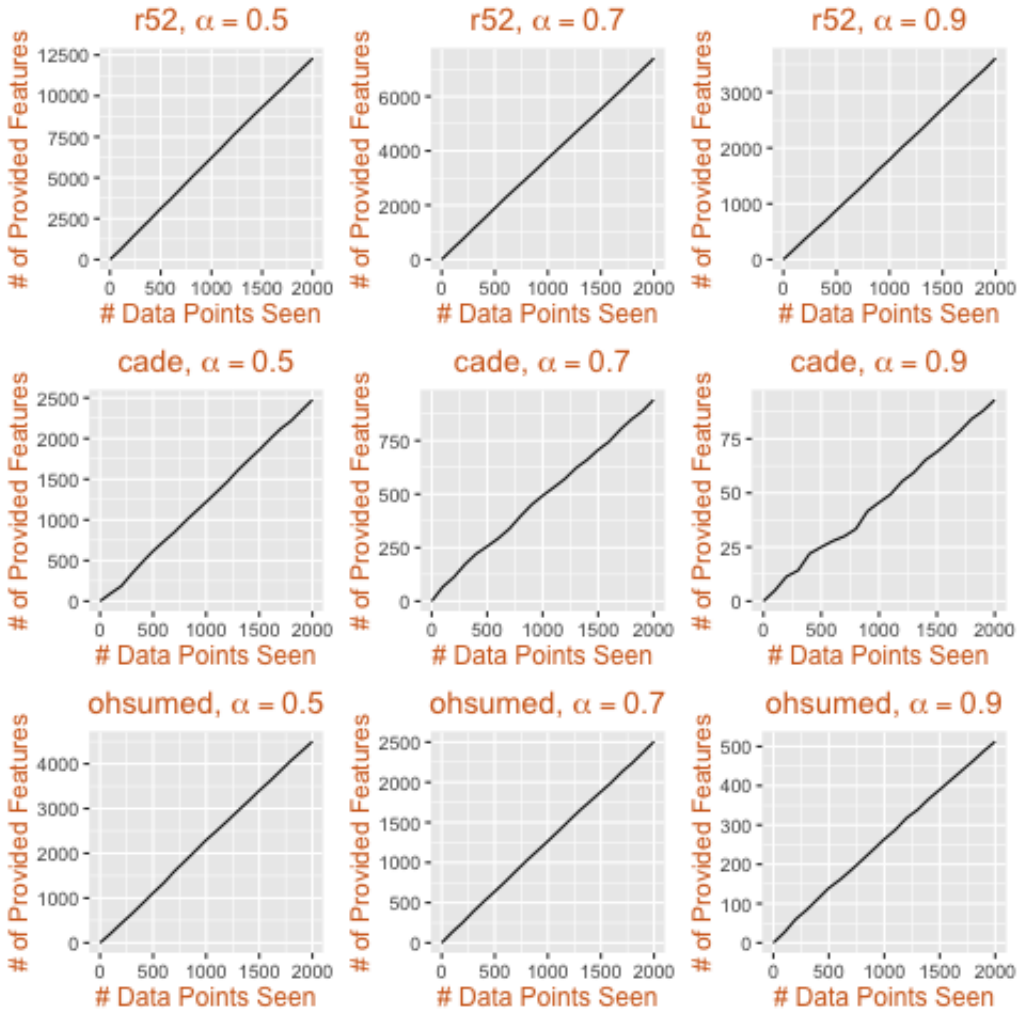


Figure 12: Amount of Feature Feedback

D.3 Human Experiment.

Figure 13 depicts the interface that was used to solicit labels and feature feedback from human annotators. Annotators were given the option to select a number of features from a list. They were also given the ability to insert a feature from the document that was not in the list.

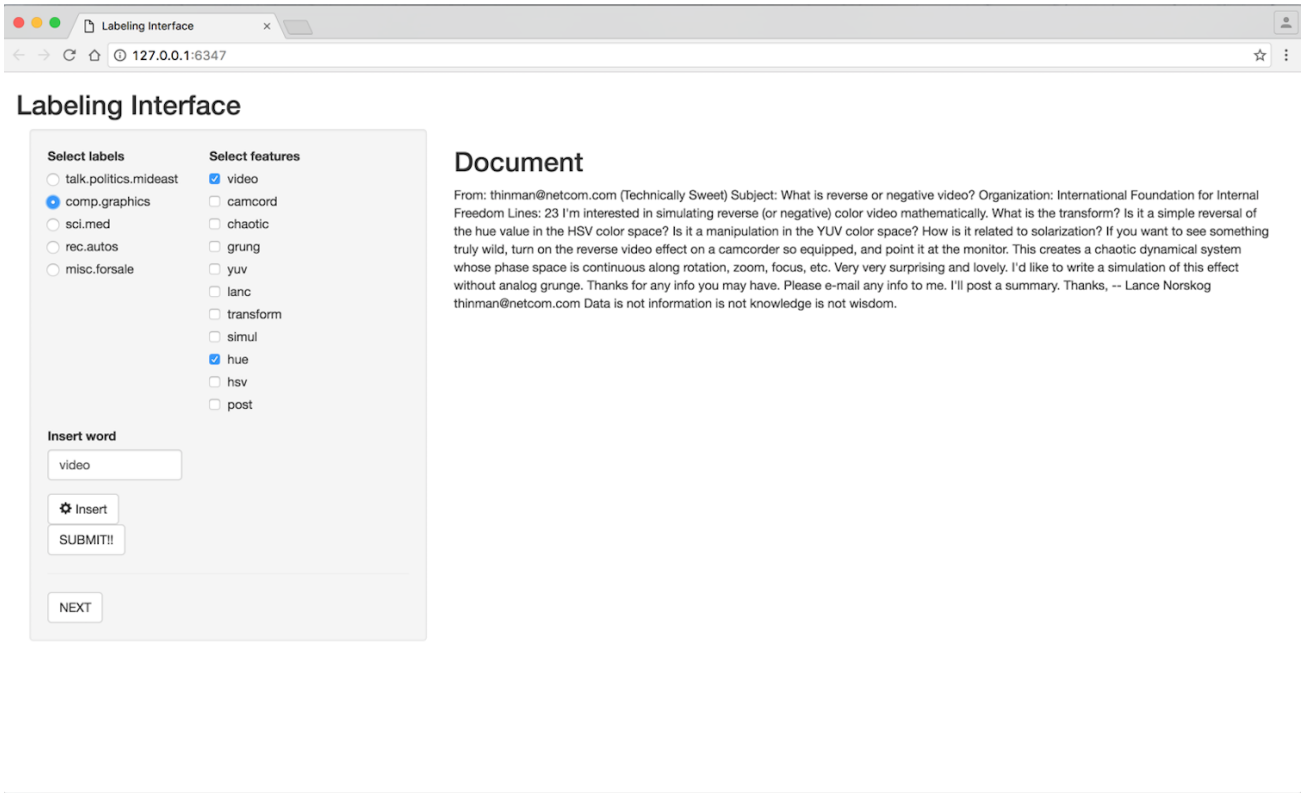


Figure 13: Interface used in Human Experiment