Online Bayesian Models for Personal Analytics in Social Media

Svitlana Volkova and Benjamin Van Durme
Center for Language and Speech Processing, Johns Hopkins University, Baltimore MD 21218, USA
svitlana@jhu.edu, vandurme@cs.jhu.edu

Abstract
Latent author attribute prediction in social media provides a novel set of conditions for the construction of supervised classification models. With individual authors as training and test instances, their associated content (“features”) are made available incrementally over time, as they converse over discussion forums. We propose various approaches to handling this dynamic data, from traditional batch training and testing, to incremental bootstrapping, and then active learning via crowdsourcing. Our underlying model relies on an intuitive application of Bayes rule, which should be easy to adopt by the community, thus allowing for a general shift towards online modeling for social media.

Introduction
The recent explosion of social media has led to an interest in predicting hidden information from the large amounts of freely available content. As compared to the earlier explosion of documents arising from the web, social media content is significantly more personalized, i.e., written in the first person, informal, and often revealing of latent properties of the author. This has become known alternatively as constructing: user demographic models, personal analytics or customer profiling services. Researchers have explored the prediction of latent attributes including gender (Rao et al. 2010; Filippova 2012; Bergsma et al. 2013), age (Nguyen et al. 2013), political preferences (Conover et al. 2011b; Cohen and Ruths 2013; Volkova, Coppersmith, and Van Dume 2014), personality traits (Bachrach et al. 2012), and so on.

However, the majority of this work treats the modeling task much as prior work on non-social media: construct a corpus of labeled materials, and perform supervised classification in a batch setting. This ignores one of the primary distinguishing characteristics of social media content: it is dynamically generated over time, and usually centered within the context of a social network (i.e., friends or other types of associates of the author). Further, different users of the medium contribute to greater or lesser extent: a given user may send one tweet a week, or one tweet an hour, etc. Prior work tends to gloss over this fact by building controlled collections with a large, fixed amount of content assumed per user e.g., 1K tweets. (Zamal, Liu, and Ruths 2012) or even 5K tweets (Rao et al. 2010).

In contrast, Burger et al. and Volkova, Coppersmith, and Van Dume showed the intuitive importance of the amount of content available per user at test time: the more content you have, the better your predictions. This was followed by Van Durme who proposed a model that allowed for incremental updating of classifier predictions over time, as users continued to author new content. This model treated each user as a sort of dynamic feature vector that evolved over time, and assumed access to a pre-trained classification model based on labeled data available a priori, akin to earlier work in the purely batch setting.

Here we go beyond the existing work and propose two novel contributions in mining streaming social media: (1) contrasting Van Durme, we treat each new message as independent evidence which is combined into an incremental user-prediction model as a straightforward application of Bayes Rule; (2) we explore model training in parallel to its application, rather than assuming a previously existing labeled dataset. Also, distinct from Van Durme, but previously explored in the batch-setting by (Zamal, Liu, and
Ruths 2012) and (Volkova, Coppersmith, and Van Dume 2014) we make use of the local user neighborhood in our dynamic model.

Our approach captures the same incremental intuitions as the work by Van Durme, but we situate it within the well understood framework of Bayesian inference: we hope this will encourage others to build upon this effort in constructing more complicated models. Further, by recognizing that both training as well as testing materials are dynamically generated in social media, then possibly coupled to dynamic model feedback via crowdsourcing, this suggests latent author attribute prediction as a rich source for methodological challenges in online and active learning. This work means to give perspective on the various ways this dynamism may be incorporated into an experimental framework. Future work may then choose a particular path and focus on models with further complexity and larger datasets.

Approach

Data

Our approach is relevant generally to multi-class prediction problems in social media. Here we focus on a binary prediction task, specifically the prediction of political preference as captured by the dominant two American political parties: Democratic and Republican. We rely on a dataset previously used for political affiliation classification by (Pennacchiotti and Popescu 2011), then (Zamal, Liu, and Ruths 2012) and (Volkova, Coppersmith, and Van Dume 2014).  

The original data consists of 200 Republican and 200 Democratic users associated with 925 tweets on average per user. Each user has on average 6155 friends with 642 tweets per friend. Sharing restrictions and rate limits on Twitter data collection only allowed us to recreate a subset of that collection. Based on the subset we were able to obtain we formed a balanced collection of 150 Democratic and 150 Republican users. For each user, we randomly sampled 20 friends with 200 tweets per friend.

Models

We assume a set of independent users $U = \{u_i\}$, and neighbors $N = \{n_j\}$, with $N^{(u)}$ the neighbors of $u$. We are concerned with models over data that changes over time: let $\tau$ be an index over discrete time-steps, where at each time-step $\tau$ we observe zero or more tweets from each user, and each user-neighbor, on which we base our predictions. A user is labeled at time $\tau$ if we know the value of the attribute function $A(u) \in \{a_i\}$.

For example, in our experiments we will model the (American) political preference attribute, defined as: $A(u) \in \{R, D\}$, with $R$ standing for Republican and $D$ for Democratic. Let $L_\tau \subseteq U$ be the labeled users at time $\tau$, and $T_\tau = U \setminus L_\tau$ the unlabeled users. Our goal is to predict the attribute value for each user in $T_\tau$ at every $\tau$ given the evidence available up to $\tau$.

Unlike previous models for latent user attribute classification, we: (1) consider updating the initial model learned at $\tau_0$ as new evidence becomes available at $\tau_k$, and (2) reestimate decision probabilities for the unlabeled users given the updated model and new content generated by these users and their neighbors by $\tau_k$.

We define two models $\Phi(u, \tau)$ and $\Phi(n, \tau)$ learned from dynamically growing streams of tweets $T^{(U)}$ and $T^{(N)}$. The user model $\Phi(u, \tau)$ is learned exclusively from user communications to be applied to user tweets $t_1^{(u)}, t_2^{(u)}, \ldots, t_m^{(u)} \in T_\tau^{(u)}$. $\Phi(u, \tau)$ is then a function mapping a user to the most likely attribute value assignment at $\tau$:

$$\Phi(u, \tau) = \arg\max_a P(A(u) = a | T_\tau^{(u)}).$$  

(1)

Neighborhood model $\Phi(n, \tau)$ is learned from neighborhood communications of Democratic and Republican users. It is defined similarly to Eq. 1 and is applied to classify friend tweets within friend or joint user-neighbor stream $t_1^{(n)}, t_2^{(n)}, \ldots, t_m^{(n)} \in T_\tau^{(n)}$.

A user is labeled at time $\tau$ if we predict the value of the attribute function $A(u)$. We apply Bayesian rule updates to dynamically revise posterior probability estimates of the attribute value $P(A(u) = R | T_\tau)$ given a prior e.g., in our case we start with a balanced prior $P(R) = P(D) = 0.5$.

$$P(A(u) = R | T_\tau) = \frac{P(A(u) = R) \cdot P(T_\tau | A(u) = R)}{\sum_{a \in A} P(A(u) = a) \cdot P(T_\tau | A(u) = a)}.$$  

(2)
We will assume tweets to be independent conditioned on attribute, which means our model factors across individual messages $T_{\tau} = (t_1, \ldots, t_m)$, allowing for simple posterior updates on a tweet by tweet basis:

$$
P(A(u) = R | t_1, \ldots, t_m) = \frac{P(A(u) = R) \cdot \prod_m P(t_m | A(u) = R)}{\sum_{a \in A} P(A(u) = a) \cdot \prod_m P(t_m | A(u) = a)}.
$$

The conditional probability of a given tweet is determined by a log-linear model trained on observations from $L_{\tau}$. We show the example updated posterior probabilities for political preference prediction $P(R | t_1, \ldots, t_m)$ in Figure 2.

The final decisions about label assignments can be made at any time $\tau_k$ e.g., if $P(R | t_1, \ldots, t_m) = 0.9$ one can label the user as $R$ with an associated 90% model confidence given the evidence available by $\tau_k$. We analyze the difference in precision and recall by making decisions based on high or low probability assignments using different thresholds $\theta$: 0.55 and 0.95. When $P(A(u) = a | t_1, \ldots, t_m)$ exceeds $\theta$ we make a decision about the label for a user at $\tau_k$.

**Experimental Setup**

We design a set of classification experiments from three types of data streams including user (U), neighbor (N) and user-neighbor (UN). We aim to explore the following prediction settings: Iterative Batch (IB), Iterative Batch with Rationale Filter (IBR), Active without Oracle (AWOO), Active with Oracle (AWO), Active with Rationale Filter (AWR).

For all settings we perform 6-fold cross validation and use a balanced prior: 50 users in the train split and 250 users in the test. For all experiments we use the LibLINEAR package integrated in the JERBOA toolkit (Van Durme 2012a). The log-linear models with dynamic Bayesian updates defined in Eq.1 and Eq.3 are learned using binary word unigram features extracted from user or neighbor content.

**Iterative Batch** We learn tweet-based models at each time stamp $\tau$ from the set of labeled users $L_{\tau}$ and their neighbors e.g., friends. We apply these models using Eq.3 to U, N and UN streams to label all unlabeled users $\hat{L}_{\tau}$ over time. The set of labeled users is constant across all values of $\tau$: we have labels on some users before hand, and no new labels are gathered; only the amount of content available for the users and their neighbors is increasing over time.

**Iterative Batch With Rationale Filtering** Prior work by (Zaidan and Eisner 2008) and (Yessenalina, Choi, and Cardie 2010) explored the utility of asking annotators to choose rationales, explicitly highlighted words or phrases in provided content, that best justified why the annotator made their labeling decision. Our batch setup with rationale filtering is equivalent to the iterative batch setup, except at every time stamp $\tau$ we modify our training data to include tweets with the rationales exclusively. For that, at every $\tau$ we estimate predictive unigrams -- potential rationale words $w \in V$ for Democratic and Republican users in $L_{\tau}$:

$$\forall a \in A(u) \exists \{ w | P(w | A(u) = a) \geq 0.55 \}$$

4Our framework generalizes to non-balanced class priors for train and test, but does assume that the prior is known a priori; estimating class priors in social media is an element of future work.

The conditional probabilities of each word $P(w | A(u) = D)$ and $P(w | A(u) = R)$ are calculated as the empirical estimates over tweets, where $w$ was constrained to have a minimum count of three. We then ask annotators on Mechanical Turk to select rationales from the strongest ranked candidates for $D$ and $R$ by showing them potential rationales and a subset of tweets up to $\tau$. We ask three redundant annotations for each unigram $w$ and take the majority vote to determine if the unigram truly reveals political preferences. For example, the Democratic rationales with $P(w | A(u) = a) > 0.9$ and 100% annotator agreement include: immigrants, students, unemployment and Republican: #tea party, dental, obamacare, arts.

**Active Without Oracle** Unlike our batch setup applied iteratively over a stream of tweets, we propose to update the $\Phi(u, \tau)$ and $\Phi(n, \tau)$ models by moving users from the test set labeled at $\tau_k$ to the training set at $\tau_{k+1}$ as shown in Figure 2. The final decisions about class labels for the unlabeled users are made based on posterior probability estimates $P(A(u) | T_{\tau})$ to exceed the threshold $\theta$. Similarly to the batch setting we experiment with two values of $\theta$ and three data streams: U, N and UN. This bootstrapping approach we refer to as active without oracle (AWOO).

**Active With Oracle** Alternatively, the final label assignments can be judged by an oracle e.g., annotators on Mechanical Turk. For example, we might show $m$ tweets produced by the user by time $\tau$ to one or more annotators. And only if one or more independent annotator judgments agree with $\Phi(u, \tau)$, then we assign a corresponding label to this user at $\tau_k$, and move this user to the training set at $\tau_{k+1}$. Here, since we know the labels we simulate turker judgments (so the oracle is 100% correct). Thus, this setup measures the upper bound for classification. But in the future, we would like to engage real turkers to make class label judgments in the loop. We refer to this setup as active with oracle (AWO) and show it in Figure 3b.

**Active With Rationale Filtering** The rationale filtering step used for IBR setup is also applied to AWOO setup at every $\tau$ as shown in Figure 3a. The difference between batch and active models with rationale filtering is that the potential rationales are estimated on a different set of training data using Eq. 4. In the active case tweets from previously unlabeled users that exceed $\theta$ at $\tau_k$ are added to the tweets of labeled users at $\tau_{k+1}$.

![Active setting with: (a) rationales and (b) oracle. Active setting without oracle (AWOO) is similar to (a) AWR except the rationale filtering step is omitted.](http://www.cs.jhu.edu/~svitlana/rationales.html)
Evaluation

We are concerned with accuracy when operating at different confidence thresholds. Let $\text{Acc}_{\tau,\theta}$ be the accuracy at $\tau$, when considering just users for which the posterior probability exceeds $\theta$. At a given value of $\tau$ and $\theta$, let:

$$\text{Acc}_{\tau,\theta} = \frac{TR + TD}{R + D}, \quad (5)$$

where $TR =$ true Republicans, $TD =$ true Democrats, and $R, D$ are the number of users labeled as Republicans or Democrats, respectively. We abbreviate this as: $\text{Acc}_{\tau,\theta} = C_{\tau,\theta}/A_{\tau,\theta}$, with $C_{\tau,\theta}$ being the number of correctly classified users, and $A_{\tau,\theta}$ being the number of users above a given threshold $\theta$. We also estimate $Q_{\tau,\theta}$ which is the total number of active users who tweeted at least once by $\tau$ (note that $C_{\tau,\theta} \leq A_{\tau,\theta} \leq Q_{\tau,\theta}$). The performance metric $\text{Acc}_{\tau,\theta}$ defined in Eq. 5 can be effectively used for targeted online advertising where one would like to send the advertisements as early as possible to only active users at time $\tau$ for whom labels are assigned with a reasonable confidence $\theta$.

Results

We first confirm that our incrementally batch-trained approach performs as would be expected. In Figure 4 (a - b), consider model U (based only on user tweets): the difference between decision thresholds 0.95 and 0.55 shows a classic precision versus recall tradeoff; at 0.95 less users are classified (x-axis) but at higher precision (y-axis), as compared to 0.55 which instead has higher recall. This pattern repeats for all models U, N and UN, trained and tested with less data (a: Jan - Apr) as well as more data (b: Jan - Sep). With more data (b), performance improves for all scenarios. U is outperformed by N and UN: having access to the content of neighbors improves performance considerably in all cases (affirming the conclusions of (Zamal, Liu, and Ruths 2012) and (Volkova, Coppersmith, and Van Dume 2014)).

Next we contrast those results to AWOO: not only do we retrain the model each month as in batch, but now we bootstrap by taking our most confident (0.95 or 0.55) predictions for users and add them into our labeled set as if their labels were known. We found that our AWOO model yields higher performance than IB model in early months (up to 1-Jul-2011), and insignificantly lower results after that. It happens because in the active setting the model accumulates noisy predictions for some users over time. In contrast, the AWO model does not have this issue and yields consistently better results over time as we show latter. In Figure 5 we present more detailed classification results for batch and active setting for two thresholds 0.55 and 0.95. These results allow us to analyze (a) the threshold and (b) data stream type influence on classification performance as shown below.

Analyzing Threshold Influence

The results in Figures 4 and 5 demonstrate that for higher $\theta$, when the models are more constrained and, therefore, more

---

6 This generalizes standard language of (True) Positive and (True) Negative to allow for non-binary scenarios, such as if adding Libertarian ($L$), Green Party ($G$), etc., to the attribute set: $\text{Acc} = (TR + TD + TG + TL)/(R + D + G + L)$.

7 Many authors don’t tweet often e.g., 85.3% of all Twitter users post less than one update per day. Thus, less tweets are generated by random users by time $\tau$ compared to the number of tweets generated by a set of their friends.
using UN (N) streams earlier at $t_k$ and only “more difficult to classify” users are left to be classified later at $t_{k+1}$.

**Utilizing Oracle Annotations**

In Figure 6 we demonstrate the upper bound for political preference classification performance with $\theta = 0.95$ using our active with oracle (AWO) experimental setup. Similar to other experiments, we report classification performance $\text{Acc}_{\tau, \theta}$ at every $\tau$ with the number of user and neighbor tweets available for training when predictions are made over U and N data streams. We find that $\text{Acc}_{\tau, \theta}$ is monotonically increasing over time and is significantly higher then for IB and AWOO settings. To give a cost estimate of requesting iterative oracle annotations, we outline the number of requests to the oracle aggregated over time in Figure 6 (top).

Active learning with iterative oracle annotations demonstrate the highest performance compared to all other classification settings. For instance, 226 out of 250 users (90%) are correctly classified by June using N stream and 230 (92%) using UN stream using AWO setup compared to 191 (76%) and 203 (81%) users using AWOO setup. Similarly, 112 (45%) users are correctly classified by June using U stream using AWO model compared to 80 (32%) using AWOO.

**Applying Rationale Filtering**

Here we analyze the impact of rationale filtering on prediction performance in batch: IB vs. IBR and active: AWOO vs. AWR settings over time. In Figure 5 we report results for models with and without rationale filtering. As before, we present the results for two thresholds $0.55$ and $0.95$ and three data streams: U, N and UN. For IBR and AWR models with rationale filtering we observe similar precision-recall trends to IB and AWOO shown in Figure 4.

During rationale filtering we select training examples with highly predictive feature norms (a.k.a. rationales) at every $\tau$. This filtering step reduces the number of training examples $L$, vocabulary size $V$ and feature space for both user $\Phi(u, \tau)$ and neighbor $\Phi(n, \tau)$ models over time as shown in Tables 1 and 2. We observe that the size of the training data $L$ is reduced at least in half at every time stamp. Therefore, we consider rationale filtering as a dimensionality reduction step for our batch and active models with incremental Bayesian updates.

Nevertheless, the size of the training data is significantly lower at every $\tau$ the quality of batch and active models trained with filtered data is better for IBR vs. IB and AWR vs. AWOO. In other words, selecting tweets with highly predictive feature norms for training leads to consistent performance improvements over time. We show the empirical results for the relative percentage gain $\Delta \text{Acc}, \%$ for batch and active models with vs. without rationale filtering in Tables 1 and 2, respectively. Models with rationale filtering yield higher precision but lower recall compared to the models without rationale filtering when the predictions are made using N or UN stream. Except we observe higher precision but comparable or higher recall when U stream is used.

To summarize, rationale filtering significantly improves classification accuracy and can be effectively used for attribute prediction that require high precision e.g., prod-

Figure 5: Comparing iterative batch IB, IBR vs. active models AWOO, AWR with and without rationale filtering for political preference prediction with two thresholds $\theta = 0.55$ and $\theta = 0.95$ applied to user, neighbor and user-neighbor communication streams. Starting on 1-Jan-2011, at each time stamp $\tau$ e.g., $\tau_2=1$-Mar-2011, $\ldots$, $\tau_{11}=1$-Dec-2011 we measure $C_{\tau, \theta}$ the # of correctly classified users, $A_{\tau, \theta}$ the # of users above the threshold $\theta$, $Q_{\tau}$ the # of users who tweeted at least once by time $\tau$. Using UN (N) streams earlier at $\tau_k$ and only “more difficult to classify” users are left to be classified later at $\tau_{k+1}$.

Figure 6: Active with oracle (AWO) classification results using U, N and UN streams and $\theta = 0.95$ (U and N stand for thousands of tweets used to train user and neighbor models).
Our approaches for making predictions over dynamically evolving social media streams based on incremental Bayesian online updates can be effectively used in: (1) real-time streaming scenarios for dynamically growing social networks; (2) limited resource training conditions e.g., iterative retraining and active learning (bootstrapping) will allow exploring new understudied attributes e.g., life satisfaction, relationship status for which no or limited labeled data exists; (3) low-resource prediction settings e.g., when no or limited user data is available at any given time, neighbor streams can be used to make predictions about the user; (4) low-cost annotation models that rely on iterative instance (assigning class labels to users) or feature annotations (highlighting predictive words in tweets) via crowdsourcing. Moreover, our batch and active models with iterative rationale filtering help to reduce storage and memory requirements when processing large feature vectors and iterative re-training models for real-world prediction in social media.

### Applications

Our key findings include: I. Active retraining with correctly classified users from test data added to the training data at every time stamp significantly outperforms iterative batch retraining setup. II. Making predictions using a joint user-neighbor or neighbor stream is more effective than using only user stream. III. Models with higher confidence yield higher precision and models with lower confidence yield higher recall for both batch and active setting. IV. Rationale annotation and filtering during iterative retraining leads up to 27.7% relative improvement in iterative batch and 20.6% in active setting. V. Active retraining with oracle annotations yields the highest recall: 85% of test users are correctly classified after the second iteration using a joint user-neighbor stream.

### Related Work

#### Batch Models for Personal Analytics

The vast majority of works on predicting latent user attributes in social media apply supervised batch models trained on large amounts of user generated content except (Burger et al. 2011; Volkova, Coppersmith, and Van Dume 2014). These models are learned using lexical bag-of-word features for classifying user: gender (Rao et al. 2010; Filippova 2012; Ciot, Sonderegger, and Ruths 2013; Bergsma and Van Durme 2013); age (Goswami and Shishodia 2012; Nguyen et al. 2013; Schwartz et al. 2013a); political orientation (Conover et al. 2011a; Penacchiotti and Popescu 2011; Zamal, Liu, and Ruths 2012; Cohen and Ruths 2013); personality (Golbeck et al. 2011; Bachrach et al. 2012; Schwartz et al. 2013b). Some works use unsupervised approaches to learn user demographics in social media including large-scale clustering (Bergsma et al. 2013) and probabilistic graphical models (Eisenstein et al. 2010).

#### Streaming Models for Personal Analytics

Van Durme proposed a streaming model with incremental updates for iteratively predicting user attributes over a stream of communications (Van Durme 2012b). Unlike Van Durme’s approach, our model suggests more straightforward application of incremental tweet-level online Bayesian updates. In addition, we explore batch vs. online retraining for incremental updates of our models. Finally, we take advantage of Twitter network structure and experiment with friend streams in addition to the user stream of communications. Moreover, our active models with iterative oracle and rationale annotations are similar to active learning techniques where the learner is in control of the data used for learning as described in details here (Olsson 2009; Settles 2010; Laws 2012; Settles 2012).

### Summary

We proposed novel approaches for making predictions over dynamically evolving social media streams based on incremental Bayesian online updates. We studied an iterative incremental retraining in batch and active settings with and without iterative oracle annotations. Moreover, we applied interactive feature annotation (rationale) technique as a filter for iterative retraining of the proposed models. Finally, we took advantage of a network structure by making predictions from neighbor and joint user-neighbor streams.
References


