Online Sentiment-based Topic Modeling for Continuous Data Streams

Gopi Chand Nutakki
Knowledge Discovery & Web Mining Lab,
University of Louisville
g0nuta01@louisville.edu

Olfa Nasraoui
Knowledge Discovery & Web Mining Lab,
University of Louisville
olfa.nasraoui@louisville.edu

ABSTRACT

Continuous social text streams, such as tweets, provide a timeline of discussions. Topic modeling techniques such as Latent Dirichlet Allocation (LDA) have been used to extract the topics being discussed on social media streams. Recently, Online LDA has been proposed as a fast alternative for topic extraction, based on on-line stochastic optimization, while sentiment analysis is often used to track the polarity of posts. In this paper, we propose an online technique, integrating Online LDA and sentiment analysis to extract more refined polarity-aware topics within an online learning framework from continuous Twitter streams.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining;
H.3.3 [**Information Search and Retrieval**]: Clustering

Keywords

Online LDA, sentiment analysis, social media, stream data.

1. INTRODUCTION

Among opinion mining tasks, sentiment classification [4] assigns a semantic orientation of a text as positive, negative or neutral. Sentiment classification models trained on one domain might not work well in another domain. Furthermore, in more fine-grained sentiment classification problems, such as involving sentiments on different topics, topic detection and sentiment classification are often performed in a two-stage pipeline process, by first detecting a topic/feature and then assigning a sentiment label to that particular topic [5, 6]. Continuous text streams provide a timeline of topics where new topics are created, while older topics evolve, decay or disappear over time. In this paper, we propose a hybrid technique that integrates sentiment detection with an Online LDA technique to extract refined topics from a continuous stream of tweets.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
ACM 978-1-4503-2622-3/14/06.
http://dx.doi.org/10.1145/2615569.2615666 .

2. BACKGROUND

The Batch Variational Bayes LDA algorithm converges faster than batch collapsed Gibbs sampling LDA, but requires a full pass through the entire corpus[2, 3]. It is therefore not suited to cases where new data is constantly arriving. Hoffman, Blei et al.[3] proposed an Online Variational Inference algorithm that is even faster than the batch version, and where a stochastic optimization algorithm optimizes an objective using noisy estimates of its gradient. Variational inference replaces sampling with optimization in a manner that is analogous to the EM algorithm. Online LDA is an approximate posterior inference algorithm that can analyze massive collections of documents and converges faster [3]. Online LDA can be extended to handle an infinite vocabulary by using a generative process that is identical to LDA, except that instead of being drawn from a finite Dirichlet[7], the topics are drawn from a Dirichlet Process with base distribution G_0 over all possible words. In the LDA framework, topics are associated with documents, and words with topics. In order to model document sentiments, Joint Sentiment Topic modeling (JST) was proposed by Lin and He[4]. JST uses a lexicon based training data set containing positive, negative and neutral scores for each lexicon. JST is effectively a four layer model, where sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics.

3. ONLINE JOINT SENTIMENT BASED TOPIC MODELING (ONLINE JST)

For online variational inference, the posterior over the per-word topic assignments z is parametrized by ϕ, the posterior over the per-document topic weights θ is parametrized by γ, and the posterior over the topic β is parametrized by λ. S is the sentiment category. Symmetric priors are assumed on θ and β. A good setting of the topics λ is one for which the Evidence Lower Bound (ELBO) $\mathcal{L}$ [1] is as high as possible after fitting the per-document variational parameters γ and ϕ with the Expectation step. Let $\gamma(n_t, \lambda)$ and $\phi(n_t, \lambda)$ be the values of γ_t and ϕ_t produced by the E-step. The goal is to set λ to maximize:

$$\mathcal{L}(n, \lambda) \triangleq \sum_t \ell\left(n_t, \gamma\left(n_t, \lambda\right), \phi\left(n_t, \lambda\right), \lambda\right) \quad (1)$$

where $\ell\left(n_t, \gamma_t, \phi_t, \lambda\right)$ is the t^{th} document's contribution to the variational bound. The Online LDA with variational

Algorithm 1 Online LDA, and Joint Sentiment based Topic Modeling Framework (Online JST).

Input: A list of documents D, α, β, sentiment lexicons
Output: Association between D×Topics, and Topics×S.

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$; // *Weight given to* $\tilde{\lambda}$
Initialize λ randomly
for t=0 $\rightarrow \infty$; // For each Tweet **do**
  E step:
  Initialize $\gamma_{tks} = 1$ (constant 1 is arbitrary)
  repeat
    Set $\phi_{twks} \propto \exp\{E_q[\log \theta_{tks}] + E_q[\log \beta_{kws}]\}$
    Set $\gamma_{tks} = \alpha + \sum_w \phi_{twks} n_{tw}$
  until $\frac{1}{KS}\sum_{ks} |change\ in\ \gamma_{tks}| < 0.00001$;
  M step (update after mini-batch to reduce noise):
  Compute $\tilde{\lambda}_{kws} = \eta + D(n_{tw})(\phi_{twks})$
  Set $\lambda = (1-\rho_t)\lambda + \rho_t \tilde{\lambda}$
end

Bayes, with joint sentiment prediction (Online JST), is listed in Algorithm 1. As the t^{th} vector of word counts n_t is observed, for each sentiment category s and topic k, an E step is performed to find locally optimal values of γ_t and ϕ_t, while holding λ fixed. In the true online case, $D \rightarrow \infty$, corresponding to empirical Bayes estimation of β. λ is updated using a weighted average of its previous value and $\tilde{\lambda}$. The weight value for $\tilde{\lambda}$ is given by $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$, where $\kappa \in (0.5, 1]$ controls the rate at which old values of $\tilde{\lambda}$ are forgotten, $\tau_0 \geq 0$ slows down the early iterations of the algorithm and is needed to guarantee convergence.

4. EXPERIMENTS

The experiments were performed on data consisting of over 300,000 tweets collected from Twitter's public API between October 2011 and December 2013, filtered using keyword *obama*. Stop words were retained to preserve the context information for sentiment detection. Figure 1 shows a sample of three topics per sentiment category extracted using the proposed Online JST. A common metric used to evaluate language models [2] is Perplexity, which is given by $perplexity\,(D') = \exp\left\{-\frac{\sum_{d=1}^{T} \ln p\left(\vec{w}^{(d)} | \vec{\alpha}, \beta\right)}{\sum_{d=1}^{T} N_d}\right\}$, for a test set of T documents $D' = \left\{\vec{w}^{(1)}, \cdots, \vec{w}^{(T)}\right\}$ with N_d keywords in the d^{th}document. Since the numerator is the held-out likelihood, a lower perplexity indicates a better generalization performance of the topic model. Figure 2 shows the improvement of the held-out likelihood for the Online JST models, with more iterations. Each iteration integrates a new window[1] of data arriving through the stream. The results show that the model matures as new data arrives.

5. CONCLUSION

The proposed Online JST model provides a faster topic modeling compared to the slower Gibbs sampling based techniques such as JST. Our experiments showed that adding a sentiment layer to the Online LDA technique can produce good quality topics while being faster and handling an infinite stream of tweets.

[1]window: a pool of 100 Tweets, ordered by timestamp.

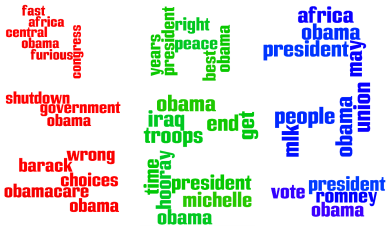

Figure 1: Word clouds of topics with different sentiments, obtained using Online JST. Red, Green and Blue denotes negative, positive, and neutral topics, respectively.

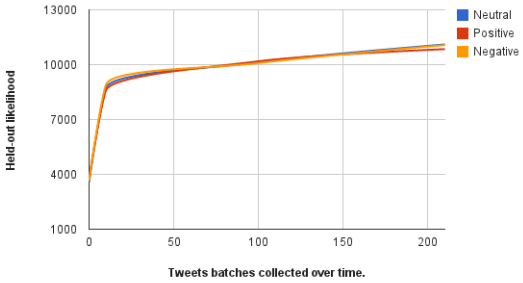

Figure 2: Held-out-likelihood measuring the topic model perplexities for the three sentiment polarities extracted using Online JST. The trend shows the perplexity improvement as new batches of tweets arrive through the stream.

6. REFERENCES

[1] Blei, D. M., and McAuliffe, J. D. Supervised topic models. *arXiv preprint arXiv:1003.0783* (2010).

[2] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[3] Hoffman, M., Blei, D. M., and Bach, F. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems 23* (2010), 856–864.

[4] Lin, C., and He, Y. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 375–384.

[5] Liu, Y., Huang, X., An, A., and Yu, X. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), ACM, pp. 607–614.

[6] Lu, Y., and Zhai, C. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World wide web* (2008), ACM, pp. 121–130.

[7] Zhai, K., and Boyd-Graber, J. Online topic models with infinite vocabulary. In *International Conference on Machine Learning* (2013).