Tracking Sentiment and Topic Dynamics from Social Media

Yulan He Chenghua Lin
Knowledge Media Institute
The Open University, UK
{y.he,c.lin}@open.ac.uk

Wei Gao
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
wgao@qf.org.qa

Kam-Fai Wong
Department of Systems Engineering
and Engineering Management
The Chinese University of Hong Kong
kfwong@se.cuhk.edu.hk

Abstract

We propose a dynamic joint sentiment-topic model (dJST) which allows the detection and tracking of views of current and recurrent interests and shifts in topic and sentiment. Both topic and sentiment dynamics are captured by assuming that the current sentiment-topic specific word distributions are generated according to the word distributions at previous epochs. We derive efficient online inference procedures to sequentially update the model with newly arrived data and show the effectiveness of our proposed model on the Mozilla add-on reviews crawled between 2007 and 2011.

Introduction

Sentiment dynamics from online contents has been shown to have a strong correlation with fluctuations in macroscopic social and economic indicators in the same time period (Bollen, Pepe, and Mao 2010). Sentiment time series extracted from Twitter messages has also been shown strongly correlate with polling data on consumer confidence and political opinion (O'Connor et al. 2010). Nevertheless, most existing work detect sentiment in isolation of topic detection and simply record sentiment in a different time granularity to form a sentiment time series.

In this paper, we propose a dynamic joint sentiment-topic model (dJST) which allows the detection and tracking of views of current and recurrent interests and shifts in topic and sentiment. The dJST model extends from the previously proposed joint sentiment-topic (JST) model which is able to extract coherent and informative topics grouped under different sentiment (Lin and He 2009; Lin et al. 2011). The only supervision required by JST learning is domain-independent polarity word prior information.

The proposal of the dJST model is motivated by two observations. First, the previously proposed JST model assumes that words in text have a static co-occurrence pattern, which may not be suitable for the task of capturing topic and sentiment shifts in a time-variant data corpus. Second, when fitting large-scale data, the standard Gibbs sampling algorithm used in JST can be computationally difficult because it has to repeatedly sample from the posterior the sentiment-topic pair assignment for each word token through the entire corpus each iteration. The time and memory costs of the batch Gibbs sampling procedure therefore scale linearly with the number of documents analysed.

As an online counterpart of JST, the proposed dJST model addresses the above issues and permits discovering and tracking the intimate interplay between sentiment and topic over time from data. To efficiently fit the model to a large corpus, we derive online inference procedures based on a stochastic expectation maximization (EM) algorithm, from which the dJST model can be updated sequentially using the newly arrived data and the parameters of the previously estimated model. Furthermore, to minimize the information loss during the online inference, we assume that the generation of documents in the current epoch is influenced by historical dependencies from the past documents.

We proceed with the proposal of the dynamic JST model and describe its online inference procedures as well as the estimation of evolutionary parameters. We demonstrate the effectiveness of our proposed approach by analyzing both sentiment and topic dynamics from review documents crawled from Mozilla review site. Finally, we conclude our work and outline future directions.

Dynamic JST (dJST) Model

In a time-stamped document collection, we assume documents are sorted in the ascending order of their time stamps. At each epoch t where the time period for an epoch can be set arbitrarily at, e.g. an hour, a day, or a year, a stream of documents $\{d_1^t, \cdots, d_M^t\}$ of variable size M are received with their order of publication time stamps preserved. A document d at epoch t is represented as a vector of word tokens, $\boldsymbol{w}_d^t = (w_{d_1}^t, w_{d_2}^t, \cdots, w_{d_{N_d}}^t)$ where the bold-font variables denote the vectors.

We assume that documents at current epoch are influenced by documents at past. Thus, the current sentiment-topic specific word distributions $\boldsymbol{\varphi}_{l,z}^t$ at epoch t are generated according to the word distributions at previous epochs. In particular, we define an evolutionary matrix of topic z and sentiment label l, $\boldsymbol{E}_{l,z}^t$ where each column is the word distribution of topic z and sentiment label l, $\boldsymbol{\sigma}_{l,z,s}^t$, generated for document streams received in the last S epochs. This is equivalent to the Markovian assumption that the current

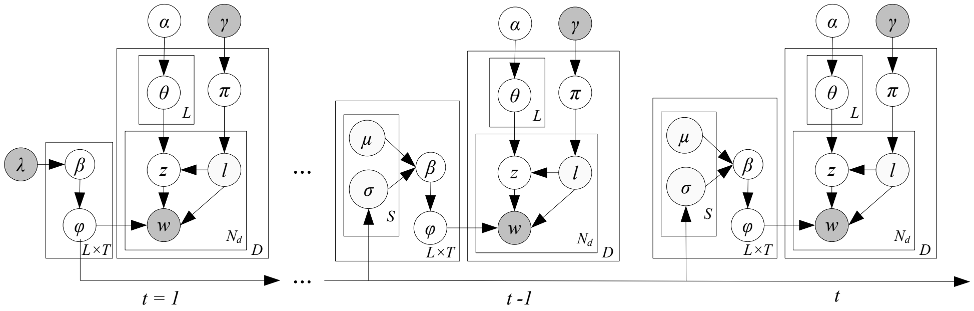


Figure 1: Dynamic JST model.

sentiment-topic-word distributions are dependent on the previous sentiment-topic specific word distributions in the last S epochs.

We then attach a vector of S weights $\boldsymbol{\mu}^t_{l,z} = [\mu^t_{l,z,0}, \mu^t_{l,z,1}, \cdots, \mu^t_{l,z,S}]^T$, each of which determines the contribution of time slice s in computing the priors of $\boldsymbol{\varphi}^t_{l,z}$. Hence, the Dirichlet prior for sentiment-topic-word distributions at epoch t is

$$\boldsymbol{\beta}^t_{l,z} = \boldsymbol{\mu}^t_{l,z} \boldsymbol{E}^t_{l,z} \tag{1}$$

Figure 1 shows the graphical model of dJST and Figure 2 illustrates how the sentiment-topic specific word distribution at current epoch is influenced by the documents in the past epoches. Here the number of historical time slices accounted for is set to 3, $\boldsymbol{\sigma}^t_{l,z,s}, s \in \{1..3\}$ is the historical word distribution of topic z and sentiment label l within the time slice specified by s. We set $\boldsymbol{\sigma}^t_{l,z,0}$ for the current epoch as uniform distribution where each element takes the value of $1/(\text{vocabulary size})$.

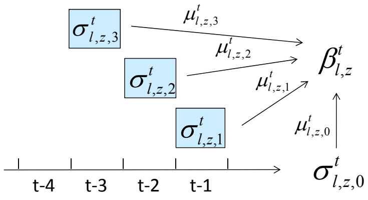


Figure 2: The dJST for $S = 3$. The evolutionary matrix $\boldsymbol{E}^t_{l,z} = [\boldsymbol{\sigma}^t_{l,z,0}, \boldsymbol{\sigma}^t_{l,z,1}, \boldsymbol{\sigma}^t_{l,z,2}, \boldsymbol{\sigma}^t_{l,z,3}]$. The weight matrix $\boldsymbol{\mu}^t_{l,z} = [\mu^t_{l,z,0}, \mu^t_{l,z,1}, \mu^t_{l,z,2}, \mu^t_{l,z,3}]^T$.

Assuming we have already calculated the evolutionary parameters $\{\boldsymbol{E}^t_{l,z}, \boldsymbol{\mu}^t_{l,z}\}$ for the current epoch t, the generative dJST model as shown in Figure 1 at epoch t is given as follows:

- For each sentiment label $l = 1, \cdots, L$
  - For each topic $z = 1, \cdots, T$
    - Compute $\boldsymbol{\beta}^t_{l,z} = \boldsymbol{\mu}^t_{l,z} \boldsymbol{E}^t_{l,z}$
    - Draw $\boldsymbol{\varphi}^t_{l,z} \sim \text{Dir}(\boldsymbol{\beta}^t_{l,z})$.
- For each document $d = 1, \cdots, D^t$
  - Choose a distribution $\boldsymbol{\pi}^t_d \sim \text{Dir}(\gamma)$.
  - For each sentiment label l under document d, choose a distribution $\boldsymbol{\theta}^t_{d,l} \sim \text{Dir}(\boldsymbol{\alpha}^t)$.
  - For each word $n = 1, \cdots, N_d$ in document d
    - Choose a sentiment label $l_n \sim \text{Mult}(\boldsymbol{\pi}^t_d)$,
    - Choose a topic $z_n \sim \text{Mult}(\boldsymbol{\theta}^t_{d,l_n})$,
    - Choose a word $w_n \sim \text{Mult}(\boldsymbol{\varphi}^t_{l_n,z_n})$.

At epoch 1, the Dirichlet priors $\boldsymbol{\beta}$ of size $L \times T \times V$ are first initialized as symmetric priors of 0.01, and then modified by a transformation matrix $\boldsymbol{\lambda}$ of size $L \times V$ which encodes the word prior sentiment information (Lin et al. 2011). For subsequent epochs, if there are any new words encountered, the word prior polarity information will be incorporated in a similar way. But for existing words, their Dirichlet priors for sentiment-topic-word distributions are obtained using Equation 1.

Online Inference

We present a stochastic EM algorithm to sequentially update model parameters at each epoch using the newly obtained document set and the derived evolutionary parameters. At each EM iteration, we infer latent sentiment labels and topics using the collapsed Gibbs sampling and estimate the hyperparameters using maximum likelihood (Lin et al. 2011).

There are two sets of evolutionary parameters to be estimated, the weight parameters $\boldsymbol{\mu}$ and the evolutionary matrix $\boldsymbol{E}$. We assign different weight to each element in $\boldsymbol{E}^t$ by estimating $\boldsymbol{\mu}$ using the fixed-point iteration method (Minka 2003) through maximizing the joint distribution in Equation 2.

$$\begin{aligned} & P(\boldsymbol{W}^t, \boldsymbol{L}^t, \boldsymbol{Z}^t | \gamma^t, \boldsymbol{\alpha}^t, \boldsymbol{E}^t, \boldsymbol{\mu}^t) = \\ & \qquad P(\boldsymbol{L}^t|\gamma^t) P(\boldsymbol{Z}^t|\boldsymbol{L}^t, \boldsymbol{\alpha}^t) P(\boldsymbol{W}^t|\boldsymbol{L}^t, \boldsymbol{Z}^t, \boldsymbol{E}^t, \boldsymbol{\mu}^t) \end{aligned} \tag{2}$$

The update formula is:

$$(\mu^t_{l,z,s})^{\text{new}} \leftarrow \frac{\mu^t_{l,z,s} \sum_w \sigma^t_{l,z,s,w} A}{B}, \tag{3}$$

where $A = \Psi(N^t_{l,z,w} + \sum_{s'} \mu^t_{l,z,s'} \sigma^t_{l,z,s',w}) - \Psi(\sum_{s'} \mu^t_{l,z,s'} \sigma^t_{l,z,s',w})$ and $B = \Psi(N^t_{l,z} + \sum_{s'} \mu^t_{l,z,s'}) -$

$\Psi(\sum_{s'} \mu^t_{l,z,s'})$, $N^t_{l,z,w}$ is the number of times word w was assigned to sentiment label l and topic z at epoch t, $N^t_{l,z} = \sum_w N^t_{l,z,w}$.

The derivation of the evolutionary matrix requires the estimation of each of its elements, $\sigma_{l,z,s,w}$, the word distribution of word w in topic z and sentiment label l at time slice s. This can be defined as follows:

$$\sigma^t_{l,z,s,w} = \frac{N^s_{l,z,w}}{\sum_{w'} N^s_{l,z,w'}} \quad (4)$$

Experiments

We crawled review documents between March 2007 and January 2011 from the Mozilla Add-ons web site[1]. These reviews are about six different add-ons, Adblock Plus, Video DownloadHelper, Firefox Sync, Echofon for Twitter, Fast Dial, and Personas Plus. All text were downcased and non-English characters were removed. We further pre-processed the documents by stop words removal based on a stop words list[2] and stemming. The final dataset contains 9,114 documents, 11,652 unique words, and 158,562 word tokens in total.

The unit epoch was set to quarterly and there were a total of 16 epochs. At the beginning, there were only reviews on Adblock Plus and Video DownloadHelper. Reviews for Fast Dial and Echofon for Twitter started to appear at Epoch 3 and 4 respectively. And reviews on Firefox Sync and Personas Plus only started to appear at Epoch 8. We also notice that there were a significantly high volume of reviews about Fast Dial at Epoch 8. As for other add-ons, reviews on Adblock Plus and Video DownloadHelper peaked at Epoch 6 while reviews on Firefox Sync peaked at Epoch 15. Each review is also accompanied with a user rating in the scale of 1 to 5. The average user rating across all the epochs for Adblock Plus, Video DownloadHelper, and Firefox Sync are 5-star, 4-star, and 2-star respectively. The reviews of the other three add-ons have an average user rating of 3-star.

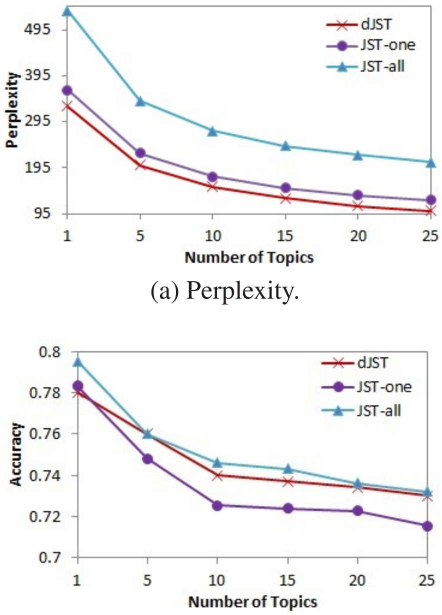
We incorporated word polarity prior information into model learning where polarity words were extracted from the two sentiment lexicons, the MPQA subjectivity lexicon and the appraisal lexicon[3]. These two lexicons contain lexical words whose polarity orientations have been fully specified. We extracted the words with strong positive and negative orientation and performed stemming. Duplicate words and words with contradictory polarities after stemming were removed automatically. The final sentiment lexicon consists of 1,511 positive and 2,542 negative words.

Comparison with Other Models

We compare the performance of dJST models with the non-dynamic version of JST models in terms of perplexity and sentiment classification accuracy. Perplexity measures a model's prediction ability on unseen data. Lower perplexity implies better predictiveness and hence a better model. We fix the number of historical time slices to 4 for dJST. The models we compare against include JST-one and JST-all. JST-one only use the data in the previous epoch for training and hence it does not model dynamics. JST-all uses all past data for model learning.

Figure 3(a) shows the average per-word perplexity over epochs with different number of topics. JST-all has higher perplexities than all the other models. The dJST model outperforms both JST-all and JST-one. Figure 3(b) shows the average document-level sentiment classification accuracy over epochs with different number of topics. dJST has similar performance as JST-one when the topic number is set to 1. Increasing the number of topics leads to a slight drop in accuracy though it stabilises at the topic number 10 and beyond for all the models. dJST outperforms JST-one and performs similarly as JST-all beyond topic number 1.

In conclusion, dJST outperforms JST-one in both predictive perplexity and sentiment classification accuracy, which demonstrates the effectiveness of modelling dynamics. On the other hand, while dJST achieves similar sentiment classification accuracies as JST-all, it has much lower perplexities. More importantly, by avoiding the modelling of all the past documents as JST-all, the computational time of dJST is just in a fraction of JST-all.



(a) Perplexity.

(b) Classification accuracy.

Figure 3: Perplexity and sentiment classification accuracy versus number of topics.

Example Topics

We list in Figure 4 some example topics extracted by dJST with the number of topics set to 10 and the number of historical time slices set to 4. We can observe that the topic-sentiments revealed by the dJST model correlate well with the actual review ratings. At the beginning, the positive sentiment topics were more about Video DownloadHelper (upper panel of Figure 4). Indeed, there are only reviews on

[1] https://addons.mozilla.org/

[2] http://ir.dcs.gla.ac.uk/resources/linguistic˙utils/stop˙words/

[3] http://lingcog.iit.edu/arc/appraisal˙lexicon˙2007b.tar.gz

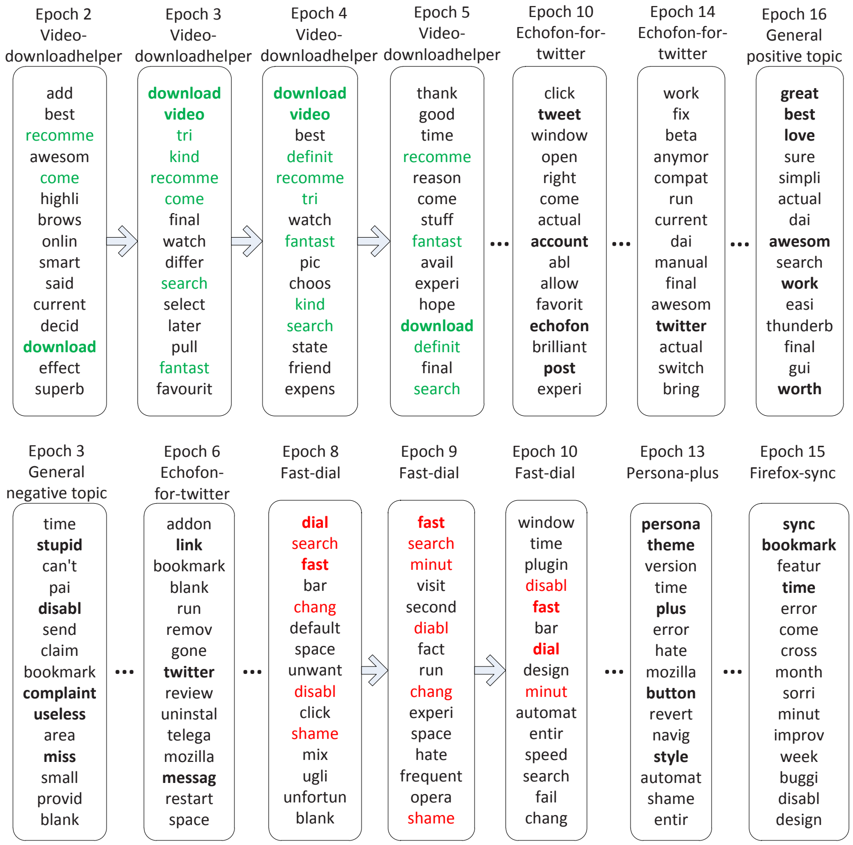


Figure 4: Example topics evolved over time. Topic labels were derived from bold-face words. The upper and lower panels show the topics under positive and negative sentiment respectively. Words that remain the same in consecutive epochs are highlighted in green or red colors.

Adblock Plus or Video DownloadHelper and their average ratings are over 4.5 stars. At Epoch 8, there were a significantly high volume of reviews about Fast Dial and the average rating is about 2 stars. We observe that the negative sentiment topics about Fast Dial start to emerge at Epoch 8 (Lower panel of Figure 4). We also see the positive sentiment topic about Echofon for Twitter at Epoch 10, which aligns with the actual average user rating (over 4 stars) on this add-on.

Conclusions

In this paper, we have proposed the dynamic joint sentiment-topic (dJST) model which models dynamics of both sentiment and topics over time by assuming that the current sentiment-topic specific word distributions are generated according to the word distributions at previous epochs. We demonstrated the effectiveness of dJST on a real-world data set in terms of predictive likelihood and sentiment classification accuracy in comparison to the non-dynamic versions of JST.

Acknowledgements

The authors would like to thank Dong Liu for crawling Mozilla add-ons review data. This work was partially supported by the EC-FP7 project ROBUST (grant number 257859) and the Short Award funded by the Royal Academy of Engineering, UK.

References

Bollen, J.; Pepe, A.; and Mao, H. 2010. *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena*. http://arxiv.org/abs/0911.1583.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *CIKM*, 375–384.

Lin, C.; He, Y.; Everson, R.; and Rueger, S. 2012. Weakly-supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*.

Minka, T. 2003. Estimating a Dirichlet distribution. Technical report.

O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*, 122–129.