A Bayesian Modeling Approach to Multi-Dimensional Sentiment Distributions Prediction

Yulan He
Knowledge Media Institute
The Open University, UK
y.he@open.ac.uk

ABSTRACT

Sentiment analysis has long focused on binary classification of text as either positive or negative. There has been few work on mapping sentiments or emotions into multiple dimensions. This paper studies a Bayesian modeling approach to multi-class sentiment classification and multi-dimensional sentiment distributions prediction. It proposes effective mechanisms to incorporate supervised information such as labeled feature constraints and document-level sentiment distributions derived from the training data into model learning. We have evaluated our approach on the datasets collected from the confession section of the Experience Project website where people share their life experiences and personal stories. Our results show that using the latent representation of the training documents derived from our approach as features to build a maximum entropy classifier outperforms other approaches on multi-class sentiment classification. In the more difficult task of multi-dimensional sentiment distributions prediction, our approach gives superior performance compared to a few competitive baselines.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithms,Experimentation

Keywords

Sentiment analysis, Opinion mining, Joint sentiment/topic model (JST), Latent Dirichlet Allocation (LDA)

1. INTRODUCTION

Sentiment analysis has long focused on binary classification of text as either positive or negative, or projection of text into a one-dimensional scale such as star ratings. Recognizing emotions such as *happiness, anger, sadness, fear, frustration*, etc. from text finds wide applications in human-computer interactions, computer-mediated communication, psychiatric diagnosis, lie detection, etc. Moreover, sentiment analysis that is beyond binary polarity classification offers much greater business insight and usability.

Classifying text into multiple emotion categories can be cast as a multi-class single-label classification problem. Approaches to predicting emotion categories from text include supervised machine learning methods [1, 12], rule-based algorithms [17], knowledge-based methods [30], etc.

Apart from classifying text into multiple emotion categories, there has been few work on mapping text into multi-dimensional sentiment or emotion space. For example, Bollen et al. [5, 4] extracted public mood pattern from Twitter by mapping each tweet message to a six-dimensional mood vector. Tumasjan et al. [31] analyzed tweets published in the weeks leading up to German federal election to predict election results. They mapped tweets into 12 emotional dimensions. Dimensional approaches can produce a much richer representation in emotions. Most importantly, they can naturally capture the smooth change of emotions over time and could thus potentially track trajectories of emotions.

Existing dimensional approaches mostly depend on a pre-built dictionary or lexicon which consists of a list of words corresponding to some emotion dimensions for the mapping of text into multi-dimensional emotion space [5, 4, 31]. Such approaches obviously lack robustness in handling text with unseen words in the dictionary or lexicon. Also, they cannot be used in languages with scarce lexicon resources. More recently, Socher et al. [29] proposed using semi-supervised recursive autoencoders to predict a multi-dimensional distribution over several complex and interconnected sentiments. Nevertheless, training the recursive autoencoders from 5000 documents took about 12 hours until convergence on a 4-core machine.

This paper studies a Bayesian modeling approach to tackle the problem of multi-dimensional sentiment prediction where it aims to map text into multiple-dimensional emotion space with different intensity in each dimension. It derives labeled feature constraints and document-level sentiment distributions from the training data and proposes efficient mecha-

WISDOM '12, August 12 2012, Beijing, China.

nisms to incorporate such supervised information into Bayesian model learning. The proposed approach has been evaluated on the two datasets crawled from the confession section of the Experience Project website. The first dataset, called EPAuthor, contains confessions labeled with authors' self-classified emotion categories. The second dataset, called EPReader, contains confessions labeled with readers' reaction categories. We have performed multi-class single-label sentiment classification on EPAuthor and multi-dimensional sentiment prediction on EPReader. Results obtained from both tasks show that our method outperforms a few competitive baselines and requires only a fraction of training time compared to the previous approach to multi-dimensional sentiment prediction.

We proceed with related work on mapping text into multiple emotion dimensions. Since the Bayesian model studied here is closely related to the Latent Dirichlet Allocation (LDA) model [3], we also review existing approaches of incorporating supervised information into LDA training. We then describe the datasets used in our experiments and present our proposed mechanisms for incorporating labeled feature constraints and document-level sentiment distributions into model learning. Following that, we discuss experimental results on EPAuthor and EPReader. Finally, we conclude the paper.

2. RELATED WORK

There exists a large body of theories of emotion representations. A survey of research on identification of basic emotions can be found in [18]. Ekman et al. [7] defined six basic emotions, *Anger, Disgust, Fear, Joy, Sadness*, and *Surprise*. Parrot [19] started with six primary emotions, *Love, Joy, Surprise, Anger, Sadness*, and *Fear*, and further defined secondary and tertiary emotions for each of them where emotions were categorized into a short tree structure. In research on intelligent tutoring systems (ITS), D'Mello et al. (2007) proposed five categories (*Boredom, Confusion, Delight, Flow*, and *Frustration*) for describing the affect states in students' interactions with ITS.

Classifying text into multiple emotion categories can be cast as a multi-class single-label classification problem. Supervised machine learning methods can be trained from labeled training data to predict sentence-level emotions from children's fairy tales [1] and infer readers' emotions from online news articles [12]. Neviarouskaya et al. [17] developed a rule-based algorithm for analysis of emotion expressed by blog posts at various grammatical levels. Strapparava and Mihalcea [30] proposed and evaluated several knowledge-based and corpus-based methods for the automatic identification of six basic emotions, *anger, disgust, fear, joy, sadness* and *surprise*, from text.

Apart from representing emotions in different categories, there has also been research on plotting emotions along several descriptive axes. Russell [26, 27] proposed to represent emotions in a 2D space, the valence dimension (pleasant vs. unpleasant) and the arousal dimension (relaxed vs. aroused). Mehrabian [14] added a third dominance dimension (dominance vs. submissiveness) to indicate whether the subject feels in control of the situation or not. Appraisal theories of emotion state that emotions result from people's interpretations and explanations of their circumstances [24, 25, 28]. However, it remains a research challenge on how to use the appraisal-based approach for automatic measurement of affect as it requires complex and sophisticated measurements of change.

There has been few work on mapping text into multi-dimensional emotion space. Bollen et al. [5, 4] extracted public mood pattern from Twitter by mapping each tweet message to a six-dimensional mood vector (*Tension, Depression, Anger, Vigour, Fatigue*, and *Confusion*) as defined in the Profile of Mood States (POMS) [13]. Tumasjan et al. [31] analyzed tweets published in the weeks leading up to German federal election to predict election results. They concatenated tweets published over the relevant timeframe into one text sample and mapped into 12 emotional dimensions using the LIWC (Linguistic Inquiry and Word Count) software [20]. More recently, Socher et al. [29] proposed using semi-supervised recursive autoencoders to predict a multi-dimensional distribution over several complex and interconnected sentiments.

Existing dimensional approaches either rely on a pre-built emotion lexicon which consists of a list of words corresponding to some emotion dimensions [5, 4, 31], or require substantial training time to train recursive autoencoders for multi-dimensional sentiment prediction [29]. In this paper, we explore incorporating supervised information extracted from the training data into the learning process of the joint sentiment-topic (JST) model [10, 11], which is extended from the LDA model.

Various approaches have been investigated to incorporate supervised information into LDA model learning. Blei and McAuliffe [2] proposed supervised LDA (sLDA) which uses the empirical topic frequencies as a covariant for a regression on document labels such as movie ratings. Mimno and McCallum [15] proposed a Dirichlet-multinomial regression which uses a log-linear prior on document-topic distributions that is a function of observed features of the document, such as author, publication venue, references, and dates. DiscLDA [9] and Labeled LDA [22] assume the availability of document class labels and utilize a transformation matrix to modify Dirichlet priors. While Labeled LDA simply defines a one-to-one correspondence between LDA's latent topics and observed document labels and hence does not support latent topics within a give document label, Partially Labeled LDA (PLDA) extends Labeled LDA to incorporate per-label latent topics [23]. MedLDA [32], a max-margin supervised topic model, integrates max-margin learning with hierarchical Bayesian topic models by optimizing a single objective function with a set of expected margin constraints.

3. DATASETS

The experience project (EP) dataset was firstly introduced in [21] and was later experimented by Socher et al. [29] using semi-supervised recursive autoencoders for predicting sentiment distributions. The Experience Project (EP) website[1] allows people sharing their life experiences or personal stories anonymously. The EP dataset was crawled from the confessions section of the EP website. Once a confession is

[1] `http://www.experienceproject.com`

Corpus	EPReader	EPAuthor
S	5	4
Senti. dist.	.23/.19/.12/.37/.1	.26/.26/.26/.22
$\parallel \mathcal{D} \parallel$	5,479	9,515
Vocab.	44,696	61,126
Avg. $\parallel d \parallel$	115	81.7

Table 1: Copora statistics. S is the number of sentiment classes or emotion categories. Senti. dist. is the distribution of different classes. $\parallel \mathcal{D} \parallel$ is the total number of confession entries in the dataset. Vocab. is the vocabulary size of the corpus. Avg. $\parallel d \parallel$ is the average number of words per confession entry.

posted to the website, readers can vote in one of the five reaction categories, *you rock*, *teehee*, *I understand*, *sorry, hugs*, and *wow, just wow*. According to the EP website, these correspond to *Inspirational*, *Funny*, *Sympathetic*, *Sad*, and *Angry* confessions.

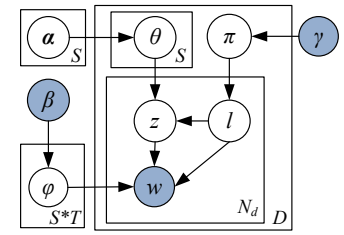
For comparison purposes, we used the same dataset[2] as reported in [29]. The original dataset contains a total of 6,129 URL links to confession entries. However, at the time of downloading, 650 URL links were missing from the EP website. Hence, we were only able to retrieve 5,479 confession entries. Using the original split of training and test sets, we ended up with 3,828 entries for training, and 1,651 entries for testing. The average length of entries is 115 words. Since the emotion categories were derived from the readers' perspectives, we denote this dataset as EPReader. As each confession entry could receive votes from multiple reaction categories, it is essentially labeled with sentiment distributions over the five reaction categories. Hence, EPReader is a dataset where each instance is labeled with a multi-dimensional sentiment distribution.

Apart from enabling readers to label each confession entry with reaction categories, the EP website also allows authors to classify their confessions into one of the 13 categories such as *Embarrassing*, *Family*, *Friend*, *Humor*, *School*, etc. We picked up four emotion-related categories, *Embarrassing*, *Humor*, *Love*, and *Revenge*, and crawled 2,500 confession entries for each of them except for *Revenge* where only 2,150 entries were available in the EP website. We denote this multi-class single-label dataset as EPAuthor. The statistics of the datasets used in our paper is listed in Table 1.

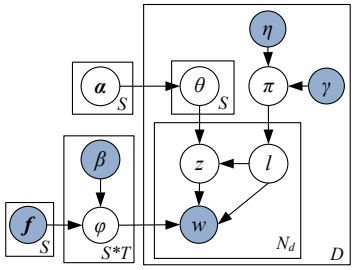
4. INCORPORATING SUPERVISED INFORMATION INTO THE JST MODEL

Assume that we have a corpus with a collection of D documents denoted by $\mathcal{D} = \{d_1, d_2, ..., d_D\}$; each document in the corpus is a sequence of N_d words denoted by $d = (w_1, w_2, ..., w_{N_d})$, and each word in the document is an item from a vocabulary index with V distinct terms denoted by $\{1, 2, ..., V\}$. Also, let S be the number of distinct sentiment labels, and T be the total number of topics. The generative process in the joint sentiment-topic (JST) model which corresponds to the graphical model shown in Figure 1(a) is as follows:

- For each document d, choose a distribution $\pi_d \sim \text{Dir}(\gamma)$.
- For each sentiment label l under document d, choose a distribution $\theta_{d,l} \sim \text{Dir}(\alpha)$.
- For each word w_t in document d
  - choose a sentiment label $l_t \sim \text{Mult}(\pi_d)$,
  - choose a topic $z_t \sim \text{Mult}(\theta_{d,l_t})$,
  - choose a word w_t from $\varphi^{l_t}_{z_t}$, a Multinomial distribution over words conditioned on topic z_t and sentiment label l_t.



(a) JST.



(b) Modified JST.

Figure 1: The Joint Sentiment-Topic (JST) model and the modified JST with supervised information incorporated.

In the multi-dimensional sentiment prediction task here, there are two sets of prior information we could explore to incorporate into the JST model. One is the word-class association probabilities, the other is the sentiment label distribution for each training document. Previous work on utilizing the JST model for sentiment classification employed word prior polarity knowledge extracted from some sentiment lexicons. However, existing sentiment lexicon resources mostly only contain words marked with positive or negative polarities. Hence, they are not suitable in our task of mapping text into much richer emotion dimensions.

4.1 Labeled Features

Assume that we have some labeled features where words are given with their prior sentiment orientation, we could

[2] http://www.socher.org

construct a set of real-valued features of the observation to express some characteristic of the empirical distribution of the training data that should also hold of the model distribution.

$$f_{kj}(w,s) = \sum_{d=1}^{D}\sum_{t=1}^{N_d} \delta(w_{d,t}=k)\delta(s_{d,t}=j) \quad (1)$$

where $\delta(x)$ is an indicator function which takes a value of 1 if x is true, 0 otherwise. Equation 1 calculates how often feature k and sentiment label j co-occur in the corpus.

By adding a normalization term into f_{kj}, we get the predicted label distribution on the feature k, i.e.

$$f_{kj}(w,s) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{N_d} \delta(w_{d,t}=k)\delta(s_{d,t}=j)}{\sum_{d=1}^{D}\sum_{t=1}^{N_d} \delta(w_{d,t}=k)} \quad (2)$$

Hence, $\mathbf{f}(w,s)$ is a matrix of size $K \times S$ where K is the total number of features or constraints used in model learning, and S is the total number of sentiment labels. The kjth entry denotes the expected number of times that feature k is assigned with label j.

In a multi-class single-label dataset, $\mathbf{f}(w,s)$ can be estimated directly from the labeled training data by maximum likelihood and the labeled feature constraints can be further selected according to their predictive power as measured by the information gain of the features with the class labels. However, in a multi-class multi-label dataset where each instance is assigned with label distributions, $\mathbf{f}(w,s)$ is calculated differently taking into account label distributions.

In the EPReader dataset, each instance has votes associated with five reaction categories. Considering each reaction category as a sentiment label, the distribution of label j on the feature k can be calculated by:

$$f_{kj}(w,s) = \frac{\sum_{d=1}^{D}[\text{Votes}_d(j) : k \in d]}{\sum_{d=1}^{D}[\sum_{s=1}^{S}\text{Votes}_d(s) : k \in d]} \quad (3)$$

where the denominator counts the total number of votes for the sentiment label or reaction category j assigned to documents containing feature k, the numerator counts the total number of votes received by documents containing feature k.

The matrix $\mathbf{f}(w,s)$ essentially captures word prior sentiment knowledge and can be used to modify the Dirichlet prior $\boldsymbol{\beta}$ of sentiment-topic-word distributions. We initialize each element of the matrix $\boldsymbol{\beta}$ of size $S \times T \times V$ to 0.01 and then perform element-wise multiplication between $\boldsymbol{\beta}$ and $\mathbf{f}(w,s)$ with the topic dimension ignored.

4.2 Labeled documents

The sentiment distribution of each training instance can be derived directly from the labeled training data. For the multi-class single-label dataset such as EPAuthor, the original single class label for each training instance is converted into a binary vector which takes value 1 for the current class and value 0 for other classes. For the multi-dimensional sentiment dataset such as EPReader, the sentiment distribution is calculated from the votes of the five reaction categories it received. For example, a confession entry with the votes [2, 5, 0, 2, 1] is assigned with the sentiment distribution [0.2, 0.5, 0, 0.2, 0.1].

A straightforward way to extend the JST model to labeled documents is to simply substitute the document-level sentiment distributions $\boldsymbol{\pi}$ with the observed sentiment distributions. If we additionally set the number of topics to 1, then the JST model is reduced to the LDA model with each document's distribution over topics being restricted to the set of observed sentiment labels for that document. This is in fact equivalent to the labeled LDA model [22] where during training, words can only be assigned to the observed sentiment labels in the document. Such a model implies a different generative process where sentiment distribution for each document is observed.

We propose a more principled way to incorporate the observed sentiment distributions into the model by updating the Dirichlet prior, γ, of the document-level sentiment distribution. In the original JST model, γ is a uniform prior and is set to $\gamma = (0.05\times \parallel d \parallel)/S$, where $\parallel d \parallel$ is the average document length, S is the total number of sentiment labels, and the value of 0.05 on average allocates 5% of probability mass for mixing. In our modified model here, a transformation matrix $\boldsymbol{\eta}$ of size $D \times S$ is used to capture the sentiment distributions as soft constraints.

With the transformation matrix $\boldsymbol{\eta}$, the original symmetric Dirichlet prior of the document-level sentiment distribution for document d is replaced by

$$\boldsymbol{\gamma}'_d = [\eta_{d,0} \times \gamma \times S, \eta_{d,1} \times \gamma \times S, ..., \eta_{d,S} \times \gamma \times S]$$

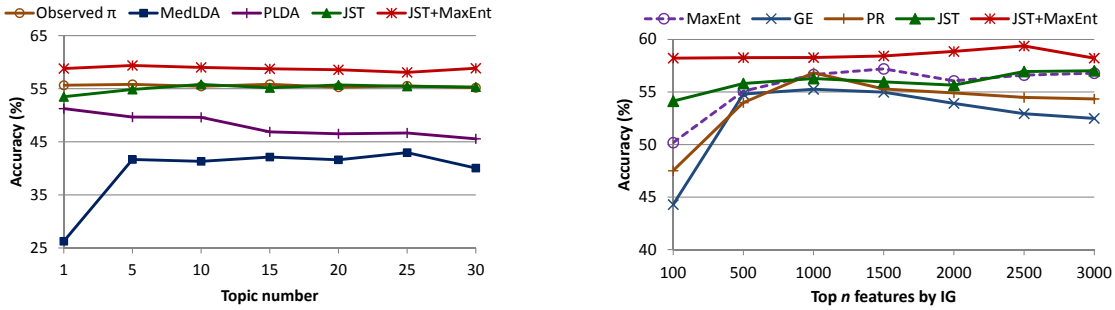
It is worth noting that with the proposed approach, any other side information indicating preference of certain sentiment labels can be incorporated into the model learning in a similar way. For example, in Twitter sentiment analysis, emoticons or hashtags sometimes give indications of the polarity of tweet messages. A tweet message containing emoticons such as ":-)" or ":)" is likely to be positive. In the tweets collected during the UK General Election 2010, the hashtag "#torywin" might represent a positive feeling towards the Tory (Conservative) Party, while "#labourout" could imply a negative opinion about the Labour Party. For such cases, we can initialize each element of $\boldsymbol{\eta}$ to 1. If the side information of a document d is available, then its corresponding vector $\boldsymbol{\eta}_d$ is updated as:

$$\eta_{ds} = \begin{cases} 0.9 & \text{For the inferred sentiment label} \\ \frac{0.1}{S-1} & \text{otherwise} \end{cases},$$

where S is the total number of sentiment classes. Thus, the probability of an instance belonging to the inferred sentiment category is set to a higher value such as 0.9. The remaining probability mass is then equally distributed among the rest sentiment classes.

5. EXPERIMENTS

In our experiments, we used asymmetric prior $\boldsymbol{\alpha}$ over the topic proportions which is learned directly from data using a fixed-point iteration method [16] and updated every 25 iterations during the Gibbs sampling procedure. For other hyperparameters β and γ, we updated them with the



(a) With different topic numbers.

(b) With top n features selected by IG.

Figure 2: Results of multi-class classification accuracy on EPAuthor.

prior knowledge from the labeled feature constraints and the document-level sentiment distributions as discussed in Section 4. It is worth noting that JST without prior information incorporated performs significantly worse than the modified JST model proposed here. Hence, we omit the results from JST. In this section, whenever JST is mentioned, it refers to the modified JST with both the labeled feature and labeled document constraints incorporated.

5.1 EPAuthor

For comparison purposes, we have tested the following baselines:

Supervised classifier. We train the the maximum entropy (MaxEnt) model from MALLET[3] on document vectors with each term weighted according to its frequency[4].

Observed π. Since the document-level sentiment label distributions are given for the training set, we can assume that the multinomial sentiment distribution π is observed instead of being drawn from the Dirichlet prior γ.

MedLDA. We also tested a max-margin supervised topic model, MedLDA [32][5]. In multi-class classification on the 20 Newsgroup dataset, MedLDA was shown to outperform several other supervised LDA models including sLDA and discLDA. For MedLDA, we chose the regularization constant C via 5-fold cross-validation during the training from $1, 4, 9, 16, 25, 36, 49, 64$. We also varied the number of topics from 1 to 100 and chose the one giving the best result.

PLDA. If we assume that the multinomial sentiment distribution π of the training data is observed and we don't incorporate the labeled feature constraints, then our JST model is reduced to Partially-Labeled LDA (PLDA) [23].

Learned from labeled features. The labeled feature constraints can be incorporated into the MaxEnt classifier training with Generalized Expectation (GE) constraints [6] or Posterior Regularization (PR) [8]. In both cases, the training instance labels are ignored. We used the implementation provided in MALLET for our experiments.

Figure 2(a) shows the classification accuracy results using various approaches. For all the results reported here, we performed 5-fold cross validation and averaged over ten different runs. Although MedLDA has previously shown performing well on multi-class classification on the 20 Newsgroup data, it only gave mediocre results here with the best accuracy of 43% obtained at topic number 25. This confirms that multi-class sentiment classification is a much more difficult task compared to topical text classification as sentiment or emotion categories are subtly inter-connected. PLDA gave the best accuracy of 51% at topic number 1. Increasing the number of topics hurts PLDA's performance. It is worth noting that PLDA with topic number 1 is in fact equivalent to Labeled LDA.

Both *Observed* π and the JST with supervised information incorporated perform similarly. However, training MaxEnt from the bag-of-words feature space augmented with the latent sentiment-topics generated from JST (JST+MaxEnt) improves over JST by almost 4%. The JST-based methods perform quite stably with different topic number settings beyond topic number 1.

We then fixed the topic number to 10 for JST and JST+MaxEnt, and performed feature selection by information gain (IG). Figure 2(b) shows the sentiment classification accuracies with different top n features selected by IG where n ranges between 100 and 3000. When the number of features is small (less than 500), GE and PR have quite low accuracies compared to other methods. With the increased number of features, JST performs similarly as the MaxEnt baseline. PR consistently outperforms GE. But they both perform worse than MaxEnt. JST+MaxEnt gives quite stable results regardless of the number of features used and it performs best among all the methods here.

We compare in Table 2 the top 10 words selected by IG and the example topics learned by JST under each of the four sentiment classes. Each JST topic is represented by the top 10 topic words. It can be observed that the latent

[3] `http://mallet.cs.umass.edu/`

[4] We also tested Naïve Bayes and support vector machines. But they perform consistently worse than MaxEnt on average.

[5] `http://www.cs.cmu.edu/~junzhu/medlda.htm`

topics inferred by JST correlate much better with sentiment classes. For example, the topic associated with *Embarrassing* is about embarrassing fat bodies possibly due to alcohol and eating disorders, while the topic under *Love* obviously indicates this specific emotion category. Our proposed method starts with labeled feature constraints extracted by IG and learns latent topics under each of the sentiment class from data. Indeed, as seen from Table 2, the proposed method is able to extract coherent and informative sentiment-bearing topics.

	Embarrassing	*Funny*	*Love*	*Revenge*
IG	work	fuck	good	act
	get	yeah	caus	better
	said	honei	doe	left
	dont	stuffi	turn	blame
	drive	laugh	sure	phone
	go	ha	new	decid
	week	epa	face	month
	mayb	cat	date	us
	put	blue	answer	dog
	children	betcha	final	mother
JST	drink	laugh	feel	hate
	eat	loud	heart	peopl
	food	danc	kiss	life
	bodi	sing	lip	suffer
	smell	song	soul	person
	face	music	bodi	kill
	weight	rock	passion	abus
	alcohol	haha	touch	god
	fat	love	close	help
	disord	listen	sweet	die

Table 2: Words associated with each sentiment class. The upper panel lists the top 10 words selected by information gain. The lower panel lists the top 10 topic words in example latent topics learned by JST.

5.2 EPReader

We evaluate our approach in two aspects. One is to predict the sentiment class with the most votes, the other is to measure the difference between our predicted sentiment distribution with the actual distribution of votes people assign to a confession story. The Kullback-Leibler (KL) divergence has been widely used to measure the difference between two probability distributions. For probability distributions P and Q of a discrete random variable, their KL divergence is defined as $D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$. This measure taking values from 0 to ∞, and $D_{\mathrm{KL}}(P \parallel Q) = 0$ if $P = Q$. The KL divergence, however, has some drawbacks, since it is asymmetric, and it is undefined if $Q = 0$. As such, we use a symmetric version of the KL-divergence, the Jensen-Shannon (JS) divergence, which is defined as:

$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2}(D_{\mathrm{KL}}(P \parallel M) + D_{\mathrm{KL}}(Q \parallel M))$$

where $M = \frac{1}{2}(P+Q)$ is the average of the two distributions. The JS-divergence is bounded between 0 and 1, and is 0 if and only if the two distributions are identical.

In our experiments here, we fixed the number of topics to 10 for JST. Table 3 shows the evaluation results on EPReader

Method	*Accuracy (%)*	*JS-Divergence*
MaxEnt	48.4**	0.350***
MedLDA	40.6***	0.382***
PLDA	49.5**	0.285**
GE	52.7*	0.299**
PR	52.5*	0.301**
Observed π	52.3*	0.285*
Our method	**54.5**	**0.260**

Table 3: Evaluation results on EPReader. Numbers marked with "*"s denote that our method performs statistically significantly better than the baseline models according to a paired t-test with $p < 0.001$ (***), $p < 0.01$ (**), or $p < 0.05$ (*).

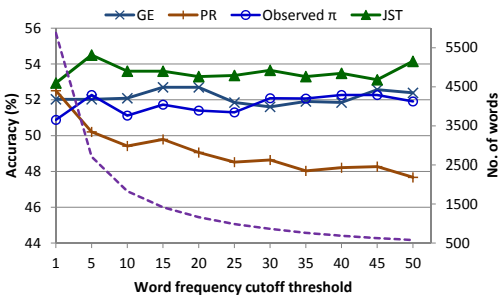
where a similar observation holds here as compared to EPAuthor. Model without the direct incorporation of supervised information (MedLDA) do not perform well. Models learned from the labeled feature constraints only (GE and PR), or learned from the observed document-level labels only (PLDA) improve upon the baseline MaxEnt. Incorporating both the labeled feature constraints and the observed document-level label distribution (*Observed* π) further improves the performance. If both sets of constraints are incorporated as prior information (our method), then we obtain the best results in both predicting the class with most votes and sentiment distributions for testing instances.

We also conducted experiments by varying the labeled feature constraints filtered by the word frequency counts. Results are shown in Figure 3. It can be observed that PR performs the best with all the constraints incorporated. Decreasing the number of constraints hurts the performance of PR. GE outperforms PR and peaked around the word frequency threshold 15. Observed π performs better than GE and PR in JS-divergence. JST gives the best results by only including labeled features which occur more than 5 times in the training data. It gives superior performance than all the other methods.

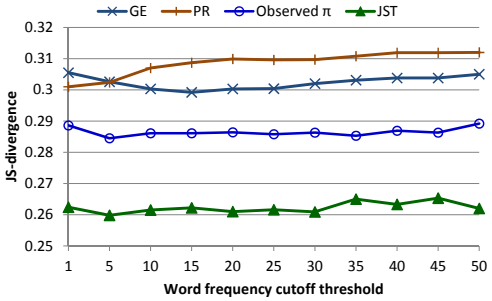
Socher et al. [29] proposed using semi-supervised recursive autoencoders for predicting sentiment distributions. On the EP dataset, they achieved 50.1% accuracy of predicting the class with most votes. Although not directly comparable to their method, our method gave 54.5% accuracy on the subset of their EP dataset. Also, training autoencoders took around 12 hours until convergence on a 4-core machine. Our method only requires the average training time of 12 minutes for 10 topics and 21 minutes for 20 topics.

6. CONCLUSION

This paper has proposed a simple and yet effective mechanism on incorporating supervised information into JST model learning. Experiments have been conducted for multi-class single-label sentiment classification and multi-dimensional sentiment prediction on EPAuthor and EPReader respectively. Existing supervised LDA models such as MedLDA or PLDA do not seem to be effective in both tasks. For multi-class single-label sentiment classification, using the latent representation of the training documents derived from our approach as features to build a MaxEnt classifier outper-



(a) Accuracy.



(b) JS-divergence.

Figure 3: Sentiment prediction results with different word-class constraints filtered by word frequency counts. The dash line in (a) shows the number of labeled feature constraints at different word frequency cutoff. (a) shows the accuracy of predicting the class with most vote; (b) shows the JS-divergence between gold and predicted sentiment distributions. Lower the better.

forms the baselines. For the more difficult task of multi-dimensional sentiment distributions prediction, our approach gives superior performance compared to a few competitive baselines.

Acknowledgments

This work was partially supported by the EPSRC grant EP/J020427/1 and EC-FP7 project ROBUST (grant number 257859).

7. REFERENCES

[1] C. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the HLT-EMNLP*, pages 579–586, 2005.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. Bollen, H. Mao, and A. Pepe. Determining the public mood state by analysis of microblogging posts. In *Proceedings of the Alife XII Conference*, 2010.

[5] J. Bollen, A. Pepe, and H. Mao. *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena*, 2009. Available at http://arxiv.org/abs/0911.1583.

[6] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602, 2008.

[7] P. Ekman, W. V. Friesen, and P. Ellsworth. *Emotion in the human face*, chapter What emotion categories or dimensions can observers judge from facial behavior?, pages 39–55. New York: Cambridge University Press, 1982.

[8] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

[9] S. Lacoste-Julien, F. Sha, and M. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

[10] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM*, 2009.

[11] C. Lin, Y. He, R. Everson, and S. Rueger. Weakly-supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145, 2012.

[12] K. Lin, C. Yang, and H. Chen. Emotion classification of online news articles from the reader's perspective. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 220–226, 2008.

[13] D. McNair, M. Lorr, and L. Droppleman. Profile of mood states (poms). *EdiTS, Educational and Industrial Testing Service*, 1989.

[14] A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[15] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence*, 2008.

[16] T. Minka. Estimating a Dirichlet distribution. Technical report, 2003.

[17] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, pages 278–281, 2009.

[18] A. Ortony and T. Turner. What's basic about basic emotions? *Psychological review*, 97(3):315–331, 1990.

[19] W. Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[20] J. Pennebaker, R. Booth, and M. Francis. *Linguistic inquiry and word count: LIWC 2007*, 2007. Austin, TX: LIWC.net.

[21] C. Potts. On the negativity of negation. In *Proceedings of the 20th Semantics and Linguistic Theory Conference*, volume 20, pages 636–659, 2011.

[22] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256, 2009.

[23] D. Ramage, C. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *KDD*, 2011.

[24] I. Roseman. Cognitive determinants of emotion: A structural theory. *Review of Personality & Social Psychology*, 1984.

[25] I. Roseman and A. Evdokas. Appraisals cause experienced emotions: Experimental evidence. *Cognition and Emotion*, 18(1):1–28, 2004.

[26] J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[27] J. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145–172, 2003.

[28] K. Scherer, E. Dan, and A. Flykt. What determines a feeling's position in affective space? a case for appraisal. *Cognition & Emotion*, 20(1):92–113, 2006.

[29] R. Socher, J. Pennington, E. Huang, A. Ng, and C. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the EMNLP*, 2011.

[30] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, 2008.

[31] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 14th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 178–185, 2010.

[32] J. Zhu, A. Ahmed, and E. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1257–1264, 2009.