Unsupervised Feature-Rich Clustering

Vladimir Eidelman
Dept. of Computer Science and UMIACS, University of Maryland, College Park, MD
vlad@umiacs.umd.edu

ABSTRACT
Unsupervised clustering of documents is challenging because documents can conceivably be divided across multiple dimensions. Motivated by prior work incorporating expressive features into unsupervised generative models, this paper presents an unsupervised model for categorizing textual data which is capable of utilizing arbitrary features over a large context. Utilizing locally normalized log-linear models in the generative process, we offer straightforward extensions to the standard multinomial mixture model that allow us to effectively utilize automatically derived complex linguistic, statistical, and metadata features to influence the learned cluster structure for the desired task. We extensively evaluate and analyze the model's capabilities over four distinct clustering tasks: topic, perspective, sentiment analysis, and Congressional bill survival, and show that this model outperforms strong baselines and state-of-the-art models.

KEYWORDS: Unsupervised Learning, Text Clustering, Sentiment Analysis.

Proceedings of COLING 2012: Posters, pages 275–286,
COLING 2012, Mumbai, December 2012.

1 Introduction

Partitioning documents into categories based on some criterion is an essential research area in language processing and machine learning (Sebastiani, 2002). However, documents are inherently multidimensional, thus a given set of documents can be correctly partitioned along a number of dimensions, depending on the criterion. For instance, given a set of movie reviews, we may be interested in partitioning them by genre, with horror, comedy, drama, etc. in separate categories, or we may want to partition by sentiment, with positive and negative reviews in separate categories. However, it often proves difficult to adapt a model suited for one task, such as topic analysis, to another, such as sentiment analysis.

Supervised generative and discriminative approaches for text classification have achieved remarkable success across a variety of tasks (Joachims, 1998; Kotsiantis, 2007; Pang et al., 2002). Since the partition criterion for a supervised model is encoded in the data via the class labels, even the standard information retrieval representation of a document as a vector of term frequencies is sufficient for many state-of-the-art classification models. Furthermore, for tasks where term presence may not be adequate, discriminative models have the ability to incorporate complex features, allowing them to generalize and adapt to the specific domain.

In unsupervised clustering of documents, we try to partition the documents such that those in one partition are somehow more similar to each other than they are to documents in another partition. Probabilistic clustering models internally assess the quality of clusters via an objective function, $\mathscr{L}(\theta)$, which is commonly maximizing the log-likelihood of generating the data $\mathscr{D}$ under the current parameters of the model, θ. Clustering models rely almost exclusively on a simple bag-of-words vector representation, and therefore achieve an optimum $\mathscr{L}(\theta)$ when grouping documents with similar terms together. This performs well for topic analysis, but, unfortunately, since we do not inherently know the underlying distribution which generated our data, maximizing $\mathscr{L}(\theta)$ is not guaranteed to learn a posterior distribution that performs well for a different task. One method for influencing the objective towards a desired outcome is to include additional feature functions which are able to capture pertinent domain specific information.

Berg-Kirkpatrick et al. (2010) presented an effective framework for learning unsupervised models with expressive feature sets by re-parameterizing every local multinomial in a generative model as a locally normalized log-linear model. They showed that this method allowed them to incorporate arbitrary features of the observation and label pair, and led to competitive performance with more complex models for unsupervised tasks like part-of-speech and grammar induction.

Motivated by their work, we developed a feature-enhanced unsupervised model for clustering in this framework by re-parameterizing the multinomial mixture model. The proposed model, which will serve as our baseline, allows for the integration of arbitrary features of the observations within a document. While in generative models the observed context is usually a single unigram, we extend our re-parametrized baseline model to enable the extraction of features from a context of larger size and incorporate document-level information. After presenting the model, we explore the use of automatically derived linguistic and statistical features, many of which have not been applied to unsupervised clustering. We show that by introducing domain relevant features, we can guide the model towards the task-specific partition we want to learn across four practical tasks with different criterion: topic, perspective, sentiment analysis, and Congressional bill survival. For each task, our feature-enhanced model is highly competitive with or outperforms strong baselines.

2 Related Work

Research on selecting which dimension of the data to cluster can broadly be categorized into approaches which constrain the clustering via external information, and those which cluster along multiple dimensions and then select an appropriate one. Druck (2011) presented a semi-supervised approach that uses domain knowledge in the form of labeled features, which encode affinities between features and classes, to constrain a log-linear model on unlabeled data using generalized expectation criteria (GE-FL). Andrzejewski et al. (2009) and Mimno and McCallum (2008) both attempt to incorporate generalized domain knowledge into generative topic models using priors. The Latent Semantic Model (LSM) (Lin et al., 2010) is a Bayesian model for unsupervised sentiment classification, similar to LDA, but only modeling a mixture of three sentiment labels, positive, negative, and neutral. Another recent approach to guide clustering for sentiment analysis was introduced by Dasgupta and Ng (2009), where they incorporate user feedback into a spectral clustering algorithm (DN). Generalized Weighted Cluster Aggregation (GWCA) (Wang et al., 2009) is a consensus clustering method for topic analysis which utilizes a set of different K-Means clusterings of the same data to construct a similarity matrix, on which spectral clustering is performed to create a single consensus clustering. Iterative Double Clustering (El-Yaniv and Souroujon, 2001) (IDC) is an extension of the Double Clustering approach based on the Information Bottleneck method for topic analysis.

3 Model Description

In our probabilistic generative model for categorizing documents, we assume documents are generated according to a mixture model. The generative process begins by first selecting a class for each document according to the class prior probabilities, θ_j. Each class corresponds to a mixture component, and θ_j are the mixture weights. Next, we generate the contents of the document conditioned on the class according to the class-conditional density, $P_\theta(d_i|c_j)$. Following the Naive Bayes (NB) assumption, we treat all words in a document as conditionally independent given the class, and break $P_\theta(d_i|c_j)$ into its constituent word probabilities θ_{kj}. Under this model, the objective we would like to maximize is the marginal log-likelihood of generating the documents, given by: $\mathscr{L}(\theta) = \sum_{d_i \in \mathscr{D}} \log P_\theta(d_i) = \sum_{d_i \in \mathscr{D}} \log \sum_{c_j \in \mathscr{C}} \theta_j \prod_{w_k \in d_i} \theta_{kj}^{c_{ki}}$ where θ_{kj} is the probability of observing word w_k in class c_j, and c_{ki} is the frequency of w_k in document d_i. Thus, there are two sets of parameters we need to estimate: θ_j for each class and θ_{kj} for each mixture component. The standard instantiation of this model is known as the Multinomial Mixture (MM) model, which is a generalization of the NB classifier for unsupervised learning where θ_{kj} and θ_j are computed using multinomial distributions.

3.1 Unsupervised Feature-Rich (UFR) Model

In order to incorporate features beyond those of term frequency, we can follow the procedure presented in Berg-Kirkpatrick et al. (2010) and re-parameterize the multinomial distribution as a log-linear model based on a feature weight vector $\boldsymbol{\psi}_w$. In this light, θ_{kj} is the output of a locally normalized logistic regression function that scores the word probability according to the active feature functions and weights for that context. Similarly, we can re-parameterize the class prior probability θ_j with a log-linear model with weights $\boldsymbol{\psi}_c$:

$$\theta_{kj}(\boldsymbol{\psi}_w) = \frac{\exp\langle \boldsymbol{\psi}_w, \mathbf{f}(w_k, c_j)\rangle}{\sum_{w_p \in \mathscr{V}} \exp\langle \boldsymbol{\psi}_w, \mathbf{f}(w_p, c_j)\rangle} \quad (1) \qquad \theta_j(\boldsymbol{\psi}_c) = \frac{\exp\langle \boldsymbol{\psi}_{c_j}\rangle}{\sum_{c_m \in \mathscr{C}} \exp\langle \boldsymbol{\psi}_{c_m}\rangle} \quad (2)$$

Combining $\boldsymbol{\psi}_w$ and $\boldsymbol{\psi}_c$ into a single vector $\boldsymbol{\psi}$, the objective function for this model remains the marginal log-likelihood, $\mathscr{L}(\boldsymbol{\psi}) = \sum_{d_i \in \mathscr{D}} \log P_{\boldsymbol{\psi}}(d_i) - \kappa||\boldsymbol{\psi}||^2$, to which we also incorporate a ℓ_2-norm regularization term.

Conveniently, exactly the same generative story as before applies. Thus, optimizing this objective remains straightforward with the Expectation-Maximization (EM) (Dempster et al., 1977) algorithm. The E-step remains the same as the MM model, with the exception that the multinomial probabilities are now being computed with a log-linear model. In the M-step however, instead of simply normalizing, we need to perform an optimization procedure to recompute the weight vector $\boldsymbol{\psi}$ to optimize the complete log-likelihood objective. However, Berg-Kirkpatrick et al. (2010) suggest an alternative method of optimization, the direct gradient approach, which directly optimizes the regularized marginal log-likelihood using L-BFGS (Liu and Nocedal, 1989). The gradient of $\mathscr{L}(\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ has the form:

$$\nabla\mathscr{L}(\boldsymbol{\psi}) = \sum_{w_k\in\mathscr{V}, c_j\in\mathscr{C}} e_{kj}\cdot\Delta_{kj}(\boldsymbol{\psi}) - 2\kappa\cdot\boldsymbol{\psi} \quad (3) \qquad \Delta_{kj}(\boldsymbol{\psi}) = \mathbf{f}(w_k, c_j) - \sum_{w_p\in\mathscr{V}} \theta_{pj}\mathbf{f}(w_p, c_j) \quad (4)$$

3.2 Event Context Expansion

As mentioned earlier, the observation, or event, for most generative models has predominantly been restricted to a single word; the one whose probability is being estimated. Due to the independence assumptions imposed by the naive structure of our UFR model, when computing θ_{kj}, we are only able to look at w_k. So although features can be shared among different observation and label pairs, such as a suffix 'ing' feature activating for both 'going' and 'trying', we are restricted to features of a single word. Thus, without modifying the model, we could not introduce a feature that considered a larger context around w_k, such as w_{k-1} and w_{k+1}. Intuitively, since we want to guide the model towards the partition of the data which we consider relevant for a specific task, it should be beneficial to utilize a larger context than a single word for feature extraction when estimating θ_{kj}. Therefore, we want to weaken the independence assumptions imposed by NB by introducing feature dependence - assuming independence between fewer words - while concurrently taking advantage of the tractable learning and inference that NB offers.

There has been a considerable amount of work in alleviating the independence assumptions of NB model by explicitly representing dependencies between attributes (i.e. words in our case), such as Lazy Bayesian Rules and Tree-Augmented NB (Friedman et al., 1997; Zheng and Webb, 2000). These approaches can be generally characterized as utilizing a less restrictive set of assumptions. First, they select a set of words $b \in \mathscr{N}_x(w_k)$ and then, w_k is allowed to depend on the words in b; such that $\theta_{kj} \rightarrow \theta_{kj|b} = p(w_k|c_j, b)$.

Our proposed extension to UFR, E-UFR, is similar in spirit to these approaches, as we will let each observation encompass the set of surrounding context words. At each position k in the document, instead of generating a single word event, w_k, according to θ_{kj}, we propose generating the entire context as the event, according to $\theta_{b_{k-r}^{k+q}j}$. Here, $b_{k-r}^{k+q} \in \mathscr{N}_{r+q}(w_k)$ is the context: the set of words centered at and including w_k, going q positions forward, and r positions back, and $\mathscr{N}_{r+q}(w_k)$ is the set of all possible contexts of size $(r+q)$ for all k. In another light, instead of having a single θ_{kj}, we now have a θ_{kj} for every different context of w_k. Since we now generate w_k along with its context, we modify the log-linear model from Eq. 1 to Eq. 5 and marginalize over the contexts, enabling feature extraction from its entirety. This will allow features to be active for more observations, thus tying more probability estimates together.

$$\theta_{b_{k-r}^{k+q}j}(\boldsymbol{\psi}_w) = \frac{\exp\langle\boldsymbol{\psi}_w, \mathbf{f}(b_{k-r}^{k+q}, c_j)\rangle}{\sum_{b_p\in\mathscr{N}_{r+q}(w_k)}\exp\langle\boldsymbol{\psi}_w, \mathbf{f}(b_{p-r}^{p+q}, c_j)\rangle} \quad (5)$$

Table 1 shows an example of context generation. Crucially, the context is not treated as a bag-of-words, and by preserving word order, we are able to extract linguistic features that depend on structure. This method of computing $\theta_{b_{k-r}^{k+q}j}$ can be viewed as a form of contrastive estimation (Smith and Eisner, 2005), where we condition the probability on $\mathcal{N}(w_k)$, the neighborhood of possible contexts. In practice, to make parameter estimation tractable for increased context size, we restrict $\mathcal{N}(w_k)$ to observed contexts.

The United States is failing in its mission to implement the roadmap
the united states, the **united** states is, the united **states** is failing, united states **is** failing in, states is **failing** in its, is failing **in** its mission, failing in **its** mission to, in its **mission** to implement, its mission **to** implement the, mission to **implement** the roadmap

Table 1: Contexts generated when producing the sentence above with a 5-word context; r=q=2. Bold indicates the w_k being generated, with surrounding context available for feature extraction.

4 Experiments

To measure the effectiveness of the E-UFR clustering model, we applied it to text corpora with known labels used in supervised classification. Specifically, to topic, perspective, and sentiment analysis, as well as Congressional bill survival. The details of the datasets are summarized in Table 2. All data is preprocessed by performing tokenization, downcasing, and removing non alpha-numeric characters, and stopwords, unless otherwise noted. We compare E-UFR performance on each task with three baselines, UFR, MM and LDA, and where applicable, results taken from related work. The UFR and E-UFR baseline models incorporate only word indicator features, making their feature set identical to the MM model. As the observation context in the E-UFR model, we utilize a 5-word context with q=2 and r=2. The θ_{kj} parameters in the MM model are initialized with uniform MAP estimates across classes from the data, all weights in $\boldsymbol{\psi}$ are initialized to 0, and θ_j is slightly perturbed using a random seed in both cases to allow for learning. To evaluate the accuracy of our approach we compute the cluster purity (Zhao and Karypis, 2002). Since each document can only be assigned one label, and we have the same number of clusters as classes, the measure is directly comparable with micro-averaged precision, accuracy, and F1 (Xue and Zhou, 2009; Bekkerman et al., 2006). All results reported are averaged over 5 runs. Results in bold are statistically significant improvements over the other models and indistinguishable from each other at the p <0.05 level, according to the p-test (Yang and Liu, 1999).

4.1 Topic Analysis

For topic analysis, we use several subsets of the 20-Newsgroup (NG20) (Lang, 1995), and WebKB (Craven et al., 1998) datasets. The NG20 corpus consists of messages posted to various Usenet newsgroups, of which we utilize the Politics, Sport, and Computer splits. The WebKB corpus consists of web pages from university computer science department websites, and has a skewed distribution of examples from each class. We use the WebKB4 split. We present two methods of introducing automatically derived features from LDA. In the first, LDA-A, we introduce a feature representing the per-word topic assignment for every term in the document. In the second, LDA-K, for each topic t_i, we sort terms w_k by $P(w_k|t_i)$ and introduce features for the top 100 terms. For example, given a context $generate_{14}\ a_7\ larger_7\ set_3\ of_7\ data_{18}$ with subscripts representing the per-word topic assignments, possible features are $f(w$=data,t=18) or $f(\#(t$=7)=3). We also incorporate linguistic features in the form of part-of-speech (POS) tags in the same manner, produced using a latent-variable POS tagger (Huang et al., 2009).

The results are presented in Table 3. On the NG20 set, the MM and UFR models exhibit strong performance, mostly outperforming the E-UFR model. With the addition of LDA-A features, however, the E-UFR becomes highly competitive. On WebKB4, the baseline E-UFR model is significantly better than the others. The introduction of LDA features does not enhance its performance, however, POS features reduce the error by 10% over the baseline. Also note, that in comparison to GE-FL, which is semi-supervised and uses LDA features, we achieve better performance. Interestingly, across all the sets, introducing either form of LDA feature results in significantly higher accuracies for the E-UFR model than the original LDA model from which the features are derived. In addition, the LDA-A features always outperform LDA-K.

Set	Task	Docs	Words
WebKB(4)	To	4199	1.3m
Pol(3)	To	2625	1.4m
Sprt(2)	To	1993	670k
Comp(2)	To	1943	480k
Mov(2)	Se	2000	1.5m
BL(2)	Pe	594	510k
Bills(2)	Su	1000	2.5m

Table 2: Description of datasets for Topic (To), Sentiment (Se), Perspective (Pe) analysis and Congressional bill survival (Su) tasks.

Model	Pol	Sprt	Comp	WebKB
MM	69.7	**98**	**83.9**	68.1
LDA	77.5	89.1	72.8	64.8
IDC	78	89	-	-
GWCA	-	-	-	67
UFR	71	**97.4**	69.2	60.6
GE-FL	-	91.5	81.7	61.5
E-UFR	69.3	93.9	63.4	71.2
+LDA-A	**84.1**	96.7	**82.7**	70.7
+LDA-K	77.3	95.7	76.3	68.3
+POS				**74.5**

Table 3: Results on Politics, Sport, Computer newsgroups and WebKB. Table cells marked with "-" for models from related work indicate result for that setting was not available in the literature for that model.

4.2 Perspective Analysis

The BitterLemons corpus Lin et al. (2006) is comprised of essays representing contrasting perspectives on the Israeli-Palestinian conflict, written by Editors and Guests. There are two clear partitions in this data. The first, IP, commonly applied and referred to as determining implicit sentiment, is the task of determining whether a document is written from the Israeli or Palestinian perspective. The second, EG, is whether the author of the article is a permanent Editor or Guest[1]. We extract complex linguistic information, in the form of OPUS (observable proxies for underlying semantics) features, which were shown to improve performance for supervised classification. OPUS features are meant to address implicit sentiment by focusing on syntactic framing in the form of grammatically relevant semantic properties (Greene and Resnik, 2009). We extracted these relations for a set of domain relevant verbs from parses of the corpus obtained with the Stanford parser (Klein and Manning, 2003). For example, sample features from the context *officially endorse the creation* would include f (w=endorse,*transitive*), f (dobj, w=creation), and f (w=endorse,dobj). Table 4 presents the results on these two tasks. As can be seen, the high performance of the UFR and MM models on topic analysis does not carry to the perspective task. The E-UFR model, on the other hand, achieves very impressive results on both tasks. Although the results are not directly comparable to supervised classifiers due to the training-test split, it is interesting to note that our unsupervised results are competitive with those of supervised classifiers on IP (Greene and Resnik, 2009). Unfortunately, the gain from OPUS features did not transfer to clustering. On the other hand, the fact that the performance

[1]As we are interested in differences in author writing style, we did not remove stopwords for this task.

did not degrade with the introduction is itself enticing, as the model was able to incorporate many complex linguistic features and not become obstructed by them. We further explored the use of POS information in EG, which led to a slight improvement. Table 5 presents the most highly weighted OPUS features.

Model	IP	EG
MM	51.4	55.1
LDA	54.4	62
UFR	51.1	52.3
E-UFR	**90.4**	**69.4**
+OPUS	**90.4**	**68.6**
+POS		**70.2**

Table 4: Results on IP and EG split of the BitterLemons dataset.

Weight	Feature
0.594	dobj(abandoned,n)/0
0.582	dobj(oppose,initiative)/0
0.574	subj(accept,israel)/1
0.525	dobj-failure/0
0.488	maintaining-subj/1
0.482	dobj-initiative/0
0.477	dobj(confront,them)/0

Table 5: Top OPUS features/class for IP split. Palestinian perspective class is 0, Israeli perspective is 1.

4.3 Sentiment Analysis

For sentiment analysis we use the Polarity v2.0 dataset (Pang and Lee, 2004), where we cluster movie reviews as negative or positive. We utilize the MPQA subjectivity lexicon (Wiebe and Cardie, 2005), where words which occur in the lexicon are associated with their prescribed polarity. For instance, *result is tepid and dull* would produce f (w=dull,neg) and f (w=tepid,neg), as well as total counts of negative and positive polarity carrying words. The results are presented in Table 6. As can be seen, the baseline UFR model is quite bad, but E-UFR outperforms MM, LDA, and LSM, and is comparable to DN, which uses user interaction. Incorporating the subjectivity lexicon provides a further significant gain. Table 7 presents the most highly weighted sentiment lexicon features. Examining the reviews alongside the lexicon, we noticed that terms that may generally be considered to convey a certain sentiment are inaccurate in their correlation with this domain. For instance, "war" is considered negative, but positive reviews are almost three times as likely to mention it. Thus, we created an alternative version of the lexicon, SUBJR, where we automatically filtered the lexicon to only include domain relevant terms. Impressively, the accuracy achieved with SUBJR is competitive with supervised approaches on this task.

Model	Movie
MM	68.1
LDA	66.6
UFR	51.1
DN	70.9
LSM	61.7
+MPQA	**74.1**
E-UFR	70.5
+MPQA	**72.4**
+SUBJR	79.7

Table 6: Results on Movie Review dataset.

Weight	Feature
0.2077	(pos,great)/1
0.168	(pos,love)/0
0.121	(neg,waste)/0
0.108	(neg,dull)/0
0.105	(neg,bland)/0
0.101	(pos,master/1
0.093	(neg,emotional)/1

Table 7: Top polarity features/class for Movie collection. Positive polarity class is 1, negative is 0.

4.4 Congressional Bill Survival

The recently introduced Congressional Bill Corpus (Yano et al., 2012) contains Congressional bills from the 103^{rd} to 111^{th} Congresses. The task is to predict whether a bill survived, i.e., was recommended by the Congressional committee, or died in committee. We randomly selected

1000 bills from the collection to evaluate our model. While features for the previous tasks are extracted from the content, for Congressional bill survival we incorporate document-level information, both from observable metadata and automatic predictions. The feature set is the one presented in Yano et al. (2012), and includes observable information about the bill (when it was proposed), the bill's sponsor (their party, etc.), the committee (is the sponsor on the committee, etc.), and automatically predicted urgency (trivial, recurring, and critical). Interestingly, our model replicates the results found in the supervised setting, where they found that the sponsor affiliations have the highest impact scores (Yano et al., 2012). The second set, *Spon*, is restricted to the highest weighted observable features describing the bill sponsor, namely, if the sponsor is on the committee and/or in the majority party. The restricted *Spon* set further outperforms all other models.

Model	Bills
MM	58.2
LDA	52.7
UFR	56.2
E-UFR	54.9
+All	60.4
+Spon	**64.1**

Table 8: Results on Congressional bill survival dataset.

Weight	Feature
2.051	sponsor-in-committee-majority/1
1.516	bill-cat4-function-CQ2-00/0
1.478	bill-cat4-function-RECUR-00/0
1.064	sponsor-in-committee/1
1.056	sponsor-in-majority/1

Table 9: Top features/class for Congressional bill survival. Bills which survived are class 1, those that died are class 0. bill-cat features indicate that the bill is not in the category of bills classified as CQ (critical) or RECUR (recurring).

5 Discussion

The results show that the E-UFR model is able to achieve strong performance across the four tasks. We believe this is due both to the increased context and additional features that can be leveraged. Both POS and LDA are a form of dimensionality reduction which can be viewed as categorizing words into distributional categories. As such, using them as features in our model allows us to incorporate information about a possible partition of the data. Since LDA is geared toward discovering topics, LDA features guide the E-UFR model into the correct space. Likewise, POS features assist with authorship because they relate to writing style. Extrapolating from this, any previous clustering of the data can be used as features within our model. In this work, we focused on using unsupervised learning to predict a certain externally imposed partition on the data. However, unsupervised learning is also useful as an exploratory technique for describing a document collection. In this setup, we can incorporate various features in our model to determine not whether they lead to a better accuracy, but what dimensions of the data we can discover. Previous studies on the use of linguistic features for supervised text classification have achieved mostly negative results (Moschitti and Basili, 2004), oftentimes finding that linguistic features do not improve classification accuracy. However, to the best of our knowledge no such analysis exists for the unsupervised treatment of text categorization. In this work, we have shown that linguistic features can be useful for clustering, while questions remain as to how best to incorporate these features.

6 Conclusion

We presented a feature-rich generative model for clustering. By extending the model to handle a wider context, we were able to utilize a rich set of automatically derived linguistic and statistical features, many of which have previously only been explored in supervised learning. We extensively analyzed and evaluated this model, showing that it is stable with respect to many arbitrary features. Applying the model to several challenging categorization domains, we showed that our model is able to adapt and achieve high clustering performance.

Acknowledgments

We would like to thank Chris Dyer, Zhongqiang Huang, and Philip Resnik for helpful comments and suggestions. This research was supported by a National Defense Science and Engineering Graduate Fellowship. Any opinions, findings, conclusions, or recommendations expressed are the author's and do not necessarily reflect those of the sponsors.

References

Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32.

Bekkerman, R., Eguchi, K., and Allan, J. (2006). Unsupervised Non-topical Classification of Documents. Technical report.

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless Unsupervised Learning with Features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. (1998). Learning to Extract Symbolic Knowledge from the World Wide Web. In *Proceedings of the fifteenth national/tenth conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*.

Dasgupta, S. and Ng, V. (2009). Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Dempster, A. P., Laird, M. N., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22.

Druck, G. (2011). Generalized Expectation Criteria for Lightly Supervised Learning. In *Open Access Dissertations*.

El-Yaniv, R. and Souroujon, O. (2001). Iterative Double Clustering for Unsupervised and Semi-supervised Learning. In *Proceedings of the 12th European Conference on Machine Learning*.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Mach. Learn.*, 29(2-3):131–163.

Greene, S. and Resnik, P. (2009). More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Huang, Z., Eidelman, V., and Harper, M. (2009). Improving A Simple Bigram HMM Part-of-Speech Tagger by Latent Annotation and Self-Training. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Nédellec, C. and Rouveirol, C., editors, *European Conference on Machine Learning*, pages 137–142, Berlin. Springer.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational LinguisticsS*, pages 423–430.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249268.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Lin, C., He, Y., and Everson, R. (2010). A Comparative Study of Bayesian Models for Unsupervised Sentiment Detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*.

Lin, W. H., Wilson, T., Wiebe, J., and Hauptmann, A. (2006). Which Side are You on? Identifying Perspectives at the Document and Sentence Levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Liu, D. C. and Nocedal, J. (1989). On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45:503–528.

Mimno, D. M. and McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, pages 411–418.

Moschitti, A. and Basili, R. (2004). Complex Linguistic Features for Text Classification: A Comprehensive Study. In *Proceedings of the 26th European Conference on Information Retrieval*.

Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, 34:1–47.

Smith, N. A. and Eisner, J. (2005). Contrastive estimation: training log-linear models on unlabeled data. In *Proceedings of ACL*, pages 354–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, F., Wang, X., and Li, T. (2009). Generalized Cluster Aggregation. In *Proceedings of the 21st international jont conference on Artifical intelligence*.

Wiebe, J. and Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation*.

Xue, X.-B. and Zhou, Z.-H. (2009). Distributional Features for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, 21.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the ACM SIGIR*, pages 42–49, New York, NY, USA. ACM.

Yano, T., Smith, N. A., and Wilkerson, J. D. (2012). Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 793–802, Montréal, Canada. Association for Computational Linguistics.

Zhao, Y. and Karypis, G. (2002). Criterion Functions for Document Clustering: Experiments and Analysis. Technical report.

Zheng, Z. and Webb, G. I. (2000). Lazy learning of bayesian rules. *Machine Learning*, 41(1):53–84.