

6-2006

Which Side are You on? Identifying Perspectives at the Document and Sentence Levels

Wei-Hao Lin
Carnegie Mellon University

Theresa Wilson
University of Pittsburgh

Janyce Wiebe
University of Pittsburgh

Alexander Hauptmann
Carnegie Mellon University, alex@cs.cmu.edu

Follow this and additional works at: <http://repository.cmu.edu/compsci>

Published In

The Tenth Conference on Natural Language Learning (CoNLL-X), 109- 116.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

In *Proceedings of the Tenth Conference on Natural Language Learning (CoNLL'06)*
Which Side are You on? Identifying Perspectives at the Document and Sentence Levels

Wei-Hao Lin Language Technologies Institute Carnegie Mellon University Pittsburgh, PA 15213 whlin@cs.cmu.edu	Theresa Wilson, Janyce Wiebe Intelligent Systems Program University of Pittsburgh Pittsburgh, PA 15260 {twilson,wiebe}@cs.pitt.edu	Alexander Hauptmann School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 alex@cs.cmu.edu
---	---	---

Abstract

In this paper we investigate a new problem of identifying the *perspective* from which a document is written. By perspective we mean a point of view, for example, from the perspective of Democrats or Republicans. Can computers learn to identify the perspective of a document? Not every sentence is written strongly from a perspective. Can computers learn to identify which sentences strongly convey a particular perspective? We develop statistical models to capture how perspectives are expressed at the document and sentence levels, and evaluate the proposed models on articles about the Israeli-Palestinian conflict. The results show that the proposed models successfully learn how perspectives are reflected in word usage and can identify the perspective of a document with high accuracy.

1 Introduction

In this paper we investigate a new problem of automatically identifying the *perspective* from which a document is written. By perspective we mean a “subjective evaluation of relative significance, a point-of-view.”¹ For example, documents about the Palestinian-Israeli conflict may appear to be about the same topic but reveal different perspectives:

¹The American Heritage Dictionary of the English Language, 4th ed.

- (1) The inadvertent killing by Israeli forces of Palestinian civilians – usually in the course of shooting at Palestinian terrorists – is considered no different at the moral and ethical level than the deliberate targeting of Israeli civilians by Palestinian suicide bombers.
- (2) In the first weeks of the Intifada, for example, Palestinian public protests and civilian demonstrations were answered brutally by Israel, which killed tens of unarmed protesters.

Example 1 is written from an Israeli perspective; Example 2 is written from a Palestinian perspective. Anyone knowledgeable about the issues of the Israeli-Palestinian conflict can easily identify the perspectives from which the above examples were written. However, can computers learn to identify the perspective of a document given a training corpus?

When an issue is discussed from different perspectives, not every sentence strongly reflects the perspective of the author. For example, the following sentences were written by a Palestinian and an Israeli.

- (3) The Rhodes agreements of 1949 set them as the ceasefire lines between Israel and the Arab states.
- (4) The green line was drawn up at the Rhodes Armistice talks in 1948-49.

Examples 3 and 4 both factually introduce the background of the issue of the “green line” without expressing explicit perspectives. Can we develop a

system to automatically discriminate between sentences that strongly indicate a perspective and sentences that only reflect shared background information?

A system that can automatically identify the perspective from which a document is written will be a valuable tool for people analyzing huge collections of documents from different perspectives. Political analysts regularly monitor the positions that countries take on international and domestic issues. Media analysts frequently survey broadcast news, newspapers, and weblogs for differing viewpoints. Without the assistance of computers, analysts have no choice but to read each document in order to identify those from a perspective of interest, which is extremely time-consuming. What these analysts need is to find strong statements from different perspectives and to ignore statements that reflect little or no perspective.

In this paper we approach the problem of learning individual perspectives in a statistical framework. We develop statistical models to learn how perspectives are reflected in word usage, and we treat the problem of identifying perspectives as a classification task. Although our corpus contains document-level perspective annotations, it lacks sentence-level annotations, creating a challenge for learning the perspective of sentences. We propose a novel statistical model to overcome this problem. The experimental results show that the proposed statistical models can successfully identify the perspective from which a document is written with high accuracy.

2 Related Work

Identifying the perspective from which a document is written is a subtask in the growing area of automatic opinion recognition and extraction. Subjective language is used to express opinions, emotions, and sentiments. So far, research in automatic opinion recognition has primarily addressed learning subjective language (Wiebe et al., 2004; Riloff et al., 2003), identifying opinionated documents (Yu and Hatzivassiloglou, 2003) and sentences (Yu and Hatzivassiloglou, 2003; Riloff et al., 2003), and discriminating between positive and negative language (Pang et al., 2002; Morinaga et al., 2002; Yu and

Hatzivassiloglou, 2003; Turney and Littman, 2003; Dave et al., 2003; Nasukawa and Yi, 2003; Popescu and Etzioni, 2005; Wilson et al., 2005). While by its very nature we expect much of the language that is used when presenting a perspective or point-of-view to be subjective, labeling a document or a sentence as subjective is not enough to identify the perspective from which it is written. Moreover, the ideology and beliefs authors possess are often expressed in ways other than positive or negative language toward specific targets.

Research on the automatic classification of movie or product reviews as positive or negative (e.g., (Pang et al., 2002; Morinaga et al., 2002; Turney and Littman, 2003; Nasukawa and Yi, 2003; Mullen and Collier, 2004; Beineke et al., 2004; Hu and Liu, 2004)) is perhaps the most similar to our work. As with review classification, we treat perspective identification as a document-level classification task, discriminating, in a sense, between different types of opinions. However, there is a key difference. A positive or negative opinion toward a particular movie or product is fundamentally different from an overall perspective. One’s opinion will change from movie to movie, whereas one’s perspective can be seen as more static, often underpinned by one’s ideology or beliefs about the world.

There has been research in discourse analysis that examines how different perspectives are expressed in political discourse (van Dijk, 1988; Pan et al., 1999; Geis, 1987). Although their research may have some similar goals, they do not take a computational approach to analyzing large collections of documents. To the best of our knowledge, our approach to automatically identifying perspectives in discourse is unique.

3 Corpus

Our corpus consists of articles published on the *bitterlemons* website². The website is set up to “contribute to mutual understanding [between Palestinians and Israelis] through the open exchange of ideas.”³ Every week an issue about the Israeli-Palestinian conflict is selected for discussion (e.g.,

²<http://www.bitterlemons.org>

³<http://www.bitterlemons.org/about/about.html>

“Disengagement: unilateral or coordinated?”), and a Palestinian editor and an Israeli editor each contribute one article addressing the issue. In addition, the Israeli and Palestinian editors invite one Israeli and one Palestinian to express their views on the issue (sometimes in the form of an interview), resulting in a total of four articles in a weekly edition. We choose the `bitterlemons` website for two reasons. First, each article is already labeled as either Palestinian or Israeli by the editors, allowing us to exploit existing annotations. Second, the `bitterlemons` corpus enables us to test the generalizability of the proposed models in a very realistic setting: training on articles written by a small number of writers (two editors) and testing on articles from a much larger group of writers (more than 200 different guests).

We collected a total of 594 articles published on the website from late 2001 to early 2005. The distribution of documents and sentences are listed in Table 1. We removed metadata from all articles, in-

	Palestinian	Israeli
Written by editors	148	149
Written by guests	149	148
Total number of documents	297	297
Average document length	740.4	816.1
Number of sentences	8963	9640

Table 1: The basic statistics of the corpus

cluding edition numbers, publication dates, topics, titles, author names and biographic information. We used OpenNLP Tools⁴ to automatically extract sentence boundaries, and reduced word variants using the Porter stemming algorithm.

We evaluated the subjectivity of each sentence using the automatic subjective sentence classifier from (Riloff and Wiebe, 2003), and find that 65.6% of Palestinian sentences and 66.2% of Israeli sentences are classified as subjective. The high but almost equivalent percentages of subjective sentences in the two perspectives support our observation in Section 2 that a perspective is largely expressed using subjective language, but that the amount of subjectivity in a document is not necessarily indicative of

its perspective.

4 Statistical Modeling of Perspectives

We develop algorithms for learning perspectives using a statistical framework. Denote a training corpus as a set of documents W_n and their perspectives labels $D_n, n = 1, \dots, N$, where N is the total number of documents in the corpus. Given a new document \tilde{W} with a unknown document perspective, the perspective \tilde{D} is calculated based on the following conditional probability.

$$P(\tilde{D}|\tilde{W}, \{D_n, W_n\}_{n=1}^N) \quad (5)$$

We are also interested in how strongly each sentence in a document conveys perspective information. Denote the intensity of the m -th sentence of the n -th document as a binary random variable $S_{m,n}$. To evaluate $S_{m,n}$, how strongly a sentence reflects a particular perspective, we calculate the following conditional probability.

$$P(S_{m,n}|\{D_n, W_n\}_{n=1}^N) \quad (6)$$

4.1 Naïve Bayes Model

We model the process of generating documents from a particular perspective as follows:

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \\ \theta &\sim \text{Dirichlet}(\alpha_\theta) \\ D_n &\sim \text{Binomial}(1, \pi) \\ W_n &\sim \text{Multinomial}(L_n, \theta_d) \end{aligned}$$

First, the parameters π and θ are sampled once from prior distributions for the whole corpus. Beta and Dirichlet are chosen because they are conjugate priors for binomial and multinomial distributions, respectively. We set the hyperparameters α_π, β_π , and α_θ to one, resulting in non-informative priors. A document perspective D_n is then sampled from a binomial distribution with the parameter π . The value of D_n is either d^0 (Israeli) or d^1 (Palestinian). Words in the document are then sampled from a multinomial distribution, where L_n is the length of the document. A graphical representation of the model is shown in Figure 1.

⁴<http://sourceforge.net/projects/opennlp/>

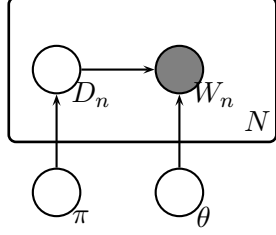


Figure 1: Naïve Bayes Model

The model described above is commonly known as a naïve Bayes (NB) model. NB models have been widely used for various classification tasks, including text categorization (Lewis, 1998). The NB model is also a building block for the model described later that incorporates sentence-level perspective information.

To predict the perspective of an unseen document using naïve Bayes, we calculate the posterior distribution of \tilde{D} in (5) by integrating out the parameters,

$$\int \int P(\tilde{D}, \pi, \theta | \{(D_n, W_n)\}_{n=1}^N, \tilde{W}) d\pi d\theta \quad (7)$$

However, the above integral is difficult to compute. As an alternative, we use Markov Chain Monte Carlo (MCMC) methods to obtain samples from the posterior distribution. Details about MCMC methods can be found in Appendix A.

4.2 Latent Sentence Perspective Model

We introduce a new binary random variable, S , to model how strongly a perspective is reflected at the sentence level. The value of S is either s^1 or s^0 , where s^1 indicates a sentence is written strongly from a perspective while s^0 indicates it is not. The whole generative process is modeled as follows:

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \\ \tau &\sim \text{Beta}(\alpha_\tau, \beta_\tau) \\ \theta &\sim \text{Dirichlet}(\alpha_\theta) \\ D_n &\sim \text{Binomial}(1, \pi) \\ S_{m,n} &\sim \text{Binomial}(1, \tau) \\ W_{m,n} &\sim \text{Multinomial}(L_{m,n}, \theta) \end{aligned}$$

The parameters π and θ have the same semantics as in the naïve Bayes model. S is naturally modeled as a binomial variable, where τ is the parameter of S . S represents how likely it is that a sentence strongly conveys a perspective. We call this model the Latent Sentence Perspective Model (LSPM) because S is not directly observed. The graphical model representation of LSPM is shown in Figure 2.

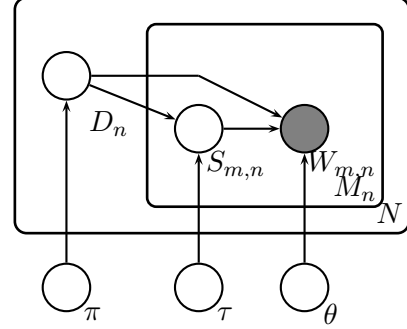


Figure 2: Latent Sentence Perspective Model

To use LSPM to identify the perspective of a new document \tilde{D} with unknown sentence perspectives \tilde{S} , we calculate posterior probabilities by summing out possible combinations of sentence perspective in the document and parameters.

$$\int \int \int \sum_{S_{m,n}} \sum_{\tilde{S}} P(\tilde{D}, S_{m,n}, \tilde{S}, \pi, \tau, \theta | \{(D_n, W_n)\}_{n=1}^N, \tilde{W}) d\pi d\tau d\theta \quad (8)$$

As before, we resort to MCMC methods to sample from the posterior distributions, given in Equations (5) and (6).

As is often encountered in mixture models, there is an identifiability issue in LSPM. Because the values of S can be permuted without changing the likelihood function, the meanings of s^0 and s^1 are ambiguous. In Figure 3a, four θ values are used to represent the four possible combinations of document perspective d and sentence perspective intensity s . If we do not impose any constraints, s^1 and s^0 are exchangeable, and we can no longer strictly interpret s^1 as indicating a strong sentence-level perspective and s^0 as indicating that a sentence carries little or no perspective information. The other problem of this parameterization is that any improvement from LSPM over the naïve Bayes model is not necessarily

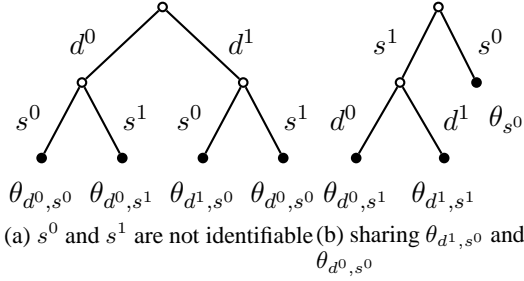


Figure 3: Two different parameterization of θ

due to the explicit modeling of sentence-level perspective. S may capture aspects of the document collection that we never intended to model. For example, s^0 may capture the editors’ writing styles and s^1 the guests’ writing styles in the *bitterlemons* corpus.

We solve the identifiability problem by forcing θ_{d^1, s^0} and θ_{d^0, s^0} to be identical and reducing the number of θ parameters to three. As shown in Figure 3b, there are separate θ parameters conditioned on the document perspective (left branch of the tree, d^0 is Israeli and d^1 is Palestinian), but there is single θ parameter when $S = s^0$ shared by both document-level perspectives (right branch of the tree). We assume that the sentences with little or no perspective information, i.e., $S = s^0$, are generated independently of the perspective of a document. In other words, sentences that are presenting common background information or introducing an issue and that do not strongly convey any perspective should look similar whether they are in Palestinian or Israeli documents. By forcing this constraint, we become more confident that s^0 represents sentences of little perspectives and s^1 represents sentences of strong perspectives from d^1 and d^0 documents.

5 Experiments

5.1 Identifying Perspective at the Document Level

We evaluate three different models for the task of identifying perspective at the document level: two naïve Bayes models (NB) with different inference methods and Support Vector Machines (SVM)

(Cristianini and Shawe-Taylor, 2000). NB-B uses full Bayesian inference and NB-M uses Maximum a posteriori (MAP). We compare NB with SVM not only because SVM has been very effective for classifying topical documents (Joachims, 1998), but also to contrast generative models like NB with discriminative models like SVM. For training SVM, we represent each document as a V -dimensional feature vector, where V is the vocabulary size and each coordinate is the normalized term frequency within the document. We use a linear kernel for SVM and search for the best parameters using grid methods.

To evaluate the statistical models, we train them on the documents in the *bitterlemons* corpus and calculate how accurately each model predicts document perspective in ten-fold cross-validation experiments. Table 2 reports the average classification accuracy across the 10 folds for each model. The accuracy of a baseline classifier, which randomly assigns the perspective of a document as Palestinian or Israeli, is 0.5, because there are equivalent numbers of documents from the two perspectives.

Model	Data Set	Accuracy	Reduction
Baseline		0.5	
SVM	Editors	0.9724	
NB-M	Editors	0.9895	61%
NB-B	Editors	0.9909	67%
SVM	Guests	0.8621	
NB-M	Guests	0.8789	12%
NB-B	Guests	0.8859	17%

Table 2: Results for Identifying Perspectives at the Document Level

The last column of Table 2 is error reduction relative to SVM. The results show that the naïve Bayes models and SVM perform surprisingly well on both the Editors and Guests subsets of the *bitterlemons* corpus. The naïve Bayes models perform slightly better than SVM, possibly because generative models (i.e., naïve Bayes models) achieve optimal performance with a smaller number of training examples than discriminative models (i.e., SVM) (Ng and Jordan, 2002), and the size of the *bitterlemons* corpus is indeed small. NB-B, which performs full Bayesian inference, improves

on NB-M, which only performs point estimation. The results suggest that the choice of words made by the authors, either consciously or subconsciously, reflects much of their political perspectives. Statistical models can capture word usage well and can identify the perspective of documents with high accuracy.

Given the performance gap between Editors and Guests, one may argue that there exist distinct editing artifacts or writing styles of the editors and guests, and that the statistical models are capturing these things rather than “perspectives.” To test if the statistical models truly are learning perspectives, we conduct experiments in which the training and testing data are mismatched, i.e., from different subsets of the corpus. If what the SVM and naïve Bayes models learn are writing styles or editing artifacts, the classification performance under the mismatched conditions will be considerably degraded.

Model	Training	Testing	Accuracy	
Baseline			0.5	
SVM	Guests	Editors	0.8822	
NB-M	Guests	Editors	0.9327	43%
NB-B	Guests	Editors	0.9346	44%
SVM	Editors	Guests	0.8148	
NB-M	Editors	Guests	0.8485	18%
NB-B	Editors	Guests	0.8585	24%

Table 3: Identifying Document-Level Perspectives with Different Training and Testing Sets

The results on the mismatched training and testing experiments are shown in Table 3. Both SVM and the two variants of naïve Bayes perform well on the different combinations of training and testing data. As in Table 2, the naïve Bayes models perform better than SVM with larger error reductions, and NB-B slightly outperforms NB-M. The high accuracy on the mismatched experiments suggests that statistical models are not learning writing styles or editing artifacts. This reaffirms that document perspective is reflected in the words that are chosen by the writers.

We list the most frequent words (excluding stopwords) learned by the the NB-M model in Table 4. The frequent words overlap greatly between the Palestinian and Israeli perspectives, in-

cluding “state,” “peace,” “process,” “secure” (“security”), and “govern” (“government”). This is in contrast to what we expect from topical text classification (e.g., “Sports” vs. “Politics”), in which frequent words seldom overlap. Authors from different perspectives often choose words from a similar vocabulary but emphasize them differently. For example, in documents that are written from the Palestinian perspective, the word “palestinian” is mentioned more frequently than the word “israel.” It is, however, the reverse for documents that are written from the Israeli perspective. Perspectives are also expressed in how frequently certain people (“sharon” v.s. “arafat”), countries (“international” v.s. “america”), and actions (“occupation” v.s. “settle”) are mentioned. While one might solicit these contrasting word pairs from domain experts, our results show that statistical models such as SVM and naïve Bayes can automatically acquire them.

5.2 Identifying Perspectives at the Sentence Level

In addition to identifying the perspective of a document, we are interested in knowing which sentences of the document strongly conveys perspective information. Sentence-level perspective annotations do not exist in the *bitterlemons* corpus, which makes estimating parameters for the proposed Latent Sentence Perspective Model (LSPM) difficult. The posterior probability that a sentence strongly convey a perspective (Example (6)) is of the most interest, but we can not directly evaluate this model without gold standard annotations. As an alternative, we evaluate how accurately LSPM predicts the perspective of a document, again using 10-fold cross validation. Although LSPM predicts the perspective of both documents and sentences, we will doubt the quality of the sentence-level predictions if the document-level predictions are incorrect.

The experimental results are shown in Table 5. We include the results for the naïve Bayes models from Table 3 for easy comparison. The accuracy of LSPM is comparable or even slightly better than that of the naïve Bayes models. This is very encouraging and suggests that the proposed LSPM closely captures how perspectives are reflected at both the document and sentence levels. Examples 1 and 2 from the introduction were predicted by LSPM as likely to

Palestinian	palestinian, israel, state, politics, peace, international, people, settle, occupation, sharon, right, govern, two, secure, end, conflict, process, side, negotiate
Israeli	israel, palestinian, state, settle, sharon, peace, arafat, arab, politics, two, process, secure, conflict, lead, america, agree, right, gaza, govern

Table 4: The top twenty most frequent stems learned by the NB-M model, sorted by $P(w|d)$

Model	Training	Testing	Accuracy
Baseline			0.5
NB-M	Guests	Editors	0.9327
NB-B	Guests	Editors	0.9346
LSPM	Guests	Editors	0.9493
NB-M	Editors	Guests	0.8485
NB-B	Editors	Guests	0.8585
LSPM	Editors	Guests	0.8699

Table 5: Results for Perspective Identification at the Document and Sentence Levels

contain strong perspectives, i.e., large $\Pr(\tilde{S} = s^1)$. Examples 3 and 4 from the introduction were predicted by LSPM as likely to contain little or no perspective information, i.e., high $\Pr(\tilde{S} = s^0)$.

The comparable performance between the naïve Bayes models and LSPM is in fact surprising. We can train a naïve Bayes model directly on the sentences and attempt to classify a sentence as reflecting either a Palestinian or Israeli perspective. A sentence is correctly classified if the predicted perspective for the sentence is the same as the perspective of the document from which it was extracted. Using this model, we obtain a classification accuracy of only 0.7529, which is much lower than the accuracy previously achieved at the document level. Identifying perspectives at the sentence level is thus more difficult than identifying perspectives at the document level. The high accuracy at the document level shows that LSPM is very effective in pooling evidence from sentences that individually contain little perspective information.

6 Conclusions

In this paper we study a new problem of learning to identify the perspective from which a text is written

at the document and sentence levels. We show that much of a document’s perspective is expressed in word usage, and statistical learning algorithms such as SVM and naïve Bayes models can successfully uncover the word patterns that reflect author perspective with high accuracy. In addition, we develop a novel statistical model to estimate how strongly a sentence conveys perspective, in the absence of sentence-level annotations. By introducing latent variables and sharing parameters, the Latent Sentence Perspective Model is shown to capture well how perspectives are reflected at the document and sentence levels. The small but positive improvement due to sentence-level modeling in LSPM is encouraging. In the future, we plan to investigate how consistently LSPM sentence-level predictions are with human annotations.

Acknowledgment

This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract number NBCHC040037.

A Gibbs Samplers

Based the model specification described in Section 4.2 we derive the Gibbs samplers (Chen et al., 2000) for the Latent Sentence Perspective Model as follows,

$$\begin{aligned}
\pi^{(t+1)} &\sim \text{Beta}(\alpha_\pi + \sum_{n=1}^N d_n + \tilde{d}^{(t+1)}, \\
&\quad \beta_\pi + N - \sum_{n=1}^N d_n + 1 - \tilde{d}^{(t+1)}) \\
\tau^{(t+1)} &\sim \text{Beta}(\alpha_\tau + \sum_{n=1}^N \sum_{m=1}^{M_n} s_{m,n} + \sum_{m=1}^{\tilde{M}} \tilde{s}_m, \\
&\quad \beta_\tau + \sum_{n=1}^N M_n - \sum_{n=1}^N \sum_{m=1}^{M_n} s_{m,n} + \tilde{M} - \sum_{m=1}^{\tilde{M}} \tilde{s}_m)
\end{aligned}$$

$$\theta^{(t+1)} \sim \text{Dirichlet}(\alpha_\theta + \sum_{n=1}^N \sum_{m=1}^{M_n} w_{m,n})$$

$$\Pr(S_{n,m}^{(t+1)} = s^1) \propto P(W_{m,n} | S_{m,n} = 1, \theta^{(t)}) \\ \Pr(S_{m,n}^{(t+1)} = 1 | \tau, D_n)$$

$$\Pr(\tilde{D}^{(t+1)} = d^1) \propto \prod_{m=1}^{\tilde{M}} \text{dbinom}(\tau_d^{(t+1)}) \\ \prod_{m=1}^{\tilde{M}} \text{dmultinom}(\theta_{d,\tilde{m}^{(t)}}) \text{dbinom}(\pi^{(t)})$$

where dbinom and dmultinom are the density functions of binomial and multinomial distributions, respectively. The superscript t indicates that a sample is from the t -th iteration. We run three chains and collect 5000 samples. The first half of burn-in samples are discarded.

References

- Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. 2004. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of ACL-2004*.
- Ming-Hui Chen, Qi-Man Shao, and Joseph G. Ibrahim. 2000. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag.
- Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.
- Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW-2003*.
- Michael L. Geis. 1987. *The Language of Politics*. Springer.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD-2004*.
- Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-1998*.
- David D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-1998*.
- S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. 2002. Mining product reputations on the web. In *Proceedings of KDD-2002*.
- Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP-2004*.
- T. Nasukawa and J. Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of K-CAP 2003*.
- Andrew Y. Ng and Michael Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS-2002*, volume 15.
- Zhongdang Pan, Chin-Chuan Lee, Joseph Man Chen, and Clement Y.K. So. 1999. One event, three stories: Media narratives of the handover of hong kong in cultural china. *Gazette*, 61(2):99–112.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP-2002*.
- Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP-2005*, pages 339–346.
- Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP-2003*.
- Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of CoNLL-2003*.
- Peter Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM TOIS*, 21(4):315–346.
- T.A. van Dijk. 1988. *News as Discourse*. Lawrence Erlbaum, Hillsdale, NJ.
- Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3).
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP-2005*.
- Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP-2003*.