Unsupervised Sentiment Analysis with a Simple and Fast Bayesian Model using Part-of-Speech Feature Selection

Christian Scheible and **Hinrich Schütze**
Institute for Natural Language Processing
University of Stuttgart
scheibcn@ims.uni-stuttgart.de

Abstract

Unsupervised Bayesian sentiment analysis often uses models that are not well motivated. Mostly, extensions of Latent Dirichlet Analysis (LDA) are applied – effectively modeling latent class distributions over words instead of documents. We introduce a Bayesian, unsupervised version of Naive Bayes for sentiment analysis and show that it offers superior accuracy and inference speed.

1 Introduction

Unsupervised models for sentiment analysis remain a challenge. While sentiment is relatively easy to detect in supervised experiments (e.g. (Pang et al., 2002)), models lacking supervision are often unable to make the distinction properly. In some domains (e.g. book reviews), topics strongly interfer with sentiment (Titov and McDonald, 2008), and unsupervised models barely beat the chance baseline (Dasgupta and Ng, 2009).

Bayesian models have received considerable attention in natural language processing research. They enable the incorporation of prior knowledge in arbitrary graphical models while parameter inference can be accomplished with simple sampling techniques (Gelman et al., 2004). In this paper, we apply a Bayesian model for unsupervised sentiment analysis of documents. Although various unsupervised Bayesian sentiment models exist, almost all of them extend Latent Dirichlet Analysis (LDA, (Blei et al., 2003)) which was designed to model document-specific topic mixtures. While LDA is a well-understood and widespread model, it is not well-suited for labeling documents. Instead, we propose a model that generates a single label per document which generates all words in it, instead of generating multiple word labels per document. Naive Bayes, which is commonly used supervisedly, meets this requirement. We will show that the unsupervised version of Naive Bayes with Bayesian Dirichlet priors achieves a higher classification accuracy than LDA on standard review classification tasks. In addition, we will demonstrate a significant speed advantage over LDA.

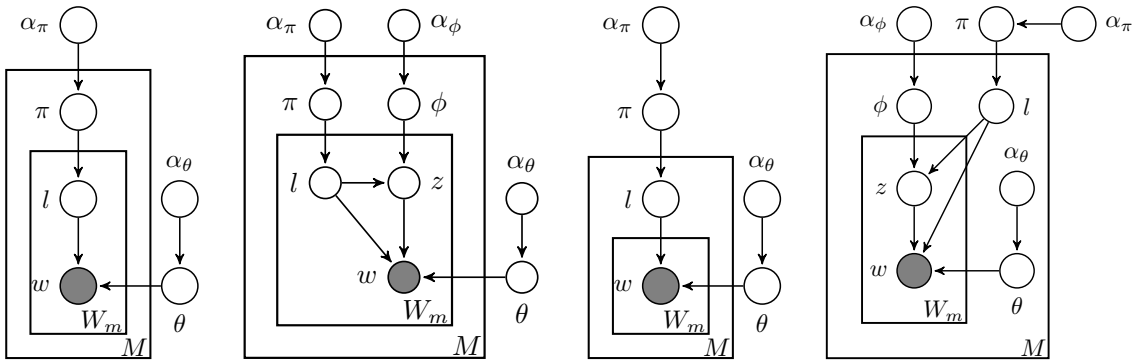
This paper is structured as follows: Section 2 describes related approaches, Section 3 contains model definition, Sections 4 and 5 presents the experimental setup and results.

2 Related Work

Most related work on Bayesian models for sentiment uses LDA-style models that predict one label per word. A notable exception is (Boyd-Graber and Resnik, 2010) who use document-level labels. Their focus however lies on supervised multilingual classification and they do not compare their model against any reference model.

Zagibalov and Carroll (2008) introduce an unsupervised model using lexical information about modifiers like negation and frequent adverbials. They automatically induce a lexicon of relevant seed words and report high classification accuracy. The model requires hand-crafted language-specific knowledge.

Dasgupta and Ng (2009) present a spectral clustering approach that yields highly competitive results. The method requires some human interaction: Selecting the best eigenvector automatically is difficult, so the authors have it manually selected, boosting the results. This constitutes a



(a) LDA (b) JST (c) BNB (d) TBNB

Figure 1: Four Bayesian topic models

form of supervised model selection.

Lin and He (2009) present the Bayesian joint sentiment-topic model (JST). It is an extension of LDA as it contains an intermediary topic layer. The authors experiment with both unsupervised and lexically supervised setups. Unfortunately, no advantage of using the additional layer could be demonstrated. We will investigate this model more closely in the following sections and compare it to our proposed model.

3 Bayesian models for document classification

We compare four Bayesian models: Latent Dirichlet Analysis (LDA, Blei et al. (2003)) presented as the latent sentiment model (LSM) by Lin et al. (2010), the joint sentiment-topic model (JST) by Lin and He (2009) that introduces an additional topic layer, Bayesian Naive Bayes (BNB) as used by Pedersen (1997), and TBNB, an extra-topic-layer version of BNB analogous to JST. We will give a short definition for each of the models in the following section. We refer the reader to the respective papers for details. Note that the terminology used in previous research is conflicting. We will thus refer to latent document or word sentiment classes as *labels* (l) and other latent classes as *topics* (z). Hyperparameters are generally referred to as α_x meaning that α is the hyperparameter for prior distribution of the multinomial with parameters x.

3.1 Model definitions

LDA. LDA (Figure 1(a)) is a model of label distributions over words in documents. Each document has a multinomial label distribution specified by θ with a Dirichlet prior with hyperparameters $\vec{\alpha}$. As labels are inferred for words, obtaining a document label requires an additional step: The document label is the most probable word label in the document by majority vote. Although presented under a different name in related work, we will refer to the model as LDA to avoid confusion.

JST. The JST model (Figure 1(b)) is an extension of LDA. Applied to reviews, LDA usually finds topics instead of sentiment (Titov and McDonald, 2008). For that reason, JST contains both labels and topics that are intended to steer the sentiment classification. The number of topics T and the number of labels K are set separately. LDA is a special case of JST since setting either $K = 1$ or $T = 1$ removes the additional latent level.

BNB and TBNB. BNB (Figure 1(c)) is the Bayesian extension of the Naive Bayes model (Pedersen, 1997). It has a global multinomial label distribution θ with a Dirichlet prior and generates one label for each document. The generative story of BNB is:

- Choose a label distribution $\pi \sim \text{Dir}(\alpha_\pi)$
- Choose a label $l_d \sim \text{Multinomial}(\pi)$ for each document document d
- Choose each word w_i in d from $p(w_i|l_d)$

Analogously to JST, we define a extension of BNB that adds a layer word-level topics (Bayesian Naive Bayes with topics, TBNB). Again, TBNB becomes BNB when the number of topics T is set to 1. We will later show that $T = 1$ is actually the best choice for JST and BNB.

3.2 Parameter estimation

We use the Hierarchical Bayes Compiler (HBC, (Daumé III, 2008)) which implements Gibbs sam-

pling (Geman and Geman, 1984) to estimate the model parameters. For the LDA and JST model, collapsed sampling is possible where the continuous parameters of the models (θ, π, ϕ) are marginalized out. For the BNB model, this would lead to a violation of independence assumptions, making explicit sampling of the parameters by prior updating necessary (cf. (Resnik and Hardisty, 2010) for details). We did not find a difference in accuracy between the collapsed and uncollapsed LDA versions.

4 Experimental setup

Data preparation. We chose two standard datasets for a comparable evaluation: the movie review (MR, (Pang et al., 2002)) and multi-domain sentiment datasets (MDS, (Blitzer et al., 2007)), each containing 1000 positive and 1000 negative documents per domain, with the MDS containing data from multiple domains (**B**ooks, **D**VDs, **E**lectronics, and **K**itchen).

We intend to reduce the feature space based on parts of speech. The processed form of multi-domain dataset is unsuitable for this as all texts were lowercased. We reconstructed the reviews from the raw files offered by the authors by extracting the full text from HTML and applying the same sentence splitter. We will make the reconstructed data publicly available.

Feature selection. Naive Bayes can be sensitive to uninformative features, making feature selection desirable. Previous work on sentiment classificaion showed that certain part-of-speech classes are highly informative for sentiment analysis, e.g. Pang et al. (2002) report high results when using only adjective features. Since the movie and product reviews differ considerably in length (746 and 158 average tokens/document, respectively), we retain more features for the MDS than for the MR dataset. Feature representations contain only adjectives for MR and adjectives, adverbs, verbs, modals, and nouns (Penn Treebank tags `JJ.*`, `MD`, `NN.*`, `RB.*`, `VB.*`) for MDS. We tag the documents with the Mate tagger (Björkelund et al., 2010). In addition, we remove stopwords and all features that occur less than 100 times in the corpus to clean the feature set and speed up computation.

Sampling. Latent class priors are set to $\frac{100}{N}$, with N being the number of classes, and all word priors are set to 0.001. Priors are symmetric. Since both datasets contain two classes, positive and negative, K is set to 2. We vary the number of topics T to study its effect. We sample for 1000 iterations and report accuracies averaged over 10 runs since model quality may vary due to random sampling.

		MR	MDS				
			B	D	E	K	*avg.*
all feat.	LDA	63.5	51.3	53.4	59.5	57.2	55.4
	BNB	61.2	51.9	53.0	61.4	62.8	57.3
selection	LDA	60.3	54.1	53.3	54.7	54.6	54.2
	BNB	**70.4**	**57.8**	**56.1**	**64.1**	**65.2**	**60.8**

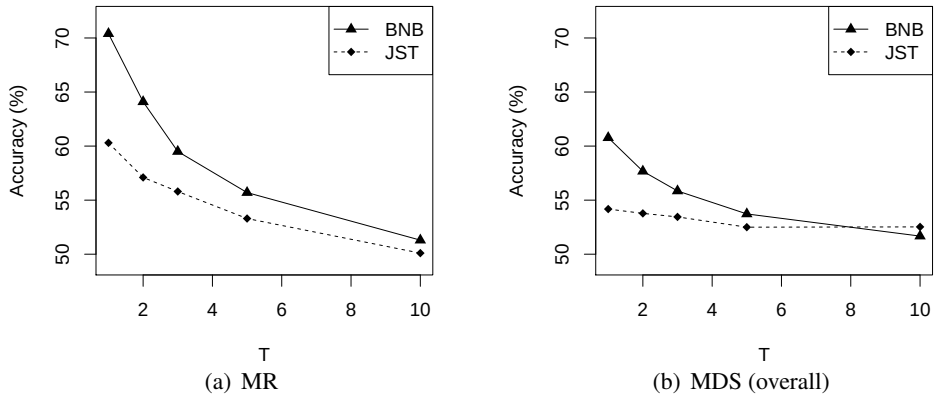
Table 1: Average accuracy (%) for each dataset

5 Experiments

This section describes our experiments on feature selection, the number of topics T, and the computation times of the models.

Feature selection We build models for (i) the full data (without stopwords and infrequent words) and (ii) with our feature selection. We try the least complex model first and set $T = 1$. As shown in Table 1, when using all features, BNB performs about equal to or slightly better than LDA except on the MR data. However, after applying feature selection BNB outperforms LDA in all cases. Feature selection shortens the documents which affects LDA negatively (Titov and McDonald, 2008), a problem whose solution would require additional computational effort (e.g. by modeling topic transitions (Blei and Moreno, 2001)). Conversely, BNB behaves as expected from a Naive Bayes model – feature selection improves the results. Errors can occur because of frequency effects: common words like *be, have, ...* receive high probabilities. Another problem is that many words are mistagged, making proper selection more difficult – particularly on the MDS where sloppy orthography is frequent. Normalization might correct this, although ungrammatical sentences might still produce erroneous results.

Number of topics. We are interested in the effects of the additional topic layer. Lin et al. (2010) do not perform an evaluation of the number of



(a) MR (b) MDS (overall)

Figure 2: Classification accuracies for different values of T

		MR	MDS				
			B	D	E	K	*avg.*
all feat.	LDA	1064	209	226	190	141	191
	BNB	728	149	156	103	110	129
selection	LDA	163	109	115	79	67	93
	BNB	109	79	87	47	50	66

Table 2: Average inference time (sec) for each dataset

topics in the unsupervised case and their results for lexically supervised classification indicate decreasing performance for higher topic numbers. To this end, we run experiments for values of T between 1 and 10 for both TBNB and JST. Note that the models simplify to BNB and LDA, respectively, if $T = 1$ since the class probabilities are then all conditioned on the same topic which essentialy leads to no conditions at all.

Figure 2 shows the classification accuracy for each dataset individually and an overall average accuracy for the MDS data. We can observe that just like in the lexically supervised case, the additional topic layer leads to a decline in accuracy when more topics are introduced, with TBNB being more sensitive than JST. We achieve the best results for models with $T = 1$, where the best TBNB setup beats the best JST setup by 6.9% on the MR data and 5.4% on the MDS. Note that the best setup for TBNB uses feature selection while the one for LDA does not. We will briefly examine implications on the computational efforts.

Computation time. BNB is the better model for document classification because it models document labels instead of word labels. Conversely, LDA needs to estimate more latent classes than needed (one per word instead of only one per document). This leads to higher computational effort which is unjustified as it is not reflected through better classification results. We measured the average time used for inference and labeling on an Intel Xeon 3.33 GHz CPU. We only report numbers for $T = 1$ as models with more topics are less accurate. Note however that these models can take significantly more time to compute since more label distributions need to be estimated. Table 2 shows the average inference time in seconds for each dataset. Using BNB saves 1/3 of computation time compared to LDA. Since LDA can only produce competitive results when run with all features, the differences become more drastic when comparing lines 1 and 4 of the table, yielding reductions up to around 90% on MR and 50% on MDS. Using a model with feature selection is thus even more desirable if efficiency is an issue.

6 Conclusion and Future Work

We presented Bayesian Naive Bayes, a Bayesian model for unsupervised document classification. We showed that BNB is superior to the LDA model on the standard unsupervised sentiment classification task. In future work, we would like to examine the behavior of our model in a semi-supervised setting where some document or feature labels are known.

Acknowledgements

This work was funded by the DFG as part of the SFB 732 project D7. We thank Thomas Müller for his helpful comments.

References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 343–348, New York, NY, USA. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 3.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 440.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 45–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sajib Dasgupta and Vincent Ng. 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*, pages 580–589, Singapore, August. Association for Computational Linguistics.

H. Daumé III. 2008. hbc: Hierarchical bayes compiler. *Pre-release version 0.7, URL http://www. cs. utah. edu/˜ hal/HBC*.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. 2004. *Bayesian data analysis*. CRC press.

S. Geman and D. Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Chenghua Lin, Yulan He, and Richard Everson. 2010. A comparative study of bayesian models for unsupervised sentiment detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 144–152, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL-EMNLP 2002*, pages 79–86.

Ted Pedersen. 1997. Knowledge lean word sense disambiguation. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 814–814. AAAI Press.

Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. Technical report, University of Maryland.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 111–120.

Taras Zagibalov and John Carroll. 2008. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 1073–1080, Stroudsburg, PA, USA. Association for Computational Linguistics.