Unsupervised Improving of Sentiment Analysis Using Global Target Context

Tomáš Brychcín **Ivan Habernal**
NTIS – New Technologies for the Information Society, and
Department of Computer Science and Engineering,
Faculty of Applied Sciences, University of West Bohemia,
Univerzitní 8, 306 14 Plzeň, Czech Republic
{brychcin,habernal}@kiv.zcu.cz

Abstract

Current approaches to document-level sentiment analysis rely on local information, e.g., the words within the given document. We try to achieve better performance by incorporating global context of the sentiment target (e.g., a movie or a product). We assume that sentiment labels of reviews about the same target are often consistent in some way. We model this consistency by Dirichlet distribution over sentiment labels and use it together with Maximum entropy classifier to gain significant improvement. This unsupervised extension increases the classification F-measure by almost 3% absolute on both Czech and English movie review datasets and outperforms the current state of the art.

1 Introduction

Sentiment analysis on the document level has been one of the most targeted research topic in the past decade (Liu and Zhang, 2012). Given a document (e.g. a review, a blog post, or a tweet), the goal is to automatically obtain its sentiment which is mostly considered as a binary value (positive and negative) or is more granular (e.g. positive, negative, and neutral or a number on the pre-defined scale).

Since the pioneering research by Pang et al. (2002), movie reviews have represented a very popular domain for evaluating sentiment analysis systems, mainly because of abundance of labeled data from existing on-line movie databases.[1] Large datasets are crucial for employing machine learning approaches.

Both approaches to sentiment analysis (machine learning-based and vocabulary-based) attempt to estimate the polarity of the document taking into account only its content (e.g. words, morphology patterns, syntax, and other features). Other external information, such as the sentiment target, the author, and others, are mostly ignored in the polarity estimation step. This means that the distribution of sentiment for each target is considered as random.

We assume that sentiment labels for each target are not independent of each other. This means that given a movie with the majority of positive reviews, there is a chance that the next unknown review will be positive as well. We model this assumption as a Dirichlet distribution over sentiment labels for each target. In summary, our approach to sentiment analysis consists of two steps. In the first step, we employ a supervised Maximum entropy classifier in order to estimate sentiment label probabilities for each review. In the second (unsupervised) step, these labels are iteratively updated using Gibbs sampling in order to maximize the probability of sentiments of each target.[2]

A big challenge in the sentiment analysis task are non-mainstream languages,[3] mostly because of the lack of precise polarity lexicons, annotated datasets, and other resources. Morphologically rich languages may also require different treatment than English, because of their rich vocabulary. Therefore, we report our result on two movie review datasets in two languages — the English IMDB and Czech CSFD datasets.

[1] One might argue that if movie or product databases already contain reviews labeled with e.g. number of stars, it is useless to try to estimate it automatically; however, not all databases are alike, e.g., the Polish movie database has no such star rating and contains only pure text reviews.

[2] Through the rest of the paper, we will use *target* and *movie* interchangeably.

[3] Majority of research in sentiment analysis focuses on English or Chinese.

2 Related work

An up-to-date survey of the entire sentiment analysis field can be found in (Liu and Zhang, 2012). Recently, there has been a shift to semi-supervised or unsupervised methods. Many of them build on graphical models, mostly adapting the topic model idea from LDA (Blei et al., 2003), such as Joint ST (Lin and He, 2009), ARO (Zhang et al., 2011), Twofold-LDA (Burns et al., 2011), NB-LDA (Zhang et al., 2013), ME-LDA (Zhao et al., 2012), and others (Li et al., 2010; Maas et al., 2011). Most of these approaches try to identify the polarity of words on the first place. Furthermore, they treat each document or target entity separately in the sentiment identification phase. The global context of documents is taken into account in cases where sentiment is conditioned on the user or topics. Some of these approaches still require a seed of sentiment-bearing words, however, they do not require large sets of labeled data as in supervised machine learning approaches (Liu and Zhang, 2012).

In Czech, sentiment analysis has gained attention only very recently. In their first attempt, Steinberger et al. (2011) used machine translation and vocabulary triangulation to obtain the Czech sentiment lexicon for entity-level analysis. They reported results on the news domain. Veselovská (2012) tested Naive Bayes classifier on two small sentence-level corpora that were manually annotated; however, the results were described only as preliminary by the author. Habernal et al. (2013) created three large labeled corpora (10k, 90k, and 130k reviews/posts) and tested various preprocessing techniques suitable for Czech, as well as various features and classifiers. They further employed semantic spaces as a mean for reducing data sparsity in morphologically rich languages (Habernal and Brychcín, 2013) and achieved state-of-the-art performance in Czech.

Although an exhaustive amount of research is devoted to semi-supervised methods, to the best of our knowledge, no related work tried to combine a supervised approach to document-level sentiment analysis with modeling dependencies of sentiment according to their targets in an unsupervised manner.

3 Baseline

Let the data are divided into M review targets, where each target contains N_m reviews. In the following text, we will use T_{mn} for denoting the review at the position n in the m-th target.

As a baseline we used the Maximum entropy classifier (Berger et al., 1996)

$$P^{\text{ME}}(S_{mn} = s|T_{mn}) = \frac{1}{Z(T_{mn})} \prod_{i=1}^{I} e^{\lambda_i f_i(T_{mn},s)}, \quad (1)$$

where s is a sentiment label (a member from a finite set $\mathfrak{S}$) for a review, T_{mn} is our knowledge about review (the review itself at n-th position in m-th target), $f_i\,(T_{mn}, s)$ is an i-th feature function, λ_i is corresponding weight and $Z(T_{mn})$ is a normalization factor. For estimating parameters of Maximum entropy model we used limited memory BFGS (L-BFGS) method (Nocedal, 1980).

In the baseline classifier, we rely on two kinds of binary features, namely the presence of word unigrams and bigrams in the review text (the same baseline that was used in (Habernal et al., 2013)). This model is denoted as *ME* in following text.

We also extend the feature set by presence of word clusters (derived from semantic spaces) in the same way as in (Habernal and Brychcín, 2013). We refer to this model as *ME+sspace*.

4 Global context extension

Our idea is that the final label decision would take into account both the score from Maximum entropy classifier as well as the likelihood of appropriate sentiment label in whole context of a review target (global context). Each sentiment label classification S_{mn} on each position n affects the probability of the sentiment labels of all other reviews in target $\boldsymbol{T}_m$. The selection of the most probable sequence of sentiment labels leads to exponential complexity.

We provide approximation of this problem by Gibbs sampling in the generative model defined bellow. The complete overview of our approach is depicted in Figure 1. The generative process for sentiment labels sequence is as follows:

1. For each target $\boldsymbol{T}_m \in \boldsymbol{T}$ sample a distribution $\boldsymbol{\theta}_m \sim \mathit{Dirichlet}\,(\boldsymbol{\alpha})$ over all sentiment labels $s \in \mathfrak{S}$, where $\boldsymbol{\alpha}$ is a vector of hyperparameters of Dirichlet distribution.

2. For each review $T_{mn} \in \boldsymbol{T}_m$, where $1 \leq n \leq N_m$ sample a sentiment label S_{mn} according to

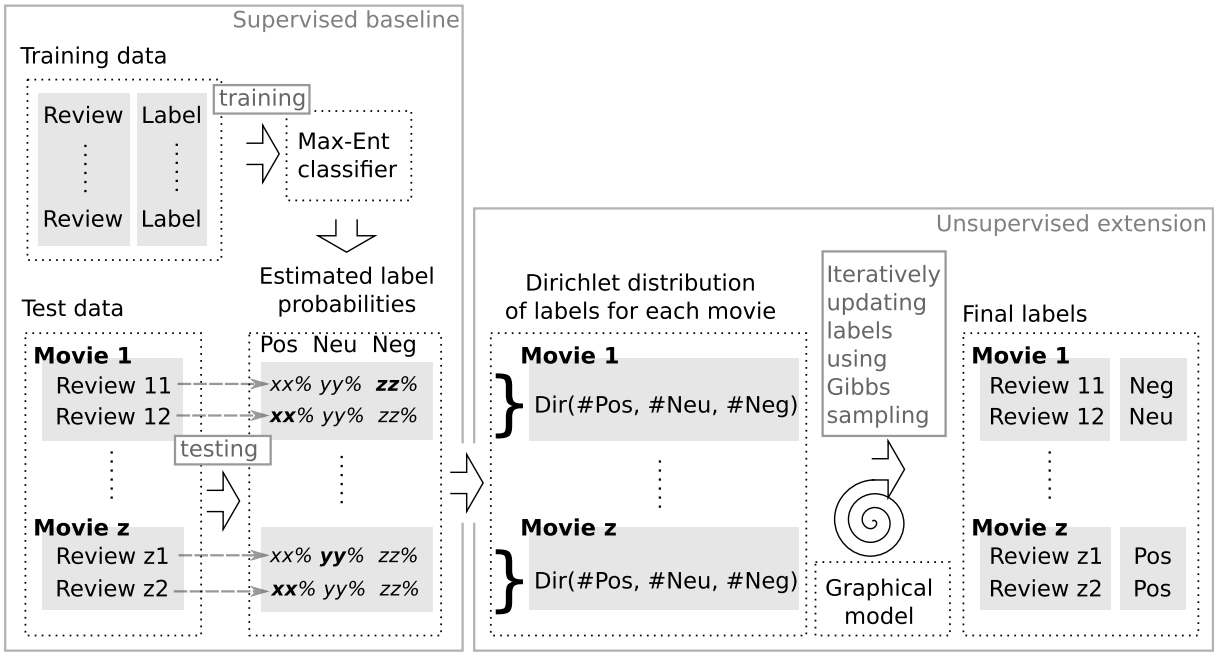


Figure 1: Diagram describing our sentiment model.

$$S_{mn} \sim \frac{\theta_m^{(s)} P^{\mathrm{ME}}\left(S_{mn}=s|T_{mn}\right)}{\sum\limits_{i\in\mathfrak{S}} \theta_m^{(i)} P^{\mathrm{ME}}\left(S_{mn}=i|T_{mn}\right)}, \quad (2)$$

where $\theta_m^{(s)}$ is the probability of sentiment label s in target $\boldsymbol{T}_m$ and $P^{\mathrm{ME}}\left(s|T_{mn}\right)$ is the label probability of the current review given by Maximum entropy model. The probability distribution, from which the labels S_{mn} are sampled, is given by probability $\theta_m^{(s)}$ rescaled by the score from Maximum entropy classifier.

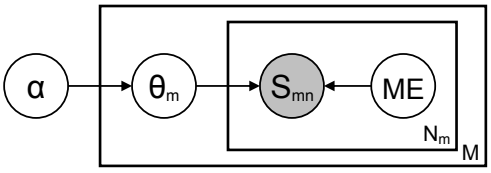


Figure 2: Plate notation representing our sentiment model. *ME* circle means the output from Maximum entropy classifier.

Plate representation of our generative model is shown in figure 2.

The Gibbs sampler needs to compute $P\left(S_{mn}|\boldsymbol{S}_{\neg mn}, T_{mn}, \boldsymbol{\alpha}\right)$, the probability of a sentiment label S_{mn} that is being assigned to a review T_{mn}, given all other labels assignments to all other reviews in appropriate review target $\boldsymbol{T}_m$.

Gibbs sampling of the Dirichlet-multinomial distribution, already derived for LDA by Griffiths and Steyvers (2004), results in simple formula

$$\begin{aligned} &P\left(S_{mn}=s|\boldsymbol{S}_{\neg mn}, \boldsymbol{\alpha}\right) \\ &\quad = \frac{c^{(s)}_{\neg mn}+\alpha_s}{\sum\limits_{i\in\mathfrak{S}} c^{(i)}_{\neg mn}+\alpha_i} \propto c^{(s)}_{\neg mn}+\alpha_s, \end{aligned} \quad (3)$$

where $\boldsymbol{S}_{\neg mn}$ means all sentiment labels except the one at position n in m-th review target. The $c^{(s)}_{\neg mn}$ denotes the number of times that the sentiment label s was assigned to the review in m-th target except the position n.

We use Maximum entropy classifier to rescale these probabilities. Final formula for sampling sentiment labels combines the information from particular review as well as contextual information about other reviews in appropriate review target

$$\begin{aligned} &P\left(S_{mn}=s|\boldsymbol{S}_{\neg mn}, T_{mn}, \boldsymbol{\alpha}\right) \\ &\quad \propto \frac{\left(c^{(s)}_{\neg mn}+\alpha_s\right) P^{\mathrm{ME}}\left(s|T_{mn}\right)}{\sum\limits_{i\in\mathfrak{S}} \left(c^{(i)}_{\neg mn}+\alpha_i\right) P^{\mathrm{ME}}\left(i|T_{mn}\right)} \\ &\quad \propto \left(c^{(s)}_{\neg mn}+\alpha_s\right) P^{\mathrm{ME}}\left(s|T_{mn}\right). \end{aligned} \quad (4)$$

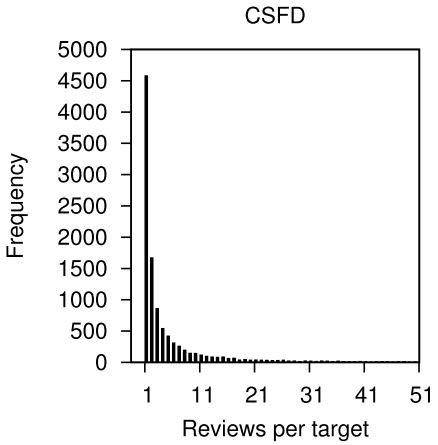


Figure 3: Histogram of reviews per target on CSFD dataset. Frequency (y axis) means how many targets have the given number of reviews (x axis).

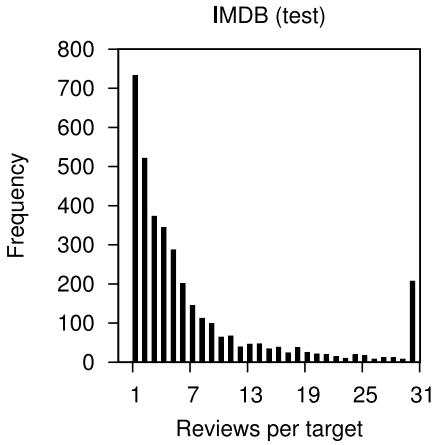


Figure 4: Histogram of reviews per target on IMDB test dataset.

5 Datasets

We perform our experiments on two datasets in the movie review domain. An English dataset from the Internet Movie Database (IMDB), provided by (Maas et al., 2011), contains 25k training and 25k test examples labeled with either positive or negative sentiment. There are also another 50k additional unlabeled reviews. All reviews are accompanied with their corresponding movies' URLs.

A Czech dataset from the Czech Movie Database (CSFD), provided by (Habernal et al., 2013), consists of $\approx$ 90k reviews equally split into positive, negative, and neutral ones. As the provided dataset did not contain information about the target movies, we tried to match the reviews and movies automatically. Unfortunately, in few cases we were not able to find the appropriate movie given the review, thus the resulting dataset slightly differs from the one from (Habernal et al., 2013). However, we report all results on the new dataset (where the reviews are paired with their movies) and also provide it for any further research.[4]

5.1 Data statistics

Figures 3 and 4 display statistics for the CSFD and IMDB test datasets, respectively, in terms of the frequency of targets with a particular number of reviews. In both datasets, the overall trend is that most of the movies have 1–10 reviews. The mean is 8.6 reviews per movie in CSFD and 7.0 in IMDB, respectively. The reason of the large peak at 30 in IMDB is the restriction of maximum reviews per movie to 30 by Maas et al. (2011).

To support our idea of some consistency in sentiment related to one target, we captured the percentage of the major sentiment label for each target, as shown in Figure 5. Each 'bin' on the Y axis deals with targets having a certain number of reviews, i.e., 1–10, 11–20, etc. For each bin, we compute the ratio of the major sentiment (i.e., if a movie has 7 positive, 2 neutral, and 1 negative review, the ratio is 0.7) and plot it as a probability distribution. It actually corresponds to consistency of reviews per target. Obviously, for targets with 1–5 or 1–10 reviews (the first Y axis bin), the graph is skewed towards 1.0. This is caused by targets with only a single review, thus the probability of major sentiment for these targets is always 1.0. With increasing number of reviews per target, the sentiment becomes a mixture where the prevalence of the major sentiment declines, yet it remains dominant (as can be seen in Figure 5).

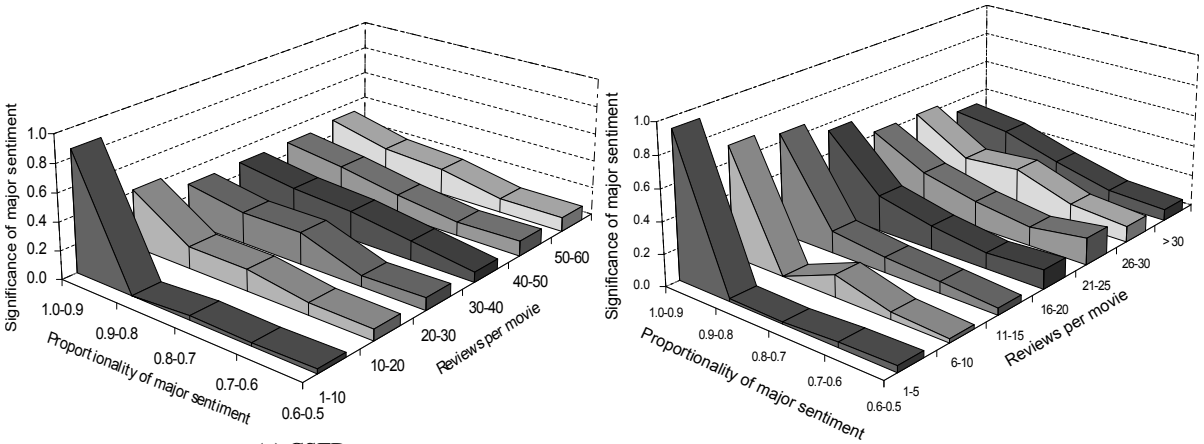
Note that we show these statistics only on test data in the IMDB dataset, as our extension does not involve the training data.

6 Results and discussion

We perform our experiments in 10-fold cross validation manner on the CSFD dataset. For the IMDB dataset, the training and test data are already separated.

In our experiment we used symmetric Dirichlet distribution, which do not favor any sentiment label over another. Results obtained by 100 iterations of Gibbs sampling and hyper-parameters $\alpha_s = 0.0001$ are shown in Table 1.

[4] http://liks.fav.zcu.cz/sentiment



(a) CSFD

(b) IMDB

Figure 5: Proportionality of major sentiments for various numbers of reviews per target.

model \ dataset	CSFD	IMDB
(Maas et al., 2011)		88.89
(Habernal and Brychcín, 2013)	78.92	89.46
(Trivedi and Eisenstein, 2013)		91.36
ME baseline	77.58	89.34
ME + sspace	78.72 (+1.14)	89.46 (+0.12)
ME + Dir	80.57 (+2.99)	92.09 (+2.75)
ME + sspace + Dir	**81.53** (+3.95)	**92.24** (+2.90)

95% confidence interval for CZ = ± 0.3.
95% confidence interval for EN = ± 0.4.

Table 1: F-measure achieved on both datasets. The improvements are measured against baseline. Note that improvement given by semantic spaces extension on English dataset is not statistically significant.

We also experimented with the number of iterations needed for sufficient inference (Figure 6) and concluded that 100 iterations is far enough. Note that the improvements in Figures 6 and 7 are always taken against the same model without global context, i.e. *ME+sspace+Dir* is compared to the *ME+sspace*, not to the *ME*.

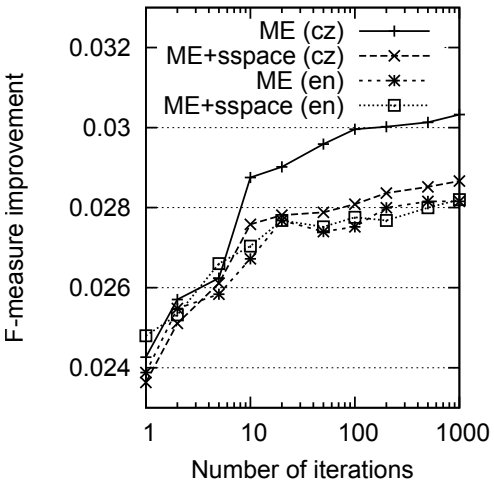


Figure 6: Improvement in F-measure depending on number of iterations of Gibbs sampling.

The selection of appropriate hyper-parameters of Dirichlet distribution can be important for such a task. The improvements in F-measure depending on different α_s are shown in figure 7. Lower α_s achieves higher improvement in performance. With lower α_s, the Dirichlet distribution is sharper and also the more consistent the review labels are expected to be in average.

We suppose this is caused mainly by the fact that many review targets have only one review (100% consistency). See Figures 3 and 4 for detailed statistics on datasets. Thus the global context should help in widely reviewed targets. In cases where the target has only one review, our extension has no effect on the final sentiment label (the label is only determined by Maximum entropy classifier).

7 Summary

7.1 Future work

In future work we would like to investigate another combinations of document level information together with global context information. We expect that linear interpolation with weights tuned on held-out data would be an efficient combination of such sources of information.

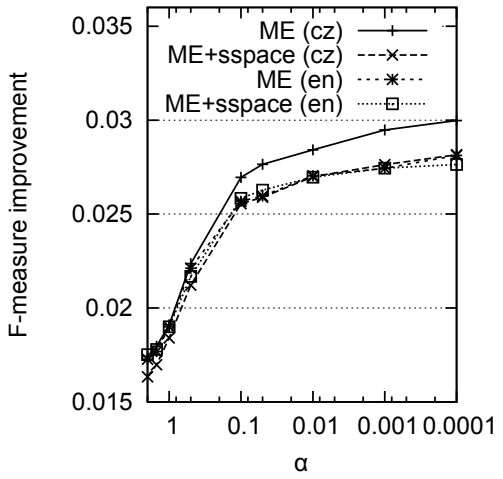


Figure 7: Improvement in F-measure depending on the parameter of Dirichlet distribution.

Another interesting idea is to use Dirichlet distribution with different hyper-parameters for targets with different number of reviews, as the Dirichlet distribution is supposed to have different shape for sparsely reviewed targets, compared to the targets with many reviews.

7.2 Conclusion

In this work we investigated global target context as a new source of information for sentiment analysis. We placed the Dirichlet distribution on sentiment labels belonging to the same review target. We combined the global target context information together with the document level classification (Maximum entropy classifier) and used Gibbs sampling for inference the sentiment labels. Our extension satisfies the unsupervised fashion and significantly improves classification F-measure by almost 3% which yields new state-of-the-art results.

Acknowledgments

This work was supported by grant no. SGS-2013-029 Advanced computing and information systems, by the POSTDOC grant from University of West Bohemia, and by the European Regional Development Fund (ERDF), project "NTIS – New Technologies for Information Society", European Center of Excellence, CZ.1.05/1.1.00/02.0090. Access to the MetaCentrum computing facilities provided under the program "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005, funded by the Ministry of Education, Youth, and Sports of the Czech Re-

public, is highly appreciated. The access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144 is acknowledged. We also thank the reviewers for their detailed and insightful comments.

References

A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71, March.

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

Nicola Burns, Yaxin Bi, Hui Wang, and Terry Anderson. 2011. A twofold-lda model for customer review analysis. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, pages 253–256, Washington, DC, USA. IEEE Computer Society.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April.

Ivan Habernal and Tomáš Brychcín. 2013. Semantic spaces for sentiment analysis. In *Text, Speech and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 482–489, Berlin Heidelberg. Springer.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2013. Sentiment analysis in czech social media using supervised machine learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–74, Atlanta, Georgia, June. Association for Computational Linguistics.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA*. AAAI Press.

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.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Com putational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jorge Nocedal. 1980. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, 35(151):773–782.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Josef Steinberger, Polina Lenkova, Mijail Alexandrov Kabadjov, Ralf Steinberger, and Erik Van der Goot. 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, RANLP'11, pages 770–775.

Rakshit Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 808–813, Atlanta, Georgia, June. Association for Computational Linguistics.

Kateřina Veselovská. 2012. Sentence-level sentiment analysis in czech. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 65:1–65:4, New York, NY, USA. ACM.

Yong Zhang, Dong-Hong Ji, Ying Su, and Cheng Sun. 2011. Sentiment analysis for online reviews using an author-review-object model. In *Proceedings of the 7th Asia conference on Information Retrieval Technology*, AIRS'11, pages 362–371, Berlin, Heidelberg. Springer-Verlag.

Yong Zhang, Dong-Hong Ji, Ying Su, and Hongmiao Wu. 2013. Joint naive bayes and lda for unsupervised sentiment analysis. In Jian Pei, VincentS. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, volume 7818 of *Lecture Notes in Computer Science*, pages 402–413. Springer Berlin Heidelberg.

Tong Zhao, Chunping Li, Qiang Ding, and Li Li. 2012. User-sentiment topic model: refining user's topics with sentiment information. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 10:1–10:9, New York, NY, USA. ACM.