Semi-supervised latent variable models for sentence-level sentiment analysis

Oscar Täckström
SICS, Kista / Uppsala University, Uppsala
oscar@sics.se

Ryan McDonald
Google, Inc., New York
ryanmcd@google.com

Abstract

We derive two variants of a semi-supervised model for fine-grained sentiment analysis. Both models leverage abundant natural supervision in the form of review ratings, as well as a small amount of manually crafted sentence labels, to learn sentence-level sentiment classifiers. The proposed model is a fusion of a fully supervised structured conditional model and its partially supervised counterpart. This allows for highly efficient estimation and inference algorithms with rich feature definitions. We describe the two variants as well as their component models and verify experimentally that both variants give significantly improved results for sentence-level sentiment analysis compared to all baselines.

1 Sentence-level sentiment analysis

In this paper, we demonstrate how combining coarse-grained and fine-grained supervision benefits sentence-level sentiment analysis – an important task in the field of opinion classification and retrieval (Pang and Lee, 2008). Typical supervised learning approaches to sentence-level sentiment analysis rely on sentence-level supervision. While such fine-grained supervision rarely exist naturally, and thus requires labor intensive manual annotation effort (Wiebe et al., 2005), coarse-grained supervision is naturally abundant in the form of online review ratings. This coarse-grained supervision is, of course, less informative compared to fine-grained supervision, however, by combining a small amount of sentence-level supervision with a large amount of document-level supervision, we are able to substantially improve on the sentence-level classification task. Our work combines two strands of research: models for sentiment analysis that take document structure into account; and models that use latent variables to learn unobserved phenomena from that which can be observed.

Exploiting document structure for sentiment analysis has attracted research attention since the early work of Pang and Lee (2004), who performed minimal cuts in a sentence graph to select subjective sentences. McDonald et al. (2007) later showed that jointly learning fine-grained (sentence) and coarse-grained (document) sentiment improves predictions at both levels. More recently, Yessenalina et al. (2010) described how sentence-level latent variables can be used to improve document-level prediction and Nakagawa et al. (2010) used latent variables over syntactic dependency trees to improve sentence-level prediction, using only labeled sentences for training. In a similar vein, Sauper et al. (2010) integrated generative content structure models with discriminative models for multi-aspect sentiment summarization and ranking. These approaches all rely on the availability of fine-grained annotations, but Täckström and McDonald (2011) showed that latent variables can be used to learn fine-grained sentiment using only coarse-grained supervision. While this model was shown to beat a set of natural baselines with quite a wide margin, it has its shortcomings. Most notably, due to the loose constraints provided by the coarse supervision, it tends to only predict the two dominant fine-grained sentiment categories well for each document sentiment category, so that almost all sentences in positive documents are deemed positive or neutral, and vice versa for negative documents. As a way of overcoming these shortcomings, we propose to fuse a coarsely supervised model with a fully supervised model.

Below, we describe two ways of achieving such a combined model in the framework of structured conditional latent variable models. Contrary to (generative) topic models (Mei et al., 2007; Titov and

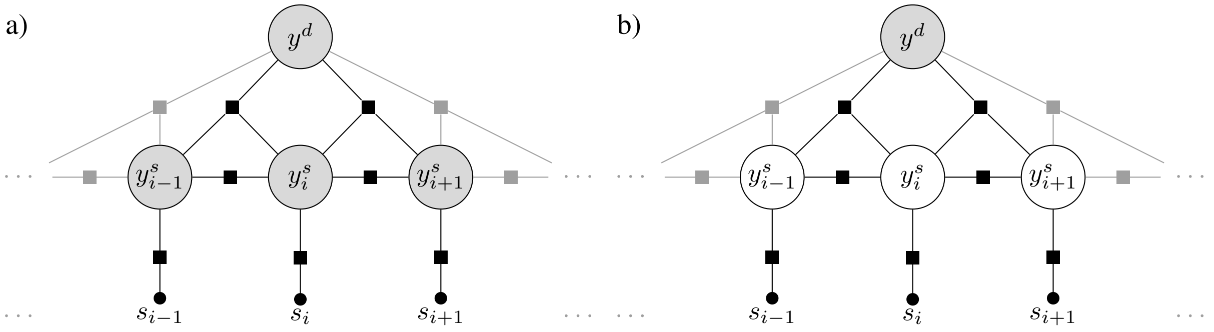

Figure 1: a) Factor graph of the fully observed graphical model. b) Factor graph of the corresponding latent variable model. During training, shaded nodes are observed, while non-shaded nodes are unobserved. The input sentences s_i are always observed. Note that there are no factors connecting the document node, y^d, with the input nodes, $\boldsymbol{s}$, so that the sentence-level variables, $\boldsymbol{y}^s$, in effect form a bottleneck between the document sentiment and the input sentences.

McDonald, 2008; Lin and He, 2009), structured conditional models can handle rich and overlapping features and allow for exact inference and simple gradient based estimation. The former models are largely orthogonal to the one we propose in this work and combining their merits might be fruitful. As shown by Sauper et al. (2010), it is possible to fuse generative document structure models and task specific structured conditional models. While we do model document structure in terms of sentiment transitions, we do not model topical structure. An interesting avenue for future work would be to extend the model of Sauper et al. (2010) to take coarse-grained task-specific supervision into account, while modeling fine-grained task-specific aspects with latent variables.

Note also that the proposed approach is orthogonal to semi-supervised and unsupervised induction of context independent (prior polarity) lexicons (Turney, 2002; Kim and Hovy, 2004; Esuli and Sebastiani, 2009; Rao and Ravichandran, 2009; Velikovich et al., 2010). The output of such models could readily be incorporated as features in the proposed model.

1.1 Preliminaries

Let d be a document consisting of n sentences, $\boldsymbol{s} = (s_i)_{i=1}^n$, with a document–sentence-sequence pair denoted $\boldsymbol{d} = (d, \boldsymbol{s})$. Let $\boldsymbol{y}^d = (y^d, \boldsymbol{y}^s)$ denote random variables[1] – the document level sentiment, y^d, and the sequence of sentence level sentiment, $\boldsymbol{y}^s = (y_i^s)_{i=1}^n$. In what follows, we assume that we have access to two training sets: a small set of fully labeled instances, $\mathcal{D}_F = \{(\boldsymbol{d}_j, \boldsymbol{y}_j^d)\}_{j=1}^{m_f}$, and a large set of coarsely labeled instances $\mathcal{D}_C = \{(\boldsymbol{d}_j, y_j^d)\}_{j=m_f+1}^{m_f+m_c}$. Furthermore, we assume that y^d and all y_i^s take values in $\{\text{POS}, \text{NEG}, \text{NEU}\}$.

We focus on structured conditional models in the exponential family, with the standard parametrization

$$p_\theta(y^d, \boldsymbol{y}^s | \boldsymbol{s}) = \exp\left\{\langle \phi(y^d, \boldsymbol{y}^s, \boldsymbol{s}), \theta\rangle - A_\theta(\boldsymbol{s})\right\} ,$$

where $\theta \in \Re^n$ is a parameter vector, $\phi(\cdot) \in \Re^n$ is a vector valued feature function that factors according to the graph structure outlined in Figure 1, and A_θ is the $\log$-partition function. This class of models is known as conditional random fields (CRFs) (Lafferty et al., 2001), when all variables are observed, and as hidden conditional random fields (HCRFs) (Quattoni et al., 2007), when only a subset of the variables are observed.

1.2 The fully supervised fine-to-coarse model

McDonald et al. (2007) introduced a fully supervised model in which predictions of coarse-grained (document) and fine-grained (sentence) sentiment are learned and inferred jointly. They showed that learning both levels jointly improved performance at both levels, compared to learning each level individually, as well as to using a cascaded model in which the predictions at one level are used as input to the other.

Figure 1a outlines the factor graph of the corre-

[1] We are abusing notation throughout by using the same symbols to refer to random variables and their particular assignments.

sponding conditional random field.[2] The parameters, θ_F, of this model can be estimated from the set of fully labeled data, $\mathcal{D}_F$, by maximizing the joint conditional likelihood function

$$L_F(\theta_F) = \sum_{j=1}^{m_f} \log p_{\theta_F}(y_j^d, \boldsymbol{y}_j^s | \boldsymbol{s}_j) - \frac{\|\theta_F\|^2}{2\sigma_F^2} ,$$

where σ_F^2 is the variance of the $\mathrm{Normal}(0, \sigma_F^2)$ prior. Note that L_F is a concave function and consequently its unique maximum can be found by gradient based optimization techniques.

1.3 Latent variables for coarse supervision

Recently, Täckström and McDonald (2011) showed that fine-grained sentiment can be learned from coarse-grained supervision alone. Specifically, they used a HCRF model with the same structure as that in Figure 1a, but with sentence labels treated as latent variables. The factor graph corresponding to this model is outlined in Figure 1b.

The fully supervised model might benefit from factors that directly connect the document variable, y^d, with the inputs $\boldsymbol{s}$. However, as argued by Täckström and McDonald (2011), when only document-level supervision is available, the document variable, y^d, should be independent of the input, $\boldsymbol{s}$, conditioned on the latent variables, $\boldsymbol{y}^s$. This prohibits the model from bypassing the latent variables, which is crucial, since we seek to improve the sentence-level predictions, rather than the document-level predictions.

The parameters, θ_C, of this model can be estimated from the set of coarsely labeled data, $\mathcal{D}_C$, by maximizing the marginalized conditional likelihood function

$$L_C(\theta_C) = \sum_{j=m_f+1}^{m_f+m_c} \log \sum_{\boldsymbol{y}^s} p_{\theta_C}(y_j^d, \boldsymbol{y}^s | \boldsymbol{s}_j) - \frac{\|\theta_C\|^2}{2\sigma_C^2} ,$$

where the marginalization is over all possible sequences of latent sentence label assignments $\boldsymbol{y}^s$.

Due to the introduction of latent variables, the marginal likelihood function is non-concave and thus there are no guarantees of global optimality, however, we can still use a gradient based optimization technique to find a local maximum.

2 Combining coarse and full supervision

The fully supervised and the partially supervised models both have their merits. The former requires an expensive and laborious process of manual annotation, while the latter can be used with readily available document labels, such as review star ratings. The latter, however, has its shortcomings in that the coarse-grained sentiment signal is less informative compared to a fine-grained signal. Thus, in order to get the best of both worlds, we would like to combine the merits of both of these models.

2.1 A cascaded model

A straightforward way of fusing the two models is by means of a cascaded model in which the predictions of the partially supervised model, trained by maximizing $L_C(\theta_C)$ are used to derive additional features for the fully supervised model, trained by maximizing $L_F(\theta_F)$.

Although more complex representations are possible, we generate meta-features for each sentence based solely on operations on the estimated distributions, $p_{\theta_C}(y^d, y_i^s | \boldsymbol{s})$. Specifically, we encode the following probability distributions as discrete features by uniform bucketing, with bucket width 0.1: the joint distribution, $p_{\theta_C}(y^d, y_i^s | \boldsymbol{s})$; the marginal document distribution, $p_{\theta_C}(y^d | \boldsymbol{s})$; and the marginal sentence distribution, $p_{\theta_C}(y_i^s | \boldsymbol{s})$. We also encode the argmax of these distributions, as well as the pairwise combinations of the derived features.

The upshot of this cascaded approach is that it is very simple to implement and efficient to train. The downside is that only the partially supervised model influences the fully supervised model; there is no reciprocal influence between the models. Given the non-concavity of $L_C(\theta_C)$, such influence could be beneficial.

2.2 Interpolating likelihood functions

A more flexible way of fusing the two models is to interpolate their likelihood functions, thereby allowing for both coarse and joint supervision of the same model. Such a combination can be achieved by constraining the parameters so that $\theta_I = \theta_F = \theta_C$ and taking the mean of the likelihood functions L_F and L_C, appropriately weighted by a hyper-parameter λ.

[2] Figure 1a differs slightly from the model employed by McDonald et al. (2007), where they had factors connecting the document label y^d with each input s_i as well.

The result is the interpolated likelihood function

$$L_I(\theta_I) = \lambda L_F(\theta_I) + (1 - \lambda) L_C(\theta_I)\,.$$

A simple, yet efficient, way of optimizing this objective function is to use stochastic gradient ascent with learning rate η. At each step we select a fully labeled instance, $(\boldsymbol{d}_j, \boldsymbol{y}_j^d) \in \mathcal{D}_F$, with probability λ and a coarsely labeled instance, $(\boldsymbol{d}_j, y_j^d) \in \mathcal{D}_C$, with probability $(1 - \lambda)$. We then update the parameters, θ_I, according to the gradients ∂L_F and ∂L_C, respectively. In principle we could use different learning rates η_F and η_C as well as different prior variances σ_F^2 and σ_C^2, but in what follows we set them equal.

Since we are interpolating conditional models, we need at least partial observations of each instance. Methods for blending discriminative and generative models (Lasserre et al., 2006; Suzuki et al., 2007; Agarwal and Daumé, 2009; Sauper et al., 2010), would enable incorporation of completely unlabeled data as well. It is straightforward to extend the proposed model along these lines, however, in practice coarsely labeled sentiment data is so abundant on the web (e.g., rated consumer reviews) that incorporating completely unlabeled data seems superfluous. Furthermore, using conditional models with shared parameters throughout allows for rich overlapping features, while maintaining simple and efficient inference and estimation.

3 Experiments

For the following experiments, we used the same data set and a comparable experimental setup to that of Täckström and McDonald (2011).[3] We compare the two proposed hybrid models (Cascaded and Interpolated) to the fully supervised model of McDonald et al. (2007) (FineToCoarse) as well as to the soft variant of the coarsely supervised model of Täckström and McDonald (2011) (Coarse).

The learning rate was fixed to $\eta = 0.001$, while we tuned the prior variances, σ^2, and the number of epochs for each model. When sampling according to λ during optimization of $L_I(\theta_I)$, we cycle through $\mathcal{D}_F$ and $\mathcal{D}_C$ deterministically, but shuffle these sets between epochs. Due to time constraints, we fixed the interpolation factor to $\lambda = 0.1$, but tuning this could potentially improve the results of the interpolated model. For the same reason we allowed a maximum of 30 epochs, for all models, while Täckström and McDonald (2011) report a maximum of 75 epochs.

To assess the impact of fully labeled versus coarsely labeled data, we took stratified samples without replacement, of sizes 60, 120, and 240 reviews, from the fully labeled folds and of sizes 15,000 and 143,580 reviews from the coarsely labeled data. On average each review consists of ten sentences. We performed 5-fold stratified cross-validation over the labeled data, while using stratified samples for the coarsely labeled data. Statistical significance was assessed by a hierachical bootstrap of 95% confidence intervals, using the technique described by Davison and Hinkley (1997).

3.1 Results and analysis

Table 1 lists sentence-level accuracy along with 95% confidence interval for all tested models. We first note that the interpolated model dominates all other models in terms of accuracy. While the cascaded model requires both large amounts of fully labeled and coarsely labeled data, the interpolated model is able to take advantage of both types of data on its own and jointly. Still, by comparing the fully supervised and the coarsely supervised models, the superior impact of fully labeled over coarsely labeled data is evident. As can be seen in Figure 2, when all data is used, the cascaded model outperforms the interpolated model for some recall values, and vice versa, while both models dominate the supervised approach for the full range of recall values.

As discussed earlier, and confirmed by Table 2, the coarse-grained model only performs well on the predominant sentence-level categories for each document category. The supervised model handles negative and neutral sentences well, but performs poorly on positive sentences even in positive documents. The interpolated model, while still better at capturing the predominant category, does a better job overall.

These results are with a maximum of 30 training iterations. Preliminary experiments with a maximum of 75 iterations indicate that all models gain from more iterations; this seems to be especially true for the supervised model and for the cascaded model with less amount of course-grained data.

[3] The annotated test data can be downloaded from http://www.sics.se/people/oscar/datasets.

	$\lvert\mathcal{D}_C\rvert = 15{,}000$			$\lvert\mathcal{D}_C\rvert = 143{,}580$		
	$\lvert\mathcal{D}_F\rvert = 60$	$\lvert\mathcal{D}_F\rvert = 120$	$\lvert\mathcal{D}_F\rvert = 240$	$\lvert\mathcal{D}_F\rvert = 60$	$\lvert\mathcal{D}_F\rvert = 120$	$\lvert\mathcal{D}_F\rvert = 240$
FineToCoarse	49.3 (-1.3, 1.4)	53.4 (-1.8, 1.7)	54.6 (-3.6, 3.8)	49.3 (-1.3, 1.4)	53.4 (-1.8, 1.7)	54.6 (-3.6, 3.8)
Coarse	49.6 (-1.5, 1.8)	49.6 (-1.5, 1.8)	49.6 (-1.5, 1.8)	**53.5 (-1.2, 1.4)**	53.5 (-1.2, 1.4)	53.5 (-1.2, 1.4)
Cascaded	39.7 (-6.8, 5.7)	45.4 (-3.1, 2.9)	42.6 (-6.5, 6.5)	**55.6 (-2.9, 2.7)**	55.0 (-3.2, 3.4)	**56.8 (-3.8, 3.6)**
Interpolated	**54.3 (-1.4, 1.4)**	**55.0 (-1.7, 1.6)**	**57.5 (-4.1, 5.2)**	**56.0 (-2.4, 2.1)**	54.5 (-2.9, 2.8)	**59.1 (-2.8, 3.4)**

Table 1: Sentence level results for varying numbers of fully labeled ($\mathcal{D}_F$) and coarsely labeled ($\mathcal{D}_C$) reviews. Bold: significantly better than the FineToCoarse model according to a hierarchical bootstrapped confidence interval, $p < 0.05$.

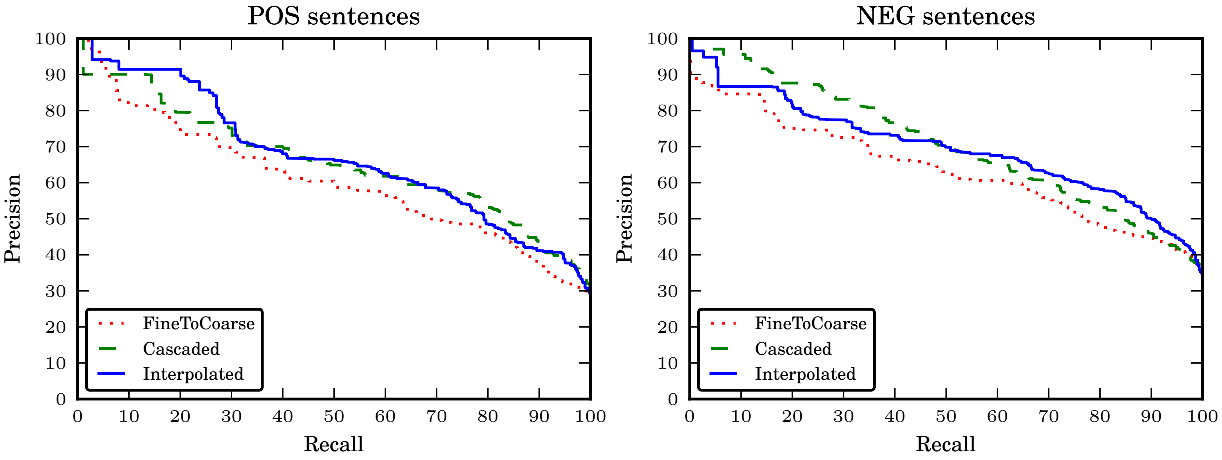

Figure 2: Interpolated POS / NEG sentence-level precision-recall curves with $\lvert\mathcal{D}_C\rvert = 143{,}580$ and $\lvert\mathcal{D}_F\rvert = 240$.

	POS docs.	NEG docs.	NEU docs.
FineToCoarse	35 / 11 / 59	33 / 76 / 42	29 / 63 / 55
Coarse	70 / 14 / 43	11 / 71 / 34	43 / 47 / 53
Cascaded	43 / 17 / 61	0 / 75 / 49	10 / 64 / 50
Interpolated	73 / 16 / 51	42 / 72 / 48	54 / 52 / 57

Table 2: POS / NEG / NEU sentence-level F_1-scores per document category ($\lvert\mathcal{D}_C\rvert = 143{,}580$ and $\lvert\mathcal{D}_F\rvert = 240$).

4 Conclusions

Learning fine-grained classification tasks in a fully supervised manner does not scale well due to the lack of naturally occurring supervision. We instead proposed to combine coarse-grained supervision, which is naturally abundant but less informative, with fine-grained supervision, which is scarce but more informative. To this end, we introduced two simple, yet effective, methods of combining fully labeled and coarsely labeled data for sentence-level sentiment analysis. First, a cascaded approach where a coarsely supervised model is used to generate features for a fully supervised model. Second, an interpolated model that directly optimizes a combination of joint and marginal likelihood functions. Both proposed models are structured conditional models that allow for rich overlapping features, while maintaining highly efficient exact inference and robust estimation properties. Empirically, the interpolated model is superior to the other investigated models, but with sufficient amounts of coarsely labeled and fully labeled data, the cascaded approach is competitive.

Acknowledgments

The first author acknowledges the support of the Swedish National Graduate School of Language Technology (GSLT). The authors would also like to thank Fernando Pereira and Bob Carpenter for early discussions on using HCRFs in sentiment analysis.

References

Arvind Agarwal and Hal Daumé. 2009. Exponential family hybrid semi-supervised learning. In *Proceedings of the International Jont conference on Artifical Intelligence (IJCAI)*.

Anthony C. Davison and David V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK.

Andrea Esuli and Fabrizio Sebastiani. 2009. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Language Resource and Evaluation Conference (LREC)*.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Julia A. Lasserre, Christopher M. Bishop, and Thomas P. Minka. 2006. Principled hybrids of generative and discriminative models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the Conference on Information and Knowledge Management (CIKM)*.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the International Conference on World Wide Web (WWW)*.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis*. Now Publishers.

Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. 2007. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2010. Incorporating content structure into text analysis applications. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jun Suzuki, Akinori Fujino, and Hideki Isozaki. 2007. Semi-supervised structured output learning based on a hybrid generative and discriminative approach. In *Porceedings of the Conference on Emipirical Methods in Natural Language Processing (EMNLP)*.

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the European Conference on Information Retrieval (ECIR)*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the Annual World Wide Web Conference (WWW)*.

Peter Turney. 2002. Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation (LREC)*.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.