Improving Claim Stance Classification with Lexical Knowledge Expansion and Context Utilization

Roy Bar-Haim
IBM Research - Haifa
Mount Carmel
Haifa, 3498825, Israel
roybar@il.ibm.com

Lilach Edelstein
IBM Research - Haifa
Mount Carmel
Haifa, 3498825, Israel
lilache@il.ibm.com

Charles Jochim
IBM Research - Ireland
Damastown Industrial Estate
Dublin 15, Ireland
charlesj@ie.ibm.com

Noam Slonim
IBM Research - Haifa
Mount Carmel
Haifa, 3498825, Israel
noams@il.ibm.com

Abstract

Stance classification is a core component in on-demand argument construction pipelines. Previous work on *claim stance classification* relied on background knowledge such as manually-composed sentiment lexicons. We show that both accuracy and coverage can be significantly improved through automatic expansion of the initial lexicon. We also developed a set of contextual features that further improves the state-of-the-art for this task.

1 Introduction

Debating technologies aim to help humans debate and make better decisions. A core capability for these technologies is the on-demand construction of pro and con arguments for a given controversial topic. Most previous work was aimed at detecting topic-dependent argument components, such as claims and evidence (Levy et al., 2014; Rinott et al., 2015). Recently, Bar-Haim et al. (2017) introduced the related task of *claim stance classification*. For example, given the topic

(1) ***The monarchy*** *should be abolished.*$\ominus$

and the following two claims

(2) ***Social traditions or hierarchies*** *are essential for social order.* $\oplus$ $\Leftrightarrow$

(3) *People feel greater dignity when* ***choosing their head of state****.* $\oplus$ $\not\Leftrightarrow$

the goal is to classify (2) as *Con* and (3) as *Pro* with respect to (1).

Bar-Haim et al. proposed a model that breaks this task into several sub-tasks: (a) Identify the sentiment targets of the topic and the claim (b) Determine the sentiment of the topic and the claim towards their sentiment targets, and (c) Determine the relation between the targets. Target A is *consistent/contrastive* with target B if the stance towards A implies the same/opposite stance towards B, respectively.

In (1)–(3), targets are marked in bold, positive/negative sentiment is indicated as $\oplus$/$\ominus$ and consistent/contrastive relation is marked as $\Leftrightarrow$/$\not\Leftrightarrow$. For instance, (3) has positive sentiment towards its target, *choosing their head of state*, which implies negative sentiment towards *the monarchy*, since the targets are contrastive. The topic's sentiment towards *the monarchy* is also negative, hence it is a *Pro* claim.

On-demand argument generation is inherently an open-domain task, so one cannot learn topic-specific features for stance classification from the training data. Furthermore, claims are short sentences, and the number of claims in the training data is relatively small as compared to common sentiment analysis and stance classification benchmarks. Consequently, external knowledge such as sentiment lexicons is crucial for this task. However, the coverage of manually-constructed sentiment lexicons is often incomplete. As reported by Bar-Haim et al., the sentiment lexicon they used was able to match sentiment terms in fewer than 80% of the claims. Moreover, manually composed sentiment lexicons lack the notion of (numeric) sentiment strength.

A more general limitation of sentiment-based approaches is that some claims express stance but do not convey explicit sentiment. As an example,

Proceedings of the 4th Workshop on Argument Mining, pages 32–38
Copenhagen, Denmark, September 8, 2017. ©2017 Association for Computational Linguistics

consider the following *Pro* claim for (1):

(4) *The people, not the members of one family, should be sovereign.*

In this work we present several improvements to the system of Bar-Haim et al. (2017) (henceforth, the *baseline system*), which address the above limitations. First, we present a method for automatic expansion of a given sentiment lexicon, which leads to a substantial performance increase. Second, while the baseline system only considers the claim itself, we developed a set of contextual features that further boosts the performance of the system. In particular, these contextual features allow classification of claims with no explicit sentiment. Overall, we outperformed the best published results for this task by a large margin.

2 Baseline System

We first give a high-level description of the Bar-Haim et al. system, which we build upon in this work. Given a topic t and a claim c, let x_t and x_c be their sentiment targets, respectively, and let s_t, $s_c \in [-1, 1]$ be the sentiment of the topic and the claim towards their respective targets. Positive/negative values indicate positive/negative sentiment. Let $\mathcal{R}(x_c, x_t) \in [-1, 1]$ denote the relation between the claim target and the topic target. Positive/negative values indicate consistent/contrastive targets (as defined in the previous section). The absolute value of both scores indicates confidence. The stance of c towards t is predicted as:

$$Stance(c,t) = s_c \times \mathcal{R}(x_c, x_t) \times s_t \qquad (1)$$

Positive/negative prediction indicates *Pro*/*Con* stance. As before, the absolute value indicates confidence. Having an effective confidence measure is important for on-demand argument construction, where we typically want to present to the user only high-confidence predictions, or rank them higher in the output.

Bar-Haim et al. assumed that the topic target x_t and sentiment s_t are given as input, and developed three classifiers for predicting x_c, s_c and $\mathcal{R}(x_c, x_t)$. The system predicts the stance of the claim c towards the given topic target x_t (e.g., *the monarchy*) as $s_c \times \mathcal{R}(x_c, x_t)$. The result is multiplied by the given topic target sentiment s_t to obtain $Stance(c, t)$.[1]

Most relevant to our work is the sentiment classifier, which predicts the sentiment s_c towards the target x_c. It is based on matching sentiment terms from a lexicon, detecting polarity flips by sentiment shifters, and aggregating sentiment scores for matched terms, which decay based on their distance from the target.

The claim stance classification dataset introduced by Bar-Haim et al. includes 2,394 claims, manually found in Wikipedia articles for 55 topics, their stance (*Pro/Con*), and fine-grained annotations for targets (x_t, x_c), sentiments (s_t, s_c) and target relations ($\mathcal{R}(x_c, x_t)$).

In this dataset, 94.4% of the claims were found to be compatible with the above modeling, out of which 20% of the claims have contrastive targets. Since identifying contrastive targets with high precision is hard, the implemented relation classifier only predicts $\mathcal{R}(x_c, x_t) \in [0, 1]$, (i.e., always predicts *consistent*). Even so, multiplying by the classifier's confidence improves the accuracy of top predictions, since it ranks claims with consistent targets higher; this reduces stance classification errors caused by contrastive targets.

3 Lexicon Expansion

To obtain a wide-coverage sentiment lexicon that also includes weak sentiment, we took the following approach. Given a seed lexicon, we trained a classifier to predict the sentiment polarity for unseen words. We trained the classifier over the words in the lexicon, where the feature vector was the word embedding and the label was its polarity.

We started with the Opinion Lexicon (Hu and Liu, 2004), used in the baseline system, as a seed sentiment lexicon containing 6,789 words. For word embeddings, we trained a skip-gram model (Mikolov et al., 2013) over Wikipedia, using `word2vec`. With the 200-dimensional word embedding feature vectors and labels from the lexicon, we trained a linear SVM classifier (LIBLINEAR, Fan et al., 2008). Following Rothe et al. (2016), we only trained on high-frequency words (4,861 words with frequency > 300).

We checked the classifier's accuracy with a leave-one-out experiment over the original lexicon. For each word in the lexicon, which also had a word embedding (6,438 words), we trained our classifier on the remaining frequent words and tested the prediction of the held-out word. The resulting accuracy was 90.5%.

[1] For example, a claim in favor of *the monarchy* is *Pro* for *"The monarchy should be preserved"*, and *Con* for *"The monarchy should be abolished"* with s_t=+1/-1, respectively.

After removing single character terms and terms containing non-alphabetic characters, we predicted sentiment for the remaining 938,559 terms with word embeddings. The predicted SVM scores are roughly in $[-3, 3]$, and we adapted max-min scaling to return sentiment scores in $[-1, 1]$ (the sentiment scores in the seed lexicon are either 1 or -1).

To obtain a more compact lexicon, we applied a filtering step using WordNet relations (Miller, 1995; Fellbaum, 1998). For each term in the expanded lexicon, we looked up all its synsets. Then, for each of those synsets we collected all terms in the synset along with the terms that are derivationally related, hypernyms, or antonyms. Next, we looked up each of the terms from this collection in the seed lexicon and counted the number of positive and negative matches (the polarity of the antonyms was reversed). If the term had no matches, or the majority count did not agree with the SVM prediction, the term was discarded. This filter drastically reduced the expanded lexicon size to only 28,670 terms (including the seed lexicon), while achieving similar performance on the stance classification task.

4 Contextual Features

Following the assumption that neighboring texts tend to agree on sentiment, we enhanced the system to use the claim's context.

We trained a linear SVM classifier, which includes the baseline system (with the expanded lexicon) as a feature, together with a set of contextual features, described below. Similar to the baseline system, the classifier aims to predict the stance towards the topic target x_t, and the result is multiplied by the given s_t to obtain $Stance(c, t)$.[2]

We employed the following features.

Header Features: Each article in Wikipedia is divided into titled sections, subsections and sub-subsections. We assume the sentiment is shared by the section header and the claims presented in the section. For example, a claim under the *"Criticism"* section is usually of negative sentiment, while the header *"Advantages"* would govern positive claims. We considered the headers of the claim's enclosing section, subsection and sub-subsection. The sentiment of each header was taken as a feature. In addition, we performed a Fisher Exact Test (Agresti, 1992) on the training data and composed two short lists of prevalent header words that were found to be the most significantly associated with positive (or negative) claims in their sections. The difference between the number of positive and negative words appearing in the claim's enclosing headers was taken as an additional feature.[3]

Claim Sentence: In some cases, the claim's enclosing sentence contains helpful cues for the claim polarity (e.g., in: *"Unfortunately, it's clear that <claim>"*). Therefore, the sentiment score of the entire sentence also served as a feature.[4]

Neighboring Sentences: We computed the average sentiment score of sentences preceding and following the claim sentence in the same paragraph. Specifically, we considered the maximal set of consecutive sentences that do not contain contrastive discourse markers and terms indicating controversy (listed in Table 1, row 2). If the claim sentence itself contained certain terms indicating contrast or controversy (Table 1, row 1), the context was ignored and the feature value was set to zero.

Neighboring Claims: Neighboring claims tend to agree on sentiment : in article sections that include more than one claim in our training data, 88% of the claims shared the majority polarity. Thus, we clustered the claims so that each pair in the same paragraph shared a cluster unless a term indicating potential polarity flip was found before the two claims or between them. The polarity flip indicators considered between/before the claims are listed in Table 1, rows 2/3, respectively. For example, consider the following claim pairs:

(5) While **adoption can provide stable families to children in need**, it is also suggested that **adoption in the immediate aftermath of a trauma might not be the best option**.

(6) **Democracy is far from perfect**. However, **it's the best form of government created so far**.

In both cases, the underlined discourse marker indicates a polarity shift between the claims (shown in **bold**), so the claims are not clustered together. For each claim, we summed the sentiment scores

[2] Accordingly, the training labels were $\frac{Stance(c,t)}{s_t}$.

[3] The positive words are *support, benefit, overview, pro, growth, reform*, and the negative words are *criticism, anti, failure, abuse, dissent, corrupt, opposite, disadvantage*.

[4] Since the whole sentence is likely to have the same target x_c as the claim itself, we multiplied this feature by the consistent/contrastive relation score $\mathcal{R}(x_c, x_t)$.

#	Context	Terms
1	Claim Sentence	though, although, even if, dispute, but, while, challenge, criticize, incorrect, wrong, however
2	Surrounding Sentences/ Between Claims	dispute, disagree, although, though, nevertheless, otherwise, but, nonetheless, notwithstanding, in contrast, after all, opponent[s] claim, however, on the other hand, on the contrary, contend
3	Before Claims	though, although, even if, dispute, but, while

Table 1: Contrast and controversy indicators considered for each context type by the *neighboring sentences* feature (rows 1+2), and the *neighboring claims* feature (rows 2+3).

over all other claims in its cluster. Note that this feature requires additional information about other claims for the topic.

5 Evaluation

We followed the experimental setup of Bar-Haim et al., including the train/test split of the dataset and the evaluation measures, and predicted the majority class in the train set with a constant, very low confidence when the classifier's output was zero. The training set contained 25 topics (1,039 claims), and the test set contained 30 topics (1,355 claims).

The evaluation explored the trade-off between *accuracy* (fraction of correct stance predictions) and *coverage* (fraction of claims for which we make a non-zero prediction). This tradeoff was controlled by setting a minimum confidence threshold for making a prediction. Given a coverage level β, Accuracy@β is defined as the maximal accuracy such that the corresponding coverage is at least β, found by exhaustive search over the threshold values. Coverage and accuracy for each threshold are macro-averaged over the tested topics.

The results are summarized in Table 2. Rows (1-2) quote the two best-performing configurations reported by Bar-Haim et al. The first is the baseline configuration used in this work, which performed best on lower coverage rates. The second is a combination of the baseline system and an SVM with unigram features, which was the best performer on higher coverage rates. Row 3 is our rerun of the baseline system. The results are close to the EACL '17 results (row 1) but not identical. This is due to some changes in low-level tools used by the system, such as the wikifier.[5]

The configurations in rows 4-6 are the contributions of this work. Row 4 reports the results for the baseline system with the expanded lexicon (Section 3). Like the baseline system, this configuration only considers the claim itself. The results show substantial improvements over the baseline (row 3), as well as the best previously reported results (rows 1-2). The expanded lexicon increased the (macro-averaged) coverage of the system from 78.2% to 98.1%.

The next two configurations use increasingly richer contexts, in addition to using the expanded lexicon. Row 5 shows the results for the classifier described in Section 4, using all the contextual features except for the *neighboring claims* feature. We refer to this feature set as *local contextual features*. The results show that these features achieve further improvement.

Last, row 6 shows the results for adding the *neighboring claims* feature, which achieves the best results. This configuration requires additional knowledge about other claims in the proximity of the given claim. While in this experiment the labeled data provides perfect knowledge about neighboring claims, in actual implementations of argument construction pipelines this information is obtained from the imperfect output of a claim detection module.

Overall, our results represent significant advancement of the state-of-the-art for this task, both for lower coverage rates (top predictions) and over the whole dataset (Accuracy@1.0).

6 Related Work

Stance classification has been applied to several different means of argumentation, for example congressional debates (Thomas et al., 2006; Yessenalina et al., 2010) or online discussions (Somasundaran and Wiebe, 2009; Walker et al., 2012b; Hasan and Ng, 2013). Some previous

[5] As explained by Bar-Haim et al. (2017), the baseline results (rows 1,3) for each coverage level$\geq$ 0.8 are the same, since they all add the default majority class predictions.

		Accuracy@Coverage									
#	Configuration	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1	Baseline (EACL'17)	0.849	0.847	0.836	0.793	0.767	0.740	0.704	0.632	0.632	0.632
2	Baselne+SVM (EACL'17)	0.784	0.758	0.749	0.743	0.730	0.711	0.682	0.671	0.658	0.645
3	Baseline (Rerun)	0.846	0.841	0.823	0.787	0.771	0.742	0.706	0.633	0.633	0.633
4	+Lexicon Expansion	0.899	0.867	0.844	0.803	0.765	0.749	0.731	0.705	0.697	0.677
5	+Local Contextual Features	0.935	0.892	0.866	0.833	0.805	0.773	0.749	0.729	0.704	0.690
6	+Neighboring Claims	0.954	0.935	0.882	0.856	0.811	0.776	0.764	0.734	0.708	0.691

Table 2: Stance classification results. Majority baseline Accuracy@1.0=51.9%

work has improved stance classification by using the conversation structure (e.g., discussion reply links) (Walker et al., 2012a; Sridhar et al., 2015) or by applying classification to groups of arguments linked by citations or agreement/disagreement (Burfoot et al., 2011; Sridhar et al., 2014). However, many features used in previous works were not available for our task. Instead, we leveraged other context information present in Wikipedia articles, and assume sentiment agreement across neighboring text fragments.

A number of approaches in the literature can generate sentiment lexicons (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003), many of which rely on graph-based approaches over WordNet (Hu and Liu, 2004; Esuli and Sebastiani, 2006; Blair-Goldensohn et al., 2008) or over a graph of distributionally similar n-grams (Velikovich et al., 2010). Our approach (Section 3) differs in that we leverage larger existing sentiment lexicons, instead of relying on small seed sets. Moreover, we opt for classifying word embeddings instead of graph-based approaches, which are sensitive to parameter settings.

More similar recent work includes Amir et al. (2015), who also used manually-created sentiment lexicons (annotated with discrete sentiment levels) and word embeddings to train linear regression models that aim to predict the polarity and intensity of new terms. Out of the tested methods, *Support Vector Regression* was found to perform best. However, they did not filter the resulting lexicon.

7 Conclusion

We addressed two of the main limitations of previous work on claim stance classification: insufficient coverage of manually-composed sentiment lexicons, and ignoring the claim's context. We presented a lexicon expansion method and a set of effective contextual features, which together significantly advance the state-of-the-art. A remaining challenge is accurate prediction of contrastive targets, which seems crucial for further substantial improvement over the whole dataset.

Acknowledgments

We would like to thank Francesco Dinuzzo for his contribution to the initial stages of this work. We also thank Francesca Bonin for her helpful feedback on this work.

References

Alan Agresti. 1992. A survey of exact inference for contingency tables. *Statistical science* pages 131–153.

Silvio Amir, Ramón Astudillo, Wang Ling, Bruno Martins, Mario J. Silva, and Isabel Trancoso. 2015. Inesc-id: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 613–618. http://www.aclweb.org/anthology/S15-2102.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 251–261. http://www.aclweb.org/anthology/E17-1024.

Sasha Blair-Goldensohn, Tyler Neylon, Kerry Hannan, George A. Reis, Ryan Mcdonald, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *NLP in the Information Explosion Era*.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 1506–1515. http://www.aclweb.org/anthology/P11-1151.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC '06)*. pages 417–422.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 1348–1356. http://www.aclweb.org/anthology/I13-1191.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Madrid, Spain, pages 174–181. https://doi.org/10.3115/976909.979640.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '04, pages 168–177. https://doi.org/10.1145/1014052.1014073.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 1489–1500. http://www.aclweb.org/anthology/C14-1141.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Curran Associates Inc., USA, NIPS'13, pages 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 440–450. http://aclweb.org/anthology/D15-1050.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 767–777. http://www.aclweb.org/anthology/N16-1091.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, pages 226–234. http://www.aclweb.org/anthology/P/P09/P09-1026.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 116–125. http://www.aclweb.org/anthology/P15-1012.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*. Association for Computational Linguistics, Baltimore, Maryland, pages 109–117. http://www.aclweb.org/anthology/W14-2715.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, pages 327–335. http://www.aclweb.org/anthology/W/W06/W06-1639.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.* 21(4):315–346. https://doi.org/10.1145/944012.944013.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Los Angeles, California, pages 777–785. http://www.aclweb.org/anthology/N10-1119.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012a. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 592–596. http://www.aclweb.org/anthology/N12-1072.

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. 2012b. That is your evidence?: Classifying stance in online political debate. *Decis. Support Syst.* 53(4):719–729. https://doi.org/10.1016/j.dss.2012.05.032.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, pages 1046–1056. http://www.aclweb.org/anthology/D10-1102.