A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis

Sebastian Ruder[1,2], **Parsa Ghaffari**[2], **and John G. Breslin**[1]

[1]Insight Centre for Data Analytics
National University of Ireland, Galway
{sebastian.ruder,john.breslin}@insight-centre.org
[2]Aylien Ltd.
Dublin, Ireland
{sebastian,parsa}@aylien.com

Abstract

Opinion mining from customer reviews has become pervasive in recent years. Sentences in reviews, however, are usually classified independently, even though they form part of a review's argumentative structure. Intuitively, sentences in a review build and elaborate upon each other; knowledge of the review structure and sentential context should thus inform the classification of each sentence. We demonstrate this hypothesis for the task of aspect-based sentiment analysis by modeling the interdependencies of sentences in a review with a hierarchical bidirectional LSTM. We show that the hierarchical model outperforms two non-hierarchical baselines, obtains results competitive with the state-of-the-art, and outperforms the state-of-the-art on five multilingual, multi-domain datasets without any hand-engineered features or external resources.

1 Introduction

Sentiment analysis (Pang and Lee, 2008) is used to gauge public opinion towards products, to analyze customer satisfaction, and to detect trends. With the proliferation of customer reviews, more fine-grained aspect-based sentiment analysis (ABSA) has gained in popularity, as it allows aspects of a product or service to be examined in more detail.

Reviews – just with any coherent text – have an underlying structure. A visualization of the discourse structure according to Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) for the example review in Figure 1 reveals that sentences and clauses are connected via different rhetorical relations, such as *Elaboration* and *Background*.

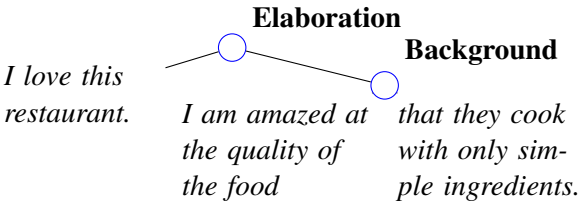

Figure 1: RST structure of an example review.

Intuitively, knowledge about the relations and the sentiment of surrounding sentences should inform the sentiment of the current sentence. If a reviewer of a restaurant has shown a positive sentiment towards the quality of the food, it is likely that his opinion will not change drastically over the course of the review. Additionally, overwhelmingly positive or negative sentences in the review help to disambiguate sentences whose sentiment is equivocal.

Neural network-based architectures that have recently become popular for sentiment analysis and ABSA, such as convolutional neural networks (Severyn and Moschitti, 2015), LSTMs (Vo and Zhang, 2015), and recursive neural networks (Nguyen and Shirai, 2015), however, are only able to consider intra-sentence relations such as *Background* in Figure 1 and fail to capture inter-sentence relations, e.g. *Elaboration* that rely on discourse structure and provide valuable clues for sentiment prediction.

We introduce a hierarchical bidirectional long short-term memory (H-LSTM) that is able to leverage both intra- and inter-sentence relations. The sole dependence on sentences and their structure

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 999–1005,
Austin, Texas, November 1-5, 2016.

within a review renders our model fully language-independent. We show that the hierarchical model outperforms strong sentence-level baselines for aspect-based sentiment analysis, while achieving results competitive with the state-of-the-art and outperforming it on several datasets without relying on any hand-engineered features or sentiment lexica.

2 Related Work

Aspect-based sentiment analysis. Past approaches use classifiers with expensive hand-crafted features based on n-grams, parts-of-speech, negation words, and sentiment lexica (Pontiki et al., 2014; Pontiki et al., 2015). The model by Zhang and Lan (2015) is the only approach we are aware of that considers more than one sentence. However, it is less expressive than ours, as it only extracts features from the preceding and subsequent sentence without any notion of structure. Neural network-based approaches include an LSTM that determines sentiment towards a target word based on its position (Tang et al., 2015) as well as a recursive neural network that requires parse trees (Nguyen and Shirai, 2015). In contrast, our model requires no feature engineering, no positional information, and no parser outputs, which are often unavailable for low-resource languages. We are also the first – to our knowledge – to frame sentiment analysis as a sequence tagging task.

Hierarchical models. Hierarchical models have been used predominantly for representation learning and generation of paragraphs and documents: Li et al. (2015) use a hierarchical LSTM-based autoencoder to reconstruct reviews and paragraphs of Wikipedia articles. Serban et al. (2016) use a hierarchical recurrent encoder-decoder with latent variables for dialogue generation. Denil et al. (2014) use a hierarchical ConvNet to extract salient sentences from reviews, while Kotzias et al. (2015) use the same architecture to learn sentence-level labels from review-level labels using a novel cost function. The model of Lee and Dernoncourt (2016) is perhaps the most similar to ours. While they also use a sentence-level LSTM, their class-level feed-forward neural network is only able to consider a limited number of preceding texts, while our review-level bidirectional LSTM is (theoretically) able to consider an unlimited number of preceding *and* successive sentences.

3 Model

In the following, we will introduce the different components of our hierarchical bidirectional LSTM architecture displayed in Figure 2.

3.1 Sentence and Aspect Representation

Each review consists of sentences, which are padded to length l by inserting padding tokens. Each review in turn is padded to length h by inserting sentences containing only padding tokens. We represent each sentence as a concatentation of its word embeddings $x_{1:l}$ where $x_t \in \mathbb{R}^k$ is the k-dimensional vector of the t-th word in the sentence.

Every sentence is associated with an aspect. Aspects consist of an entity and an attribute, e.g. `FOOD#QUALITY`. Similarly to the entity representation of Socher et al. (2013), we represent every aspect a as the average of its entity and attribute embeddings $\frac{1}{2}(x_e + x_a)$ where $x_e, x_a \in \mathbb{R}^m$ are the m-dimensional entity and attribute embeddings respectively[1].

3.2 LSTM

We use a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), which adds input, output, and forget gates to a recurrent cell, which allow it to model long-range dependencies that are essential for capturing sentiment.

For the t-th word in a sentence, the LSTM takes as input the word embedding x_t, the previous output h_{t-1} and cell state c_{t-1} and computes the next output h_t and cell state c_t. Both h and c are initialized with zeros.

3.3 Bidirectional LSTM

Both on the review and on the sentence level, sentiment is dependent not only on preceding but also successive words and sentences. A Bidirectional LSTM (Bi-LSTM) (Graves et al., 2013) allows us to look ahead by employing a forward LSTM, which processes the sequence in chronological order, and a backward LSTM, which processes the sequence in reverse order. The output h_t at a given time step is then the concatenation of the corresponding states of the forward and backward LSTM.

[1] Averaging embeddings produced slightly better results than using a separate embedding for every aspect.

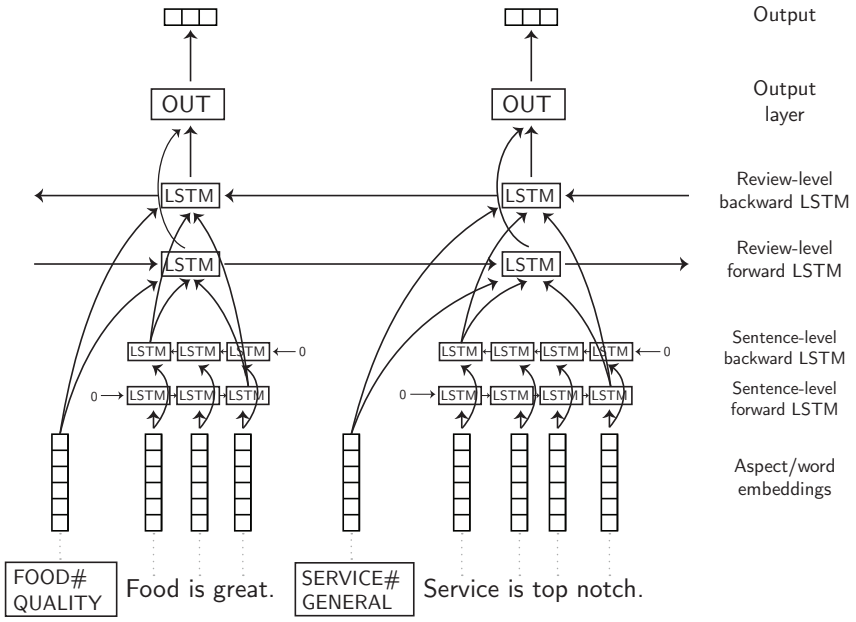

Figure 2: The hierarchical bidirectional LSTM (H-LSTM) for aspect-based sentiment analysis. Word embeddings are fed into a sentence-level bidirectional LSTM. Final states of forward and backward LSTM are concatenated together with the aspect embedding and fed into a bidirectional review-level LSTM. At every time step, the output of the forward and backward LSTM is concatenated and fed into a final layer, which outputs a probability distribution over sentiments.

3.4 Hierarchical Bidirectional LSTM

Stacking a Bi-LSTM on the review level on top of sentence-level Bi-LSTMs yields the hierarchical bidirectional LSTM (H-LSTM) in Figure 2.

The sentence-level forward and backward LSTMs receive the sentence starting with the first and last word embedding x_1 and x_l respectively. The final output h_l of both LSTMs is then concatenated with the aspect vector a[2] and fed as input into the review-level forward and backward LSTMs. The outputs of both LSTMs are concatenated and fed into a final softmax layer, which outputs a probability distribution over sentiments[3] for each sentence.

4 Experiments

4.1 Datasets

For our experiments, we consider datasets in five domains (restaurants, hotels, laptops, phones, cameras) and eight languages (English, Spanish, French, Russian, Dutch, Turkish, Arabic, Chinese) from the recent SemEval-2016 Aspect-based Sentiment Analysis task (Pontiki et al., 2016), using the provided train/test splits. In total, there are 11 domain-language datasets containing 300-400 reviews with 1250-6000 sentences[4]. Each sentence is annotated with none, one, or multiple domain-specific aspects and a sentiment value for each aspect.

4.2 Training Details

Our LSTMs have one layer and an output size of 200 dimensions. We use 300-dimensional word embeddings. We use pre-trained GloVe (Pennington et al., 2014) embeddings for English, while we train embeddings on `frWaC`[5] for French and on the Leipzig Corpora Collection[6] for all other languages.[7] Entity

[2]We experimented with other interactions, e.g. rescaling the word embeddings by their aspect similarity, an attention-like mechanism, as well as summing and multiplication, but found that simple concatenation produced the best results.

[3]The sentiment classes are *positive*, *negative*, and *neutral*.

[4]Exact dataset statistics can be seen in (Pontiki et al., 2016).

[5]`http://wacky.sslmit.unibo.it/doku.php?id=corpora`

[6]`http://corpora2.informatik.uni-leipzig.de/download.html`

[7]Using 64-dimensional Polyglot embeddings (Al-Rfou et al., 2013) yielded generally worse performance.

Language	Domain	Best	XRCE	IIT-TUDA	CNN	LSTM	H-LSTM	HP-LSTM
English	Restaurants	**88.1**	**88.1**	86.7	82.1	81.4	83.0	85.3
Spanish	Restaurants	**83.6**	-	**83.6**	79.6	75.7	79.5	81.8
French	Restaurants	**78.8**	**78.8**	72.2	73.2	69.8	73.6	75.4
Russian	Restaurants	77.9	-	73.6	75.1	73.9	**78.1**	77.4
Dutch	Restaurants	77.8	-	77.0	75.0	73.6	82.2	**84.8**
Turkish	Restaurants	**84.3**	-	**84.3**	74.2	73.6	76.7	79.2
Arabic	Hotels	82.7	-	81.7	82.7	80.5	82.8	**82.9**
English	Laptops	**82.8**	-	**82.8**	78.4	76.0	77.4	80.1
Dutch	Phones	83.3	-	82.6	83.3	81.8	81.3	**83.6**
Chinese	Cameras	**80.5**	-	-	78.2	77.6	78.6	78.8
Chinese	Phones	73.3	-	-	72.4	70.3	**74.1**	73.3

Table 1: Results of our system with randomly initialized word embeddings (H-LSTM) and with pre-trained embeddings (HP-LSTM) for ABSA for each language and domain in comparison to the best system for each pair (Best), the best two single systems (XRCE, IIT-TUDA), a sentence-level CNN (CNN), and our sentence-level LSTM (LSTM).

and attribute embeddings of aspects have 15 dimensions and are initialized randomly. We use dropout of 0.5 after the embedding layer and after LSTM cells, a gradient clipping norm of 5, and no l_2 regularization.

We unroll the aspects of every sentence in the review, e.g. a sentence with two aspects occurs twice in succession, once with each aspect. We remove sentences with no aspect[8] and ignore predictions for all sentences that have been added as padding to a review so as not to force our model to learn meaningless predictions, as is commonly done in sequence-to-sequence learning (Sutskever et al., 2014). We segment Chinese data before tokenization.

We train our model to minimize the cross-entropy loss, using stochastic gradient descent, the Adam update rule (Kingma and Ba, 2015), mini-batches of size 10, and early stopping with a patience of 10.

4.3 Comparison models

We compare our model using random (H-LSTM) and pre-trained word embeddings (HP-LSTM) against the best model of the SemEval-2016 Aspect-based Sentiment Analysis task (Pontiki et al., 2016) for each domain-language pair (Best) as well as against the two best single models of the competition: IIT-TUDA (Kumar et al., 2016), which uses large sentiment lexicons for every language, and XRCE (Brun et al., 2016), which uses a parser augmented with hand-crafted, domain-specific rules. In order to ascertain that the hierarchical nature of our model is the deciding factor, we additionally compare against the sentence-level convolutional neural network of Ruder et al. (2016) (CNN) and against a sentence-level Bi-LSTM (LSTM), which is identical to the first layer of our model.[9]

5 Results and Discussion

We present our results in Table 1. Our hierarchical model achieves results superior to the sentence-level CNN and the sentence-level Bi-LSTM baselines for almost all domain-language pairs by taking the structure of the review into account. We highlight examples where this improves predictions in Table 2.

In addition, our model shows results competitive with the best single models of the competition, while requiring no expensive hand-crafted features or external resources, thereby demonstrating its language and domain independence. Overall, our model compares favorably to the state-of-the-art, particularly for low-resource languages, where few hand-engineered features are available. It outperforms the state-of-the-art on four and five datasets using randomly initialized and pre-trained embeddings respectively.

[8] Labeling them with a NONE aspect and predicting *neutral* slightly decreased performance.

[9] To ensure that the additional parameters do not account for the difference, we increase the number of layers and dimensions of LSTM, which does not impact the results.

Id	Sentence	LSTM	H-LSTM
1.1	No Comparison	*negative*	*positive*
1.2	It has great sushi and even better service.	*positive*	*positive*
2.1	Green Tea creme brulee is a must!	*positive*	*positive*
2.2	Don't leave the restaurant without it.	*negative*	*positive*

Table 2: Example sentences where knowledge of other sentences in the review (not necessarily neighbors) helps to disambiguate the sentiment of the sentence in question. For the aspect in 1.1, the sentence-level LSTM predicts *negative*, while the context of the service and food quality in 1.2 allows the H-LSTM to predict *positive*. Similarly, for the aspect in 2.2, knowledge of the quality of the green tea crème brulée helps the H-LSTM to predict the correct sentiment.

5.1 Pre-trained embeddings

In line with past research (Collobert et al., 2011), we observe significant gains when initializing our word vectors with pre-trained embeddings across almost all languages. Pre-trained embeddings improve our model's performance for all languages except Russian, Arabic, and Chinese and help it achieve state-of-the-art in the Dutch phones domain. We release our pre-trained multilingual embeddings so that they may facilitate future research in multilingual sentiment analysis and text classification[10].

5.2 Leveraging additional information

As annotation is expensive in many real-world applications, learning from only few examples is important. Our model was designed with this goal in mind and is able to extract additional information inherent in the training data. By leveraging the structure of the review, our model is able to inform and improve its sentiment predictions as evidenced in Table 2.

The large performance differential to the state-of-the-art for the Turkish dataset where only 1104 sentences are available for training and the performance gaps for high-resource languages such as English, Spanish, and French, however, indicate the limits of an approach such as ours that only uses data available at training time.

While using pre-trained word embeddings is an effective way to mitigate this deficit, for high-resource languages, solely leveraging unsupervised language information is not enough to perform on-par with approaches that make use of large external resources (Kumar et al., 2016) and meticulously hand-crafted features (Brun et al., 2016).

Sentiment lexicons are a popular way to inject additional information into models for sentiment analysis. We experimented with using sentiment lexicons by Kumar et al. (2016) but were not able to significantly improve upon our results with pre-trained embeddings[11]. In light of the diversity of domains in the context of aspect-based sentiment analysis and many other applications, domain-specific lexicons (Hamilton et al., 2016) are often preferred. Finding better ways to incorporate such domain-specific resources into models as well as methods to inject other forms of domain information, e.g. by constraining them with rules (Hu et al., 2016) is thus an important research avenue, which we leave for future work.

6 Conclusion

In this paper, we have presented a hierarchical model of reviews for aspect-based sentiment analysis. We demonstrate that by allowing the model to take into account the structure of the review and the sentential context for its predictions, it is able to outperform models that only rely on sentence information and achieves performance competitive with models that leverage large external resources and hand-engineered features. Our model achieves state-of-the-art results on 5 out of 11 datasets for aspect-based sentiment analysis.

Acknowledgments

We thank the anonymous reviewers, Nicolas Pécheux, and Hugo Larochelle for their constructive feedback. This publication has emanated from research conducted with the financial support of the Irish Research Council (IRC) under Grant Number EBPPG/2014/30 and with Aylien Ltd. as Enterprise Partner as well as from research supported by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

[10] https://s3.amazonaws.com/aylien-main/data/multilingual-embeddings/index.html

[11] We tried bucketing and embedding of sentiment scores as well as filtering and pooling as in (Vo and Zhang, 2015)

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.

Caroline Brun, Julien Perez, and Claude Roux. 2016. XRCE at SemEval-2016 Task 5: Feedbacked Ensemble Modelling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 282–286.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2014. Extraction of Salient Sentences from Labelled Documents. *arXiv preprint arXiv:1412.6815*, pages 1–9.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (3):6645–6649.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing Deep Neural Networks with Logic Rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1–18.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13.

Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From Group to Individual Labels using Deep Features. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597—-606.

Ayush Kumar, Sarah Kohail, Amit Kumar, Asif Ekbal, and Chris Biemann. 2016. IIT-TUDA at SemEval-2016 Task 5: Beyond Sentiment Lexicon: Combining Domain Dependency and Distributional Semantics Features for Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. *Proceedings of NAACL-HLT 2016*.

Jiwei Li, Minh-Thang Luong, and Daniel Jurafsky. 2015. A Hierarchical Neural Autoencoder for Paragraphs and Documents. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1106–1115.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (September):2509–2514.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeny Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect-Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 1–14.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. *Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 1–10.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, page 9.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Target-Dependent Sentiment Classification with Long Short Term Memory. *arXiv preprint arXiv:1512.01100*.

Duy-tin Vo and Yue Zhang. 2015. Target-Dependent Twitter Sentiment Classification with Rich Automatic Features. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1347–1353.

Zhihua Zhang and Man Lan. 2015. ECNU: Extracting Effective Features from Multiple Sequential Sentences for Target-dependent Sentiment Analysis in Reviews. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 736–741.