Aspect Category Detection via Topic-Attention Network

Sajad Movahedi, Erfan Ghadery, Heshaam Faili, Azadeh Shakery
University of Tehran
Tehran, Iran
{s.movahedi, erfan.ghadery, hfaili, shakery}@ut.ac.ir

Abstract

The e-commerce has started a new trend in natural language processing through sentiment analysis of user-generated reviews. Different consumers have different concerns about various aspects of a specific product or service. Aspect category detection, as a subtask of aspect-based sentiment analysis, tackles the problem of categorizing a given review sentence into a set of pre-defined aspect categories. In recent years, deep learning approaches have brought revolutionary advances in multiple branches of natural language processing including sentiment analysis. In this paper, we propose a deep neural network method based on attention mechanism to identify different aspect categories of a given review sentence. Our model utilizes several attentions with different topic contexts, enabling it to attend to different parts of the review sentence based on different topics. Experimental results on two datasets in the restaurant domain released by SemEval workshop demonstrate that our approach outperforms existing methods on both datasets. Visualization of the topic attention weights shows the effectiveness of our model in identifying words related to different topics.

Introduction

User-generated reviews in e-commerce websites are valuable resources for both the consumers and the producers of products or services. For a potential consumer, the experiences of other consumers are helpful in making educated decisions before purchasing a product or service. On the other hand, such data can help the producers of these products or services in refining the quality of what they offer. Different customers can have different concerns about the same product. This issue raises the challenge of categorizing the reviews into pre-defined aspects of the entity under review. This challenge is tackled by aspect category detection, a subtask of aspect-based sentiment analysis. Given a review sentence, aspect category detection aims to categorize the sentence into a set of pre-defined categories like ‘FOOD’, ‘PRICE’ etc. in the restaurant review domain. For example, the sentence “It is very overpriced and not very tasty.” belongs to both ‘FOOD’ and ‘PRICE’ aspect categories.

Previous research works in this field can be divided into five approaches: frequency-based, syntax-based, supervised machine learning, unsupervised machine learning, and hybrid (Schouten and Frasincar 2016). The majority of the proposed approaches are machine learning based including classic algorithms such as SVM and Maximum Entropy (Xenos et al. 2016), (Hercig et al. 2016), and deep neural network based approaches (Tjo and Su 2016), (Xue et al. 2017). Due to the multi-label nature of this task, most of the approaches utilize the one-vs-all classification models. These methods have shown good results in performing aspect category detection. However, training several one-vs-all classifiers require a lot of resource and time, especially when there are numerous categories. Another issue is raised by the fact that for different aspect categories, different words may contribute variously. For example, words like ‘waiter’, ‘staff’, and ‘atmosphere’ have little or no relevance to the ‘FOOD’ category, while words like ‘delicious’, ‘salty’, and ‘shrimp’ do. Therefore, a mechanism that attends to different parts of a given review sentence based on different topics may help the performance of the model. In this paper, instead of training several one-vs-all models, we propose a single model, namely Topic-Attention Network (TAN), for aspect category detection.

The intuition behind TAN is that given different categories, a model should attend on different words of a given review sentence. Therefore, TAN utilizes multiple attentions with different contexts as topics to attend to different words, given their different topics. Firstly, using a bi-directional GRU layer, we obtain encoded representations of each word. Secondly, these representations are fed to the topic attention layer to acquire attentive representations of the sentence. The topic attention layer, for each topic, gives an attention sentence representation corresponding to the topic. In the next step, attentive sentence representations each are fed to a fully connected layer followed by the squash activation function (Sabour, Frosst, and Hinton 2017) to get a vector with the length of at most 1 for the topic. Then, the squashed vectors obtained from the previous layer are concatenated together and fed into a fully connected layer followed by squash for each aspect category, similar to the previous layer. In the last step, we treat the length of the output of each aspect category vector as the probability of the given sentence belonging to the aspect category. If the probability of an aspect category surpasses a threshold, the aspect category will be assigned to the review sentence. In actuality, by using the fully connected layer followed by the squash activation function, we transfer the problem into a vector space,
where the probability of a sentence belonging to a category is represented by a vector in the vector space. The longer the length of an output vector, the higher the probability of the sentence belonging to the corresponding category of the output vector.

We evaluate our proposed method by comparing it with several baselines in two freely available benchmark datasets of SemEval workshops. The results confirm the effectiveness of the proposed method, and visualization of the topic-attention weights shows that TAN is able to efficiently attend to different parts of a sentence, given different topics.

Our contributions in this paper are as follows. First, we propose a new neural architecture, Topic-Attention Network (TAN) to capture important words given different topics. Second, by converting the problem into a vector space using the squash activation function (Sabour, Frosst, and Hinton 2017) and treating the length of the output vectors as probabilities, we show the effectiveness of the squash function in the aspect category detection.

**Related Work**

Aspect-based sentiment analysis has gained much attention in recent years following the pioneering work of (Hu and Liu 2004). Based on a hypothesis that aspects are nouns or noun phrases, they used an association rule mining to extract frequent nouns and noun phrases as the candidates for aspects. In the next step, a set of rules are conducted to prune non-aspect candidates. (Qiu et al. 2011) proposed to use a double propagation technique to extract aspect terms and opinion terms in an iterative manner. They conduct a set of rules based on dependency relations to extract aspect terms from opinion terms like ‘good’ or ‘bad’, and vice versa.

Aspect category detection is a subtask of aspect-based sentiment analysis, which instead of extracting aspect terms, there are a set of pre-defined aspect categories like ‘FOOD’ and ‘PRICE’, and the goal is assigning a subset of these categories to a given review sentence. SemEval workshop has addressed aspect category detection task for three consecutive years, which attracted a lot of contestants, especially in SemEval 2016 (Pontiki et al. 2016). (Kiritchenko et al. 2014) proposed multiple features including n-grams, lexicon features, etc. to train a set of one-vs-all SVM classifiers. This model was the top contestant of SemEval 2014 (Pontiki et al. 2014). In (Xenos et al. 2016), authors train a set of one-vs-all SVM classifier with several hand-crafted features. By calculating the Precision, Recall, and F1-Score of the stemmed and un-stemmed N-grams on the train data, they create a set of lexicons for providing features to the classifiers.

In recent years, deep neural network based approaches have been used to address the aspect category detection task, achieving state-of-the-art results. In (Toh and Su 2016), authors proposed using the output of a convolutional neural network trained on the dataset as features for a set one-vs-all of linear classifiers along with several other features such as ngrams and POS tags. This work was the top contestant in SemEval 2016. In (Zhou, Wan, and Xiao 2015) two other loss functions were added to the skip-gram model introduced by (Mikolov et al. 2013) to train a word embedding specifically for aspect category detection. Using a set of multi-layer perceptrons, a set of hybrid features were extracted on the average of word embeddings to train another set of one-vs-all classifiers to extract aspect categories. In (Xue et al. 2017) a set of one-vs-all deep neural networks composed of a CNN layer on top of an LSTM layer was proposed to be trained on both aspect category and aspect term labels simultaneously. A deep neural network approach based on attention mechanism is proposed in (He et al. 2017a). In this paper, the authors use a network similar to an autoencoder in order to perform unsupervised aspect category detection. The proposed network is trained in a way that attends to the aspect-relevant terms.

**Topic-Attention Network**

Figure 1 represents the architecture of Topic-Attention Network (TAN). Our network consists of several components including a sentence encoder layer, a topic-attention layer, and two non-linear transformation layers. In the following subsections, we describe the details of different parts of our model.
Sentence Encoder Layer

We employed a bi-directional recurrent neural network to extract the sequential information for each word. We utilized the Gated Recurrent Unit (GRU) (Cho et al. 2014) for this purpose. In (Chung et al. 2014) the GRU was found to show better performance on small datasets. In order to track the state of sequences, GRU utilizes a gating mechanism without using separate memory cells. The GRU is formulated as follows:

\[ r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}) \]  
\[ z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}) \]  
\[ n_t = \tanh(W_{ix}x_t + b_{ix} + W_{hx}h_{t-1} + b_{hx}) \]  
\[ h_t = (1 - z_t) \odot n_t + z_t \odot h_{t-1} \]  

where \( x_t \) is the input, \( r_t \) is the reset gate, \( z_t \) is the update gate, \( n_t \) is the candidate state, and \( \cdot \odot \) is pair-wise multiplication. The reset gate \( r_t \) decides how of the past state \( h_{t-1} \) is used in the current state \( h_t \) and the update gate \( z_t \) decides how much of the candidate state \( n_t \) is used in the current state \( h_t \).

The input of the encoder layer is a continuous representation of the input sentence using word embeddings.

Topic-Attention Layer

Since not all words contribute equally to the representation of the sentence, we utilize an attention mechanism (Bahdanau, Cho, and Bengio 2014) to emphasize important words of the sentence. Moreover, the importance of a word highly depends on the given topic. For example, consider the review sentence “It is very overpriced and not very tasty” in the restaurant domain. In this example, the word ‘overpriced’ is more important than the other words if the given topic is ‘price’. On the other hand, if the given topic is ‘food’, the word ‘tasty’ should be given higher importance compared to the other words. Therefore, we consider a number of context vectors as topics, where they’re trained during the training process. Figure 2 represents our attention mechanism.

Given the \( i \)th topic and word representations \( \{h_i\}_{i=1}^{N} \) obtained from the encoder layer, the final attentive sentence representation \( v_i \) is computed as follows:

\[ e_{it} = h_i T_i \]  
\[ \alpha_{it} = \frac{\exp(e_{it})}{\sum_{j=1}^{N} \exp(e_{ij})} \]  
\[ v_i = \sum_{j=1}^{N} \alpha_{ij} h_j \]  

where \( e_{it} \) determines the importance of the \( t \)th word with respect to \( i \)th topic by measuring the similarity between the word and the context vector \( T_i \), and \( e_{it} \) is the normalized value of \( e_{it} \), using the softmax function, where \( N \) is the length of the given sentence. Finally, we compute attentive sentence representation \( v_i \) as a weighted sum of word representations obtained from the encoder layer based on their attention weights. Intuitively, the context vector \( T_i \) plays the role of a filter, which learns to emphasize the relevant words to the topic. The number of topics is one of the hyperparameters of the network and context vector of each topic is trained during the training process.

Non-Linear Transformation with Squash

We utilize the squash function proposed in (Sabour, Frosst, and Hinton 2017) which is a non-linear function that ensures the length of almost zero for short vectors and a length of slightly below 1 for long vectors. Given an input \( x \in R^m \) where \( m \) is the length of the input vector, the output of the squash function for \( x \) can be formalized as follows:

\[ squash(x) = \frac{||x||^2}{1 + ||x||^2} \frac{x}{||x||} \]  

In this layer, each output vector obtained from the topic attention layer is fetched into a one-layer MLP in order to extract high-level features and also perform dimensionality reduction simultaneously. The non-linear squash function is then applied to the output vector of the MLP in order to reduce the length of the vector to be no more than 1 while preserving the direction of the vector.

All \( k \) squashed vectors are then concatenated together to provide features for the next similar layer.

Category Detection

The output of the fully connected with squash layer is a squashed vector with the length between 0 and 1, which its L2-Norm can be considered as a probability value. Therefore, the last layer of the TAN consists of \( c \) parallel fully connected layers with squash where \( c \) is the number of categories. To model the probability that a review sentence belongs to each of the categories, we calculate the L2-Norm
of the output vector corresponding to each category. In other words, the longer the length of an output vector, the higher the probability of the given sentence belonging to the corresponding category.

Training Objective

The training parameters of our model \( \theta \) consists of the weights of the model and the context vectors of the topics.

In order to train our model in a multi-label fashion, we utilize the Mean Square Error loss formalized as follows:

\[
J(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]  

(9)

where \( n \) is the number of elements that the loss function is applied to, \( y_i \) is the ground truth value of the \( i \)th element, and \( \hat{y}_i \) is the predicted value for the \( i \)th element.

Following (He et al. 2017b), we also applied a regularization term to the model to keep the topic context vectors orthogonal and encourage the uniqueness of the topics. Given \( T \in \mathbb{R}^{k \times d} \) where \( k \) is the number of topics and \( d \) is the size of the context vectors, we define the regularization term \( U \) as follows:

\[
U(\theta) = ||T_n \cdot T_n^\top - I||
\]  

(10)

where \( T_n \) is the context vectors normalized to have a length of 1 and \( I \) is the identity matrix. This term encourages all the non-diagonal elements of \( T_n \cdot T_n^\top \) to have the value of zero. This means the dot product of the context vectors are encouraged towards being zero. Finally, our overall loss function will be:

\[
L(\theta) = J(\theta) + U(\theta)
\]  

(11)

Experiments

Datasets

For our experiments, we consider two datasets from SemEval workshop, SemEval-2014 task 4 (Pontiki et al. 2014) and SemEval-2016 task 5 (Pontiki et al. 2016). SemEval 2015 dataset was not used because according to (Pontiki et al. 2016), it exists in SemEval 2016 dataset. In both datasets, we used restaurant domain reviews for our experiments. Table 1 shows the statistics of datasets. In SemEval-2016 there are 12 categories which are a combination of aspect and attribute pairs, (e.g. ‘Food#Quality’, ‘Service#General’), while SemEval-2014 has 5 categories which are aspects only, (e.g. ‘Food’, ‘Price’). Review sentences that don’t contain any sentiments and therefore don’t belong to any of the aspect categories are discarded in the training and validation data.

Baseline Methods

In order to show the merit of our model, we compare TAN with multiple baselines in each dataset. The baseline methods are as follows:

• **NLANGP.** The model introduced in (Toh and Su 2016) consists of a CNN model trained on word embeddings and a set of features including POS tags, word clusters, name lists which are fed to a set of binary classifiers for each category. This method is the top-ranked contestant in SemEval 2016 Aspect Category Detection subtask.

• **MTNA.** This method which was introduced in (Xue et al. 2017) utilizes both aspect category and aspect term information to train a set of one-vs-all deep neural network models consisting of an LSTM layer followed by a CNN layer.

• **NRC-Canada.** The model proposed in (Kiritchenko et al. 2014) is the top-ranked contestant in SemEval 2014 aspect category detection subtask. This model addresses the aspect category detection task using a set one-vs-all SVM classifiers (one classifier for each category) using several features including lexicon features, n-grams, word clusters, etc.

• **RepLearn.** In (Zhou, Wan, and Xiao 2015) a semi-supervised in-domain approach for training word embeddings is introduced to capture the semantic relations between words, aspects, and sentiment words and aspects. Afterwards, multiple classifiers are trained to capture hybrid features, which are then used as features for a set of one-vs-all logistic regression classifiers to determine the aspect categories.

• **Vanilla-Attention(VA).** In order to demonstrate the effectiveness of utilizing multiple attentions in the network, we compare our method with a model consisting of an encoder layer, an attention layer, and a fully connected hidden layer with the ReLU activation function. The output of the hidden layer is then fed to another fully connected layer with the Sigmoid activation function to represent the probability of each aspect.

• **Topic-Attention-without-Squash(TAwS).** This baseline was added to demonstrate the effectiveness of the non-linear squash activation function. In this model, the output of the topic attention layers in TAN are fed to a fully connected layer followed by the ReLU activation function and then concatenated. Then, the concatenated vector is fed to another fully connected layer with the Sigmoid activation function to represent the probability of each aspect.

Experiment Settings

In our experiments, we use F1-score, Precision, and Recall as evaluation measures. Stop-words and punctuation removal is performed as a preprocessing step using the NLTK package (Bird, Klein, and Loper 2009). For the input of TAN, TAwS, and VA, we train the word embeddings on the large unlabeled Yelp challenge dataset using the genism

| Dataset       | Train | Test | Total |
|---------------|-------|------|-------|
| SemEval-2014  | 3041  | 800  | 3841  |
| SemEval-2016  | 2000  | 676  | 2676  |

Table 1: The data statistics of the two datasets used for experiments. The numbers denote the number of sentences in each dataset.
We select 10 percent of the training data as the validation set for each aspect category in a uniform manner. All the hyperparameters of the model are tuned on the validation set using grid search. The optimum hidden size of the GRU is found to be 128 for which the combination of forward and backward GRU leads to a 256 dimension vector for each word annotation. The size of the context vector of topics is therefore set to 256, and the optimum number of the topics was found to be 11 for SemEval-2016 dataset and 6 for SemEval-2014 dataset. Since SemEval-2014 has a relatively simpler data compared to SemEval-2016, this difference in their optimum topic number makes sense. We set the size of the squashed vectors of the first fully connected layer with squash to 32 and 64 for the next similar layer for SemEval-2016 dataset, and 16 and 32 for SemEval-2014 dataset.

For training the model, we use a mini-batch size of 128, and training is performed using the Adam optimizer (Kingma and Ba 2014). We use drop out with the probability of 0.6 in order to prevent the overfitting of our model. The model is trained for a maximum of 300 epochs for which early-stopping is performed with the patience set to 20.

We implemented VA, TAwS, and TAN using PyTorch (Paszke et al. 2017) version 0.4.1. All the experiments were done on a GeForce GTX 1080.

### Results and Analysis

The comparison results are shown in table 2. We extracted the results reported for MTNA, NRC-Canada, RepLearn, and NLANG from the original papers (Xue et al. 2017), (Kiritchenko et al. 2014), (Zhou, Wan, and Xiao 2015), and (Toh and Su 2016) respectively.

On the SemEval 2016 dataset, which is more complicated and at the same time smaller compared to the SemEval 2014 dataset, our model comfortably surpasses the other baselines in terms of F1 score. Compared to the MTNA baseline which utilizes aspect term information in the training process, our method outperforms MTNA by 1.96%. Compared to the other baselines, TAN also performs better, surpassing VA, TAwS, and NLANG in term of F1 measure by 3.1% and 3.54%, and 5.3% respectively. Interestingly, VA baseline which is a vanilla version of TAN - TAN without multiple attention as topics - achieves better results compared to all the other baselines in the SemEval-2014 dataset, and outperforms TAwS and NLANG in the SemEval-2016 dataset, which indicates the strength of the attention mechanism for aspect category detection. Also, we see that TAwS in both datasets achieves a worse result compared to VA, which confirms the effectiveness of utilizing squash function in our method. On the SemEval 2014 dataset, TAN also performs better than the other baselines in term of F1 measure. Our method outperforms VA, RepLearn, TAwS, MTNA, and NRC-Canada baseline by 0.44%, 0.51%, 0.81%, 1.7%, and 2.03%, respectively.

### Visualization of Topic-Attention

In this section, we visualize attention weights of sentence words for different topics. Note that, for each word we have a set of attention scores where each score shows the probability of the word belonging to a specific topic. Figure 3 shows the example of attention scores visualization for two sentences from both of the datasets. Each Column denotes a specific topic, so the sum of attention scores in every column is 1. From Figure 3a we can see that there are several topics in the sentence. Topic 1 give an attention score of 1.0 to the word ‘service’ and 0 to the other words, which shows that maybe this topic models the category ‘SERVICE#GENERAL’. Similarly, the words ‘decor’, ‘food’, ‘delicious’, ‘large’, and ‘portions’ get high attention scores each for a specific topic. Intuitively, these words represent different categories, so we expect the model to categorize these words in different topics. Figure 3b shows an example of SemEval-2014 dataset. This dataset is relatively simpler than the SemEval-2016 dataset so here we have 6 topics. According to the attention scores, we can see that obviously topics 1 and 2 models the drinks and foods respectively, which corresponds to the ‘FOOD’ category; and topic 1 models the subjects related to the ‘Anec-dote/Miscellaneous’ category.

### Conclusion

In this paper, we propose a deep neural network based model composed of an encoder layer utilizing GRU recurrent units, a topic-attention layer producing sentence representations based on the existing topics in the data, and two fully connected layers that transfer representations into a vector space, using the squash activation function. Empirical results prove the effectiveness of our model compared to several baselines, including a single attention model and another version of our model without the squash activation function. This indicates the effectiveness of utilizing a topic attention mechanism and non-linear transformation in the vector space via the squash activation function.

| Dataset     | Method | P (%) | R (%) | F1 (%) |
|-------------|--------|-------|-------|--------|
| NRC-Canada  | 91.04  | 86.24 | 88.58 |
| MTNA        | -      | -     | 88.91 |
| TAwS        | 93.24  | 86.41 | 89.70 |
| RepLearn    | -      | -     | 90.10 |
| VA          | 91.54  | 88.85 | 90.17 |
| TAN         | 91.60  | 89.63 | 90.61 |
| NLANGP      | 72.45  | 73.62 | 73.03 |
| TAwS        | 71.11  | 78.97 | 74.84 |
| VA          | 76.06  | 74.52 | 75.28 |
| MTNA        | -      | -     | 76.42 |
| TAN         | 74.78  | 82.34 | 78.38 |

Table 2: The experimental results of our method (TAN) compared with baselines.
(a) An example from SemEval 2016 restaurant dataset.

(b) An example from SemEval 2014 restaurant dataset.

Figure 3: The visualization of the attention values.

References

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[Bird, Klein, and Loper 2009] Bird, S.; Klein, E.; and Loper, E. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”.

[Cho et al. 2014] Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[Chung et al. 2014] Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[He et al. 2017a] He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2017a. An unsupervised neural attention model for aspect extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, 388–397.

[He et al. 2017b] He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2017b. An unsupervised neural attention model for aspect extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, 388–397.

[Hercig et al. 2016] Hercig, T.; Brychcín, T.; Svoboda, L.; and Konkol, M. 2016. Uwb at semeval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), 342–349.

[Hu and Liu 2004] Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177. ACM.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[Kiritchenko et al. 2014] Kiritchenko, S.; Zhu, X.; Cherry, C.; and Mohammad, S. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 437–442.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111–3119.

[Paszke et al. 2017] Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In NIPS-W.

[Pontiki et al. 2014] Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. Semeval-2014 task 5: Aspect based sentiment analysis.

[Pontiki et al. 2016] Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Mohammad, A.-S.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis.
Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), 19–30.

[Qiu et al. 2011] Qiu, G.; Liu, B.; Bu, J.; and Chen, C. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics* 37(1):9–27.

[Rehůřek and Sojka 2010] Rehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. [http://is.muni.cz/publication/884893/en](http://is.muni.cz/publication/884893/en).

[Sabour, Frosst, and Hinton 2017] Sabour, S.; Frosst, N.; and Hinton, G. E. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, 3856–3866.

[Schouten and Frasincar 2016] Schouten, K., and Frasincar, F. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge & Data Engineering* (1):1–1.

[Toh and Su 2016] Toh, Z., and Su, J. 2016. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 282–288.

[Xenos et al. 2016] Xenos, D.; Theodorakakos, P.; Pavlopoulos, J.; Malakasiotis, P.; and Androustopoulos, I. 2016. Auwb-absa at semeval-2016 task 5: Ensembles of classifiers and embeddings for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 312–317.

[Xue et al. 2017] Xue, W.; Zhou, W.; Li, T.; and Wang, Q. 2017. Mtna: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, 151–156.

[Zhou, Wan, and Xiao 2015] Zhou, X.; Wan, X.; and Xiao, J. 2015. Representation learning for aspect category detection in online reviews. In *AAAI*, 417–424.