TDAM: a Topic-Dependent Attention Model for Sentiment Analysis

Gabriele Pergola*, Lin Gui, Yulan He*

University of Warwick, Coventry, UK

Abstract

We propose a topic-dependent attention model for sentiment classification and topic extraction. Our model assumes that a global topic embedding is shared across documents and employs an attention mechanism to derive local topic embedding for words and sentences. These are subsequently incorporated in a modified Gated Recurrent Unit (GRU) for sentiment classification and extraction of topics bearing different sentiment polarities. Those topics emerge from the words' local topic embeddings learned by the internal attention of the GRU cells in the context of a multi-task learning framework. In this paper, we present the hierarchical architecture, the new GRU unit and the experiments conducted on users' reviews which demonstrate classification performance on a par with the state-of-the-art methodologies for sentiment classification and topic coherence outperforming the current approaches for supervised topic extraction. In addition, our model is able to extract coherent aspect-sentiment clusters despite using no aspect-level annotations for training.

Keywords: sentiment analysis, neural attention, topic modeling

1. Introduction

In recent years, attention mechanisms in neural networks have been widely used in various tasks in Natural Language Processing (NLP), including machine

*Corresponding authors.

Email addresses: gabriele.pergola@warwick.ac.uk (Gabriele Pergola), y.he@cantab.net (Yulan He)

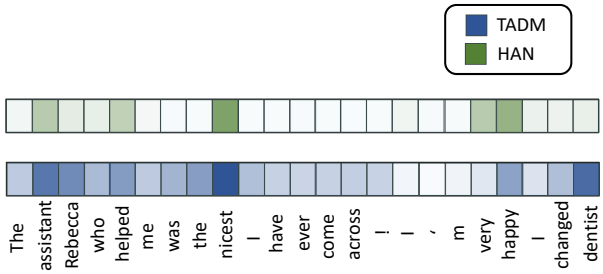

Figure 1: Attention weights from the *Topic-Dependent Attention Model* (TDAM) and *Hierarchical Attention Network* (HAN) (Yang et al., 2016). TDAM highlights and gives more relevance to both sentiment and topical words.

translation (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), image captioning (Xu et al., 2015), text classification (Yang et al., 2016; Chen et al., 2016; Ma et al., 2017) and reading comprehension (Hermann et al., 2015; Wang et al., 2017). Attention mechanisms are commonly used in models for processing sequence data that instead of encoding the full input sequence into a fixed-length vector learn to "attend" to different parts of the input sequence, based on the task at hand. This is equivalent to giving the model the access to its internal memory which consists of the hidden states of the sequence encoder. Typically soft attention is used which allows the model to retrieve a weighted combination of all memory locations.

One advantage of using attention mechanisms is that the learned attention weights can be visualized to enable intuitive understanding of what contributes the most to the model's decision. For example, in sentiment classification, the visualization of word-level attention weights can often give us a clue as to why a given sentence is classified as positive or negative. Words with higher attention weights can be sometimes indicative of the overall sentence-level polarity (for example, see Figure 1). This inspires us the development of a model for the extraction of polarity-bearing topics based on the attention weights learned by a model.

However, simply using the attention weights learned by the traditional attention networks such as the Hierarchical Attention Network (HAN) (Yang et al.,

2016) would not give good results for the extraction of polarity-bearing topics, since in these models the attention weight of each word is calculated as the similarity between the word's hidden state representation with a context vector shared across all the documents. There is no mechanism to separate words into multiple clusters representing polarity-bearing topics.

Therefore, in this paper, we propose a novel Topic-Dependent Attention Model (TDAM)[1] in which a global topic embedding (i.e., a matrix with K topic vectors) is shared across all the documents in a corpus and captures the global semantics in multiple topical dimensions. When processing each word in an input sequence, we can calculate the similarity of the hidden state of the word with each topic vector to get the attention weight along a certain topical dimension. By doing so, we can subsequently derive the local topical embedding for the word by the weighted combination of the global topic embeddings, indicating the varying strength of the association of the word with different topical dimensions. We use Bidirectional Gated Recurrent Unit (BiGRU) to model the input word sequence; we modify the GRU cells to derive a hidden state for the current word which simultaneously takes into account the current input word, the previous hidden state and local topic embedding.

Our proposed formulation of topical attention is somewhat related to the consciousness prior proposed in Bengio (2017) in which the conscious state value corresponds to the content of a thought and can be derived by a form of attention selecting a "small subset of all the information available" from the hidden states of the model. Analogously, we first assume the corpus is characterized by a global topic embedding. Then, we learn how to infer the local topic mixture for each analyzed word/sentence combining hidden states and global topic embedding with attention.

In this paper, we describe TDAM and present its application to sentiment classification in reviews by a hierarchical and multi-task learning architecture. The aim is to evaluate a review's polarity by predicting both the rating and the

[1] `https://github.com/gabrer/topic_dependent_attention_model`

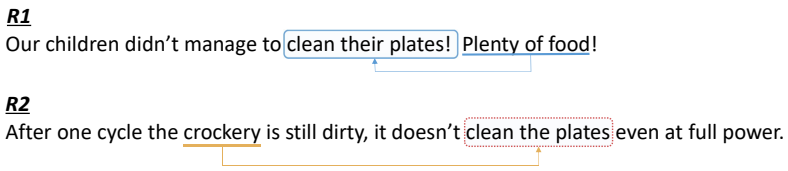

Figure 2: An example of topics bearing polarities.

domain category of the review (e.g. *restaurant*, *service*, *health*, etc.). Often these reviews contain statements that can be fully specified only by the contextual topic. To illustrate, in Figure 2 we show two review extracts, one for a restaurant and another for a dishwasher. Interestingly, the same expression "*not to clean the plates*" can be regarded as positive for food while it bears a negative polarity for kitchen equipment. Thus, it is important to jointly consider both topic and sentiment shared over words for better sentiment analysis.

In particular, we make the following contributions:

- We design a neural architecture and a novel neural unit to analyze users' reviews while jointly taking into account topics and sentiments. The hierarchical architecture makes use of a global topic embedding which encodes the shared topics among words and sentences; while the neural unit employs a new internal attention mechanism which leverages the global topic embeddings to derive a local topic representation for words and sentences.

- We assess the benefit of multi-task learning to induce representations which are based on documents' polarities and domains. Our experiments show that combining the proposed architecture with the modified GRU unit is an effective approach to exploit the polarity and domain supervision for accurate sentiment classification and topic extraction.

- As a side task to evaluate the sentence representations encoded by TDAM, we extract *aspect-sentiment* clusters using no aspect-level annotations during the training; then, we evaluate the coherence of those clusters. Experiments demonstrate that TDAM achieves state-of-the-art performance in extracting clusters whose sentences share coherent polarities and belong

to common domains.

To evaluate the performance of our model, we conduct experiments on both Yelp and Amazon review datasets (see §4.1). We compare the sentiment classification performance with state-of-the-art models (§5). Then, visualization of topical attention weights highlights the advantages of the proposed framework (§5.2). We also evaluate how meaningful are the inferred representations in term of topic coherence (§5.3) and based on their capability to cluster sentences conveying a shared sentiment about a common aspect (§5.4).

2. Related Work

Our work is related to three lines of research.

Hierarchical structure for text classification. Many works have recently proposed to incorporate prior knowledge about the document structure directly into the model architecture to enhance the model's discriminative power in sentiment analysis. A hierarchical model incorporating user and product information was first proposed by Tang et al. (2015) for rating prediction of reviews. Similarly, Chen et al. (2016) combined user and product information in a hierarchical model using attention (Bahdanau et al., 2015); here, attention is employed to generate hidden representations for both products and users. Yang et al. (2016) used a simple and effective two-level hierarchical architecture to generate document representations for text classification; words are combined in sentences and in turn, sentences into documents by two levels of attention. Liu & Lapata (2018) further empowered the structural bias of neural architectures by embedding a differentiable parsing algorithm. This induces dependency tree structures used as additional discourse information; an attention mechanism incorporates these structural biases into the final document representation. Yang et al. (2019) introduced Coattention-LSTM for aspect-based sentiment analysis which designs a co-attention encoder alternating and combining the context and target attention vectors of reviews.

Combining topics with sequence modeling. There has been research incorporating topical information into the sequence modeling of text or use variational neural inference for supervised topic learning. Dieng et al. (2017) developed a language model combining the generative story of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) with the word representations generated by a recurrent neural network (RNN). Stab et al. (2018) proposed incorporating topic information into some gates in Contextual-LSTM, improving generalization accuracy on argument mining. Abdi et al. (2019) proposed to directly incorporate word and sentence level features about contextual polarity, type of sentence and sentiment shifts by encoding prior knowledge about part-of-speech (POS) tagging and sentiment lexicons. Kastrati et al. (2019) enhanced document representations with knowledge from an external ontology and encoded documents by topic modeling approaches. Jin et al. (2018) proposed to perform topic matrix factorization by integrating both LSTM and LDA, where LSTM can improve the quality of the matrix factorization by taking into account the local context of words. Card et al. (2018) proposed a general neural topic modeling framework which allows incorporating metadata information with a flexible variational inference algorithm. The metadata information can be labels driving the topic inference and used for the classification task, analogous to what proposed in a Bayesian framework by Mcauliffe & Blei (2008) with supervised Latent Dirichlet Allocation (S-LDA).

Multi-task learning. Several variants of multi-task learning with neural networks have been recently used for sentiment analysis.

Wu & Huang (2016) proposed a multi-task learning framework for microblog sentiment classification which combines common sentiment knowledge with user-specific preferences. Liu et al. (2016) employed an external memory to allow different tasks to share information. Liu et al. (2017) proposed an adversarial approach to induce orthogonal features for each task. Chen & Cardie (2018) applied a different training scheme to the adversarial approach to minimize the distance between feature distributions across different domains. Zhang et al. (2018) proposed to use an embedded representation of labels to ease the gener-

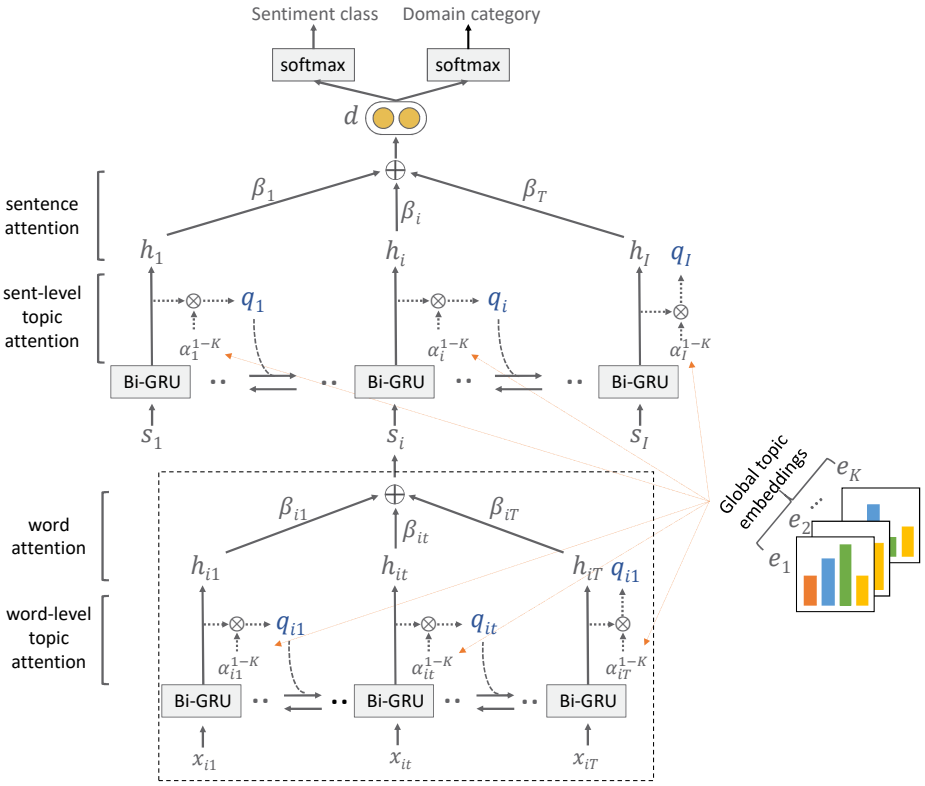

Figure 3: Topic-Dependent Attention Model (TDAM).

ation of cross-domain features. Zheng et al. (2018) proposed to share the same sentence representation for each task which in turn can select the task-specific information from the shared representation using an ad-hoc attention mecha-
140 nism. Wang et al. (2018) applied multi-task learning for microblog sentiment classification by characterizing users across multiple languages.

3. Topic-Dependent Attention Model

We illustrate the architecture of our proposed Topic-Dependent Attention Model (TDAM) in Figure 3, which is a hierarchical and multi-level attention
145 framework trained with multi-task learning.

Concretely, at the word sequence level (the bottom part of Figure 3), we add a word-level topic attention layer which computes the local topic embedding of

each word based on the global topic embedding and the current hidden state. Such word-level local topic embedding indicates how strongly each word is associated with every topic dimension, which is fed into the Bi-GRU cell in the next time step for the derivation of the hidden state representation of the next word. Bi-GRU is used to capture the topical contextual information in both the forward and backward directions. We then have a word attention layer which decides how to combine the hidden state representations of all the constituent words in order to generate the sentence representation. At the sentence-level, a similar two-level attention mechanism is used to derive the document representation, which is fed into two separate softmax layers for predicting the sentiment class and the domain category. Each of the key components of TDAM is detailed below.

3.1. Topic-Dependent Word Encoder

Given a word sequence $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iT})$, where $x_{it} \in \mathbb{R}^d$ is a word embedding vector with d dimensions, we use Bi-GRU to encode the word sequence. The hidden state at each word position, h_{it}, is represented by the concatenation of both forward and backward hidden states, $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$, which captures the contextual information of the whole sentence centred at x_{it}.

We assume there are K global topic embeddings shared across all documents, where each topic has a dense and distributed representation, $e_k \in \mathbb{R}^n$, with $k = \{1, ..., K\}$, which is initialized randomly and will be updated during model learning.

At each word position, we can calculate the word-level topic weight by measuring the distance between the word vector and each global topic vector. We first project h_{it} using a one-layer MLP and then compute the dot products between the projected h_{it} and global topic vectors $e_k, k = \{1, ..., K\}$ to generate

the weight of local topic embedding for the corresponding word position[2]:

$$u_{it} = \tanh(W_w h_{it}) \tag{1}$$

$$\alpha_{it}^k = \text{softmax}(u_{it}^\intercal e_k) \tag{2}$$

where $W_w \in \mathbb{R}^{n \times n}$ and $k \in \{1, ..., K\}$. The local topic embedding is then:

$$q_{it} = \sum_{k=1}^{K} \alpha_{it}^k \otimes e_k \tag{3}$$

with $q_{it} \in \mathbb{R}^n$, $\alpha_{it} \in \mathbb{R}^K$. Here, $\otimes$ denotes multiplication of a vector by a scalar.

We add the local topic embedding into the GRU cell to rewrite the formulae as follows:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + V_r \boldsymbol{q_{t-1}}) \tag{4}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + V_z \boldsymbol{q_{t-1}}) \tag{5}$$

$$\hat{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1} + V_h \boldsymbol{q_{t-1}})) \tag{6}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, all the W, U and Vs are weight matrices which are learned in the training process, $\odot$ denotes the element-wise product. The reset gate r_t controls how much past state information is to be ignored in the current state update. The update gate z_t controls how much information from the previous hidden state will be kept. The hidden state h_t is computed as the interpolation between the previous state h_{t-1} and the current candidate state $\hat{h}_t$.

In the above formulation, the hidden state in the current word position not only depends on the current input and the previous hidden state, but also takes into account the local topic embedding of the previous word. Since some of those words may be more informative than others in constituting the overall sentence

[2]We drop the bias terms in all the equations in our paper for simplicity.

meaning, we aggregate these representations with a final attention mechanism:

$$v_{it} = \tanh(W_v h_{it}) \tag{8}$$

$$\beta_{it} = \text{softmax}(v_{it}^{\intercal} v_w) \tag{9}$$

$$s_i = \sum_{t=1}^{t} \beta_{it} \otimes h_{it} \tag{10}$$

where β_{it} is the attention weight for the hidden state h_{it} and $s_i \in \mathbb{R}^n$ is the sentence representation for the ith sentence.

3.2. Sentence Encoder

Given each sentence representation s_i in document d where $i = \{1, ..., d_L\}$ and d_L denotes the document length, we can form the document representation using the proposed topical GRU in a similar way. For each sentence i, its context vector is $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$, which captures the contextual information of the whole document centred at s_i.

We follow an approach analogous to the topic-dependent word encoder and generate the local topic embedding for ith sentence:

$$u_i = \tanh(W_s h_i) \qquad W_s \in \mathbb{R}^{n \times n} \tag{11}$$

$$\alpha_i^k = \text{softmax}(u_i^{\intercal} e_k) \qquad k \in \{1, ..., K\} \tag{12}$$

$$q_i = \sum_{k=1}^{K} \alpha_i^k \otimes e_k \qquad q_i \in \mathbb{R}^n \tag{13}$$

where q_i is local topic embedding for sentence i. We add the local topic embedding into the GRU cell as in Eq. 4-7.

Analogously to the word encoder, those sentences contribute differently to the overall document meaning; thus, we aggregate these representations with an attention mechanism similar to the final attention mechanism described in Section 3.1.

3.3. Multi-Task Learning

Finally, for each document d, we feed its representation m_d into the task-specific softmax layers, each one defined as follows:

$$p_d = \text{softmax}(W_d m_d) \quad W_d \in \mathbb{R}^{C \times n} \tag{14}$$

where C denotes the total number of classes. The training loss is defined as the total cross-entropy of all documents computed for each task:

$$L_{task} = -\sum_{d=1}^{D}\sum_{c=1}^{C} y_{d,c} \log p_{d,c} \tag{15}$$

where $y_{d,c}$ is the binary indicator (0 or 1) if class label c is the correct classification for document d. We compute the overall loss as a weighted sum over the task-specific losses:

$$L_{total} = \sum_{j=1}^{J} \omega_j L(\hat{y}^{(j)}, y^{(j)}) \tag{16}$$

where J is the number of tasks, ω_j is the weight for each task, $y^{(j)}$ are the ground-truth labels in task j and $\hat{y}^{(j)}$ are the predicted labels in task j.

3.4. Topic Extraction

Once our model is trained, we can feed the test set and collect the local topic embedding q_{it} associated to each word (Eq. 3), collecting a set of n-dimensional vectors for each occurrence of words in text. This mechanism can be interpreted analogously to models generating deep contextualised word representations based on language model, where each word occurrence has a unique representation based on the context in which it appears (Peters et al., 2018; Devlin et al., 2019).

The local representation q_{it} in our model results from the interaction with the global topic embeddings, which encode the word co-occurrence patterns characterizing the corpus. We posit that these vectors can give us an insight about the topic and polarity relations among words. Therefore, we first project these representations into a two-dimensional space by applying the t-SNE (Van der

Maaten & Hinton, 2008); then, the resulting word vectors are clustered by applying the *K-means* algorithm. We create a fixed number of clusters k, whose value is tuned by maximizing the topic coherence for $k \in [50, 100, 200]$. We use the distance of each word to the centroid of a topic cluster to rank words within a cluster. Similarly, we cluster sentences based on the representation resulting from the sentence-level topical attention layer. This encoding synthesises both the main topic and polarity characterizing the sentence.

4. Experimental Setup

Dataset	**Yelp18**	**Amazon**
Sentiment classes	3	3
Domain categories	5	5
Documents	75,000	75,000
Average #s	9.7	6.7
Average #w	15.9	16.7
Vocabulary	$\sim 85 \times 10^3$	$\sim 100 \times 10^3$
Tokens	$\sim 11.7 \times 10^6$	$\sim 8.5 \times 10^3$

Table 1: Dataset statistics with #s number of sentences per document and and #w of words per sentence.

4.1. Datasets

We gathered two balanced datasets of reviews from the publicly available Yelp Dataset Challenge dataset in 2018 and the Amazon Review Dataset[3] (McAuley et al., 2015), preserving the meta-information needed for a multi-task learning scenario. Each review is accompanied with one of the three ratings, *positive*, *negative* or *neutral* and comes from five of the most frequent domains[4].

[3] http://jmcauley.ucsd.edu/data/amazon/

[4] For Yelp: *restaurants*, *shopping*, *home services*, *health & medical* and *automotive*. For Amazon: *Pet supplies*, *electronics*, *health personal care*, *clothes shoes* and *home and kitchen*.

Those ratings are the human labeled review scores regarded as gold standard sentiment labels during the experimentation. For each pair of domain and rating, we randomly sample 3,000 reviews, collecting a total of 75,000 reviews. To make it possible for others to replicate our results, we make both the dataset and our source code publicly available[5]. Table 1 summarizes the statistics of the datasets.

4.2. Baselines

We train our proposed TDAM with multi-task learning to perform sentiment and domain classification simultaneously. We compare the performance of TDAM with the following baselines on both sentiment classification and topic extraction:

- **BiLSTM** (Hochreiter & Schmidhuber, 1997) or **BiGRU** (Cho et al., 2014): Both models consider a whole document as a single text sequence. The average of the hidden states is used as features for classification.
- **Hierarchical Attention Network (HAN)** (Yang et al., 2016): The hierarchical structure of this attention model learns word and sentence representations through two additive attention levels.
- **Supervised-LDA (S-LDA)** (Mcauliffe & Blei, 2008): It builds on top of the latent Dirichlet allocation (LDA) (Blei et al., 2003) adding a response variable associated with each document (e.g. review's rating or category).
- **Scholar** (Card et al., 2018): A neural framework for topic models with metadata incorporation without the need of deriving model-specific inference. When metadata are labels, the model infers topics that are relevant to those labels.

The baselines, such as BiLSTM, BiGRU and HAN, are additionally trained with multi-task learning, similar to the setup of our model.

[5] `https://github.com/gabrer/topic_dependent_attention_model`

4.3. Parameter Settings

For our experiments, we split the dataset into training, development and test set in the proportion of 80/10/10 and average all the results over 5-fold cross-validation. We perform tokenization and sentence splitting with SpaCy[6]. We do not filter any words from the dataset during the training phase; although we use the default preprocessing for models like S-LDA and SCHOLAR. Word embeddings are initialized with 200-dimensional GloVe vectors (Pennington et al., 2014). We tune the models' hyperparameters on the development set via a grid search over combinations of learning rate $\lambda \in [0.01, 0.1]$, dropout $\delta \in [0, 0.6]$ and topic vector's size $\gamma_t \in [50, 200]$. Matrices are randomly initialized to be semi-orthogonal matrix (Saxe et al., 2014); all the remaining parameters are randomly sampled from a uniform distribution in $[-0.1, 0.1]$. We adopt Adam optimizer (Kingma & Ba, 2015) and use batch size of 64, sorting documents by length (i.e. number of sentences) to accelerate training convergence; we also apply batch normalization as additional regulariser (Cooijmans et al., 2017).

Once the model is trained, we extract the local topic embedding for each word occurrence in text as its contextualized word representation. These vectors are then projected to a lower-dimensional space by means of a multi-core implementation of a Tree-Based algorithm for accelerating t-SNE[7] (Van Der Maaten, 2014). Then, we cluster these words with K-means[8].

5. Evaluation and results

We report and discuss the experimental results obtained on three evaluation tasks, sentiment classification topic extraction and sentence cluster extraction.

5.1. Sentiment Classification

We train the models under two different settings: a single and a multi-task learning scenario, where we optimize over the only review polarity or over the

[6] https://spacy.io/

[7] https://github.com/DmitryUlyanov/Multicore-TSNE

[8] http://scikit-learn.org/stable/modules/clustering.html

Methods	**Yelp 18**	**Amazon**
BiLSTM	74.5 ± 0.2	72.1 ± 0.2
BiLSTM - Mtl	74.2 ± 0.2	71.8 ± 0.1
BiGRU	75.5 ± 0.1	72.5 ± 0.3
BiGRU - Mtl	75.4 ± 0.2	72.1 ± 0.3
HAN	83.7 ± 0.2	78.4 ± 0.2
HAN - Mtl	83.6 ± 0.3	78.2 ± 0.3
S-LDA	70.8 ± 0.2	64.6 ± 0.1
SCHOLAR	77.3 ± 0.2	71.4 ± 0.2
TDAM	84.2 ± 0.2	78.9 ± 0.2
TDAM - Mtl	$\mathbf{84.5} \pm 0.3$	$\mathbf{79.1} \pm 0.2$

Table 2: Sentiment classification accuracy and standard deviation over the 5-fold cross validation.

combination of polarity and domain, respectively. For the latter, we denote the results with '-Mtl' in Table 2.

It can be observed from the table that BiLSTM and BiGRU perform similarly. With hierarchical attention mechanism at both the word level and the sentence level, HAN boosts the performance by nearly 10% on Yelp and 6% on Amazon compared to BiLSTM and BiGRU. For the neural topic modeling approaches, SCHOLAR outperforms traditional S-LDA by a large margin. However, SCHOLAR is still inferior to HAN. With our proposed topical attentions incorporated into the hierarchical network structure, TDAM further improves on HAN. Multi-task learning does not seem to bring any benefit to sentiment classification for baseline models, though it further improves the performance of TDAM slightly.

5.2. Effectiveness of Topical Attention

If we remove the topical attention and substitute our modified GRU with standard GRU, then the resulting architecture is similar to HAN (Yang et al., 2016) for a multi-task learning setting. In this section, we visualize the attention

	Yelp18			Amazon		
Topics =	50	100	200	50	100	200
HAN	-7.22	-7.05	-7.08	-13.21	-13.15	-13.14
HAN - Mtl	-7.04	-6.94	-6.93	-12.72	-12.20	-12.29
S-LDA	-6.26	-6.13	-6.15	-9.57	-9.41	-9.28
SCHOLAR	-6.24	-6.08	-6.11	-9.52	-9.46	-9.48
SCHOLAR-R	**-6.19**	-6.11	-6.08	-9.34	**-9.09**	-9.17
TDAM	-6.41	-6.12	-6.09	-9.62	-9.50	-9.46
TDAM - Mtl	-6.22	**-6.05**	**-5.93**	**-9.23**	-9.12	**-9.01**

Table 3: Topic coherence for different number of topics. The higher the better.

weights learned by HAN and TDAM to compare their results. Examples are shown in Figure 1. In TDAM, topical words such as *dentist* or the dentist's name, *Rebecca*, are regarded as relevant by the model. Along with them, it focuses on words bearing a strong sentiment, such as *nicest* or *happy*. These weights are compared with the attention weights learned by the HAN, showing that it primarily focuses sentiment words and overlooks other topical words, such as *dentist*.

5.3. Topic Coherence Evaluation

Among the baselines, S-LDA and SCHOLAR are topic modeling methods and therefore the can directly output topics from text. In addition, we can follow the topic extraction procedure described in Section 3.4 to extract topics from HAN to gain an insight into the learned representations. We thus compare the topic extraction results of TDAM with these three models. Also, as previously shown in (Card et al., 2018), higher regularisation on SCHOLAR produced better topics. Therefore, we also report the results using SCHOLAR with higher regularization, named as SCHOLAR-R.

To evaluate the quality of topics, we use the topic coherence measure[9] proposed in (Röder et al., 2015) which has been shown outperforming all the other

[9]https://github.com/dice-group/Palmetto

existing topic coherence measures. We can observe from Table 3 that HAN gives the worse topic coherence results, showing that simply extracting topics using the attention weights is not feasible. With the incorporation of domain category information through multi-task learning, HAN-Mtl gives slightly better coherence results. Among topic modeling approaches, SCHOLAR-R with higher regularization generates more coherence topics compared to SCHOLAR, which outperforms S-LDA. TDAM gives similar topic coherence results as SCHOLAR-R on some topic numbers. TDAM-Mtl improves over TDAM and generates the best coherence results on 2 out of 3 topic settings for both Yelp18 and Amazon, showing higher coherence scores overall.

5.4. Aspect-Polarity Coherence Evaluation

To assess the effectiveness of our proposed TDAM in extracting polarity-bearing topics, we use the annotated dataset provided in the SemEval 2016 Task 5 for aspect-based sentiment analysis[10]; this provides sentence-level annotations about different aspects (e.g. `FOOD#QUALITY`) and polarities (`pos`, `neut`, `neg`) in restaurant and laptop reviews.

We join the training set of restaurant and laptop reviews with the Yelp18 and Amazon dataset, respectively. With the same approach adopted for topic extraction, we use the test sets to generate sentence clusters and evaluate their *aspect-polarity coherence*, defined as the ratio of sentences sharing a common aspect and sentiment in a cluster. For the two topic modeling approaches, S-LDA and SCHOLAR, we generate sentence clusters based on the generative probabilities of sentences conditional on topics. Note that although the SemEval dataset provides the sentence-level annotations of aspects and polarities, these were NOT used for the training of the models here. We only use the gold standard annotations of aspects and polarities in the test set to evaluate the quality of the extracted polarity-bearing topics.

We generate multiple clusters, i.e. (50,100,150), representing polarity-bearing

[10] http://alt.qcri.org/semeval2016/task5/

aspects and report the results in Table 4, which shows the ratio of sentence clusters with more than *threshold* sentences sharing a common aspect (values in brackets) or a common aspect-polarity. We can observe that the topic modeling approaches struggle in generating coherent aspect-polarity clusters with at least 50% of common aspect-polarities. The two hierarchical models, HAN and TDAM, have significantly more coherent aspect-polarity clusters compared to S-LDA and SCHOLAR, and both benefit from multi-task learning. For all the models, results on SemEval-Restaurant are better than those obtained on SemEval-Laptop. This might be partly attributed to the abundant restaurant reviews on Yelp18 compared to the laptop-related reviews on Amazon. Overall, TDAM-Mtl gives the best results.

We also show some example sentence clusters produced by HAN and TDAM under multi-task learning in Table 5. HAN discriminates rather effectively positive sentences (the majority in the cluster) from negative and neutral ones. However, despite several sentences sharing the same polarity, their topics/aspects are quite heterogeneous. TADM phrases are rather coherent overall, both in terms of topics and expressed sentiment.

These results are encouraging. Our TDAM is able to detect coherent aspects and also polarity-bearing aspects despite using no aspect-level annotations at all. Considering it is very time consuming to provide aspect-level annotations, TDAM could be used to bootstrap the training of aspect-based sentiment detectors.

6. Conclusion

We have presented a new topic-dependent attention model for sentiment classification and topic extraction. The conjunction of topical recurrent unit and multi-task learning framework has been shown to be an effective combination to generate representations for more accurate sentiment classification, meaningful topics and for side task of polarity-bearing aspects detection. In future, we will extend the model to deal with discourse-level sentiments (Feng & Hirst, 2012).

Methods	**Topics**	**SemEval-Restaurant**					**SemEval-Laptop**				
	threshold	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$
HAN	50	(0.52) 0.40	(0.28) 0.24	(0.14) 0.10	(0.04) 0.02	(0.00) 0.00	(0.18) 0.15	(0.15) 0.12	(0.08) 0.03	(0.03) 0.01	(0.00) 0.00
	100	(0.64) 0.47	(0.36) 0.27	(0.14) 0.11	(0.09) 0.07	(0.03) 0.03	(0.28) 0.26	(0.21) 0.20	(0.12) 0.08	(0.05) 0.04	(0.01) 0.01
	150	(0.70) 0.59	(0.39) 0.32	(0.21) 0.16	(0.14) 0.12	(0.12) 0.08	(0.37) 0.34	(0.28) 0.23	(0.14) 0.10	(0.8) 0.07	(0.04) 0.03
HAN-Mtl	50	(0.56) 0.40	(0.36) 0.28	(0.26) 0.18	(0.12) 0.10	(0.06) 0.02	(0.24) 0.19	(0.18) 0.15	(0.12) 0.04	(0.04) 0.02	(0.03) 0.01
	100	(0.64) 0.52	(0.51) 0.40	(0.26) 0.22	(0.17) 0.13	(0.10) 0.06	(0.31) 0.27	(0.26) 0.19	(0.16) 0.09	(0.08) 0.04	(0.04) 0.03
	150	(0.72) 0.63	(0.51) 0.43	(0.30) 0.22	(**0.21**) 0.12	(0.14) 0.08	(0.41) 0.38	(0.35) 0.24	(0.17) 0.12	(0.12) 0.08	(0.05) 0.03
S-LDA	50	(0.18) 0.12	(0.08) 0.03	(0.02) 0.01	(0.00) 0.00	(0.00) 0.00	(0.09) 0.07	(0.08) 0.05	(0.03) 0.01	(0.02) 0.00	(0.00) 0.00
	100	(0.24) 0.21	(0.11) 0.10	(0.03) 0.02	(0.00) 0.00	(0.00) 0.00	(0.15) 0.14	(0.10) 0.06	(0.04) 0.01	(0.01) 0.00	(0.00) 0.00
	150	(0.39) 0.35	(0.19) 0.16	(0.05) 0.04	(0.01) 0.01	(0.01) 0.01	(0.27) 0.24	(0.14) 0.11	(0.08) 0.04	(0.2) 0.02	(0.00) 0.00
SCHOLAR-R	50	(0.31) 0.18	(0.22) 0.10	(0.04) 0.03	(0.01) 0.01	(0.00) 0.00	(0.14) 0.10	(0.08) 0.04	(0.06) 0.02	(0.03) 0.01	(0.00) 0.00
	100	(0.39) 0.24	(0.24) 0.13	(0.08) 0.04	(0.03) 0.01	(0.01) 0.01	(0.21) 0.15	(0.16) 0.08	(0.09) 0.05	(0.03) 0.01	(0.01) 0.01
	150	(0.43) 0.36	(0.28) 0.19	(0.11) 0.10	(0.04) 0.03	(0.01) 0.01	(0.34) 0.26	(0.21) 0.13	(0.12) 0.06	(0.05) 0.02	(0.02) 0.01
TDAM	50	(0.54) 0.42	(0.30) 0.24	(0.12) 0.08	(0.06) 0.04	(0.02) 0.00	(0.19) 0.17	(0.15) 0.14	(0.09) 0.06	(0.03) 0.02	(0.00) 0.00
	100	(0.63) 0.55	(0.40) 0.31	(0.21) 0.16	(0.14) 0.10	(0.10) 0.06	(0.38) 0.29	(0.24) 0.18	(0.12) 0.10	(0.05) 0.04	(0.02) 0.02
	150	(0.73) 0.65	(0.43) 0.36	(0.28) **0.26**	(0.19) 0.15	(**0.16**) 0.13	(0.39) 0.37	(0.28) 0.25	(0.16) 0.13	(0.08) 0.08	(0.05) 0.03
TDAM-Mtl	50	(0.68) 0.51	(**0.52**) 0.38	(0.24) 0.20	(0.14) 0.12	(0.06) 0.04	(0.31) 0.25	(0.26) 0.17	(0.22) 0.13	(0.13) 0.07	(0.04) 0.02
	100	(0.72) 0.58	(0.47) 0.39	(0.31) 0.24	(0.19)0.16	(0.13) 0.12	(0.39) 0.32	(0.38) 0.24	(0.26) 0.15	(0.12) 0.09	(0.02) 0.0
	150	(**0.80**) **0.68**	(0.50) **0.40**	(**0.32**) 0.25	(0.20) **0.16**	(**0.16**) **0.14**	(**0.48**) **0.43**	(**0.42**) **0.31**	(**0.26**) **0.18**	(**0.17**) **0.11**	(**0.09**) **0.05**

Table 4: Ratio of clusters where at least $x\%$ sentences sharing the same aspect (values in brackets) and sharing the same aspect-polarity (i.e. both aspect and polarity are correct).

HAN		TDAM	
Positive polarity - Food#Quality			
1) wait the half hour with a cup of joe , and enjoy more than your average breakfast .	FOOD#QUALITY pos	1) the food was all good but it was way too	FOOD#QUALITY neg
2) space was limited , but the food made up for it .	RESTAURANT#MISCELLANEOUS neg	2) the pizza 's are light and scrumptious .	FOOD#STYLE_OPTIONS pos
3)the prices should have been lower .	FOOD#STYLE_OPTIONS neg	3) the food is great and they make a mean bloody mary .	FOOD#QUALITY pos
4) the crowd is mixed yuppies , young and old .	RESTAURANT#MISCELLANEOUS neut	4) great draft and bottle selection and the pizza rocks .	FOOD#QUALITY pos
5)making the cakes myself since i was about seven - but something about these little devils gets better every time .	FOOD#QUALITY pos	5) the food is simply unforgettable !	FOOD#QUALITY pos
6) mmm ... good !	RESTAURANT#GENERAL pos	6)the food is great, the bartenders go that extra mile.	FOOD#QUALITY pos
7) the service is so efficient you can be in and out of there quickly .	SERVICE#GENERAL pos	7)the food is sinful.	FOOD#QUALITY pos
8) service was decent .	SERVICE#GENERAL neut	8)the sushi here is delicious !	FOOD#QUALITY pos
9) their specialty rolls are impressive	FOOD#QUALITY pos	9) the food was great !	FOOD#QUALITY pos
10) it was nt́ the freshest seafood ever , but the taste and presentation was ok .	FOOD#STYLE_OPTIONS neut	10) good eats .	FOOD#QUALITY pos
Negative polarity - Food#Quality			
1) the pancakes were certainly inventive but $ 8.50 for 3 - 6 " pancakes (one of them was more like 5 ")	FOOD#STYLE_OPTIONS neg	1) i may not be a sushi guru	FOOD#QUALITY neg
2)a beautiful assortment of enormous white gulf prawns , smoked albacore tuna, [..] and a tiny pile of dungeness	FOOD#STYLE_OPTIONS pos	2) rice is too dry , tuna was n't so fresh either .	FOOD#QUALITY neg
3) space was limited , but the food made up for it .	RESTAURANT#MISCELLANEOUS neg	3) the only way this place survives with such average food is because most customers are one-time customer tourists	FOOD#QUALITY neg
4) the portions are big though , so do not order too much .	FOOD#STYLE_OPTIONS neut	4) the portions are big though , so do not order too much .	FOOD#STYLE_OPTIONS neut
5) not the biggest portions but adequate .	FOOD#QUALITY pos	5) the only drawback is that this place is really expensive and the portions are on the small side .	RESTAURANT#PRICES neg
6)the waiter was a bit unfriendly and the feel of the restaurant was crowded .	SERVICE#GENERAL neg	6) but i can tell you that the food here is just okay and that there is not much else to it .	FOOD#QUALITY neg
7) food was fine , with a some little - tastier - than - normal salsa .	FOOD#QUALITY pos	7) and they give good quantity for the price .	FOOD#STYLE_OPTIONS pos
8) i got the shellfish and shrimp appetizer and it was alright .	FOOD#QUALITY neut	8) food was fine , with a some little - tastier - than - normal salsa .	FOOD#QUALITY pos
9) once seated it took about 30 minutes to finally get the meal .	FOOD#GENERAL neg	9)our drinks kept coming but our server came by a couple times .	SERVICE#GENERAL pos
10) the food is here is incredible , though the quality is inconsistent during lunch .	FOOD#QUALITY pos	10) nice food but no spice !	FOOD#QUALITY neg

Table 5: Clusters of positive and negative aspects about FOOD#QUALITY experiences from SemEval16. On the left sentences clustered with HAN, on the right the ones clustered with TDAM. The aspect and polarity label for each sentence are the gold standard annotations.

References

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, *56*, 1245 – 1259.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*.

Bengio, Y. (2017). The consciousness prior. *CoRR*, *abs/1709.08568*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Card, D., Tan, C., & Smith, N. A. (2018). Neural Models for Documents with Metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018* (pp. 2031–2040). Melbourne, Australia.

Chen, H., Sun, M., Tu, C., Lin, Y., & Liu, Z. (2016). Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016* (pp. 1650–1659). Austin, Texas, USA.

Chen, X., & Cardie, C. (2018). Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2018* (pp. 1226–1240). New Orleans, Louisiana.

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014* (pp. 1724–1734). Doha, Qatar.

Cooijmans, T., Ballas, N., Laurent, C., Gülçehre, Ç., & Courville, A. (2017). Recurrent batch normalization. In *Proceedings of the 2017 International Conference for Learning Representations, ICLR 2017*. Touloun, France.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019* (pp. 4171–4186). Minneapolis, USA.

Dieng, A. B., Wang, C., Gao, J., & Paisley, J. W. (2017). TopicRNN: A recurrent neural network with long-range semantic dependency. In *Proceedings of the 2017 International Conference for Learning Representations, ICLR 2017*. Touloun, France.

Feng, V. W., & Hirst, G. (2012). Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012* (pp. 60–68). Jeju Island, Korea.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28, NIPS 2015* (pp. 1693–1701). Montreal, Canada.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735–1780.

Jin, M., Luo, X., Zhu, H., & Zhuo, H. H. (2018). Combining deep learning and topic modeling for review understanding in context-aware recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2018* (pp. 1605–1614). New Orleans, Louisiana.

Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep

learning on document classification using semantically rich representations. *Information Processing & Management*, *56*, 1618 – 1632.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference for Learning Representations, ICLR 2015*. San Diego, USA.

Liu, P., Qiu, X., & Huang, X. (2016). Deep multi-task learning with shared memory for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016* (pp. 118–127). Austin, Texas, USA.

Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017* (pp. 1–10). Vancouver, Canada.

Liu, Y., & Lapata, M. (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, *6*, 63–75.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015* (pp. 1412–1421). Lisbon, Portugal.

Ma, D., Li, S., Zhang, X., & Wang, H. (2017). Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJACI 2017* (pp. 4068–4074). Melbourne, Australia.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2015*. Santiago, Chile.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems 20, NIPS 2008* (pp. 121–128). Vancouver, Canada.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP 2014* (pp. 1532–1543). Doha, Qatar.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2018* (pp. 2227–2237). New Orleans, Louisiana.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015* (pp. 399–408). Shanghai, China.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the 2015 International Conference for Learning Representations, ICLR 2014*. Banff, Canada.

Stab, C., Miller, T., Schiller, B., Rai, P., & Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing EMNLP 2018* (pp. 3664–3674). Brussels, Belgium.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing EMNLP 2015* (pp. 1422–1432). Lisbon, Portugal.

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, *15*, 3221–3245.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems NIPS 2017* (pp. 5998–6008). Long Beach, California, USA.

Wang, W., Feng, S., Gao, W., Wang, D., & Zhang, Y. (2018). Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing EMNLP 2018* (pp. 338–348). Brussels, Belgium.

Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics ACL 2017* (pp. 189–198). Vancouver, Canada.

Wu, F., & Huang, Y. (2016). Personalized microblog sentiment classification via multi-task learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI 2016* (pp. 3059–3065). Phoenix, Arizona.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning ICML 2015* (pp. 2048–2057). Lille, France.

Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. *Information Processing & Management*, *56*, 463 – 478.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computa-*

tional Linguistics: Human Language Technologies NAACL 2016 (pp. 1480–1489). San Diego, California.

Zhang, H., Xiao, L., Chen, W., Wang, Y., & Jin, Y. (2018). Multi-task label embedding for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing EMNLP 2018* (pp. 4545–4553). Brussels, Belgium.

Zheng, R., Chen, J., & Qiu, X. (2018). Same representation, different attentions: Shareable sentence representation learning from multiple tasks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018* (pp. 4616–4622). Stockholm, Sweden.