Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis

Zhengxuan Wu [1], **Desmond C. Ong** [2, 3]
[1] Symbolic Systems Program, Stanford University
[2] Department of Information Systems and Analytics, National University of Singapore
[3] Institute of High Performance Computing, Agency for Science, Technology, and Research, Singapore
wuzhengx@stanford.edu, dco@comp.nus.edu.sg

Abstract

Aspect-based sentiment analysis (ABSA) and Targeted ASBA (TABSA) allow finer-grained inferences about sentiment to be drawn from the same text, depending on context. For example, a given text can have different targets (e.g., neighborhoods) and different aspects (e.g., price or safety), with different sentiment associated with each target-aspect pair. In this paper, we investigate whether adding context to self-attention models improves performance on (T)ABSA. We propose two variants of Context-Guided BERT (CG-BERT) that learn to distribute attention under different contexts. We first adapt a context-aware Transformer to produce a CG-BERT that uses context-guided *softmax*-attention. Next, we propose an improved Quasi-Attention CG-BERT model that learns a compositional attention that supports subtractive attention. We train both models with pretrained BERT on two (T)ABSA datasets: SentiHood and SemEval-2014 (Task 4). Both models achieve new state-of-the-art results with our QACG-BERT model having the best performance. Furthermore, we provide analyses of the impact of context in the our proposed models. Our work provides more evidence for the utility of adding context-dependencies to pretrained self-attention-based language models for context-based natural language tasks.

1 Introduction

People are living more of their lives online, both on social media and on e-commerce platforms, and this trend was exacerbated by the recent need for social distancing during the Covid-19 pandemic. Because people are using online review platforms like Yelp and delivery platforms more frequently, understanding the types of emotional content generated on such platforms could yield business insights or provide personalized recommendations (Kang, Yoo, and Han 2012). To this end, Sentiment Analysis techniques have been applied to understand the emotional content generated on microblogs (Kouloumpis, Wilson, and Moore 2011; Severyn and Moschitti 2015), online reviews (e.g., movie and restaurant reviews) (Socher et al. 2013; Kiritchenko et al. 2014), narratives (Wu et al. 2019; Ong et al. 2019) and other online social media (Li and Wu 2010; Lwin et al. 2020).

However, user-generated reviews contain more complex information than just a single overall sentiment. A review of an upscale neighborhood (for potential renters or homebuyers) may praise the safety but express incredulity at the price. Identifying the different *aspects* (e.g., price, safety) embedded within a given text, and their associated sentiment, has been formalized as Aspect-Based Sentiment Analysis (ASBA) (Pontiki et al. 2016; Poria et al. 2016). Targeted ABSA (TABSA) is a more general version of ASBA, when there are multiple targets in a review, each with their associated aspects. For example, given a review of neighborhoods: "*LOC1* area is more expensive but has a better selection of amenities than in *LOC2*" (where *LOC1* and *LOC2* are specially masked tokens), we note that the sentiment depends on the specific target (*LOC1* or *LOC2*) and their aspect. The sentiment towards the **price** of *LOC1* may be negative—and may be more important to a price-conscious recent graduate—but positive in terms of **convenience**, while the sentiment towards *LOC2*'s aspects are reversed.

Research using neural models for (T)ABSA has mainly focused on using deep neural networks such as RNNs or attention-gated networks to generated context-dependent sentence embeddings (Saeidi et al. 2016; Tang, Qin, and Liu 2016; Wang et al. 2016; Chen et al. 2017; Ma, Peng, and Cambria 2018; Fan et al. 2018; Liu, Cohn, and Baldwin 2018). Recently, with the advent of powerful self-attention models like the Transformer and BERT, Sun, Huang, and Qiu (2019) and Li et al. (2019) both applied pretrained BERT models to (T)ABSA, and showed promising performance improvements. However, these approaches simply used a pretrained BERT model as a blackbox: either via using BERT as a emebedding layer or appending the aspect to the input sentence.

By contrast, we propose to improve the BERT model architecture to be context-aware. A context-aware BERT model should be distribute its attention weights appropriately under different contexts—in (T)ASBA, this means specific targets and aspects. Additionally, by incorporating context into the calculation of attention weights, we aim to enrich the learnt hidden representations of the models. Formally, we propose two methods to integrate context into the BERT architecture: (1) a Context-Guided BERT (CG-BERT) model adapted from a recent context-aware self-attention network (Yang et al. 2019), which we apply to (T)ASBA; and (2) a novel Quasi-Attention Based Context-Guided BERT (QACG-BERT) model that learns

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

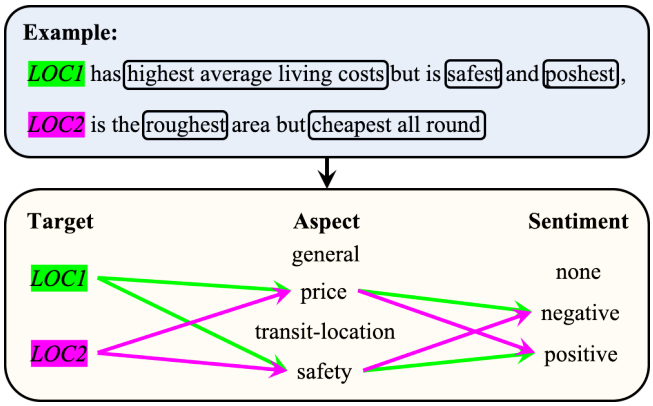


Figure 1: A labelled example from the SentiHood dataset. The sentiment of aspects not mentioned in the text are given the label *none*.

quasi-attention weights—that could be negative—in a compositional manner and which enables subtractive attention, that is lacking in *softmax*-attention. In particular, our contribution is three-fold:

1. We extend a recently-proposed context-aware self-attention network to the (T)ASBA task by formulating a Context-Guided BERT model (CG-BERT).
2. We propose a new Quasi-Attention-based Context-Guided BERT model (QACG-BERT) that achieves new state-of-the-art (SOTA) results on two (T)ASBA tasks.
3. We analyze how context influences the self-attention and decisions of our models.

2 Background and Related Work

2.1 Self-attention Networks

Self-attention networks, exemplified in the Transformer (Vaswani et al. 2017), have become the *de facto* go-to neural models for a variety of NLP tasks including machine translation (Vaswani et al. 2017), language modeling (Liu and Lapata 2018; Dai et al. 2019) and sentiment analysis (Shen et al. 2018; Wu et al. 2019). BERT, a popular and successful variant of the Transformer, has successfully been applied to various NLP tasks (Reddy, Chen, and Manning 2019; Devlin et al. 2019).

2.2 Formulation of *Quasi*-attention

Attention has been successfully applied to many NLP tasks. Various forms of attention are proposed including additive attention (Bahdanau, Cho, and Bengio 2015), dot-product attention (Luong, Pham, and Manning 2015) and scaled dot-product attention used in self-attention (Vaswani et al. 2017). Many of them rely on a *softmax* activation function to calculate attention weights for each position. As a result, the output vector is in the convex hull formed by all other hidden input vectors, preventing the attention gate from learning subtractive relations. This can be solved by allowing attention weights to be negative, using what has been called "quasi" attention, which allows input vectors to add $(+1)$, not contribute (0), and even subtract from (-1) the output vector (Tay et al. 2019).

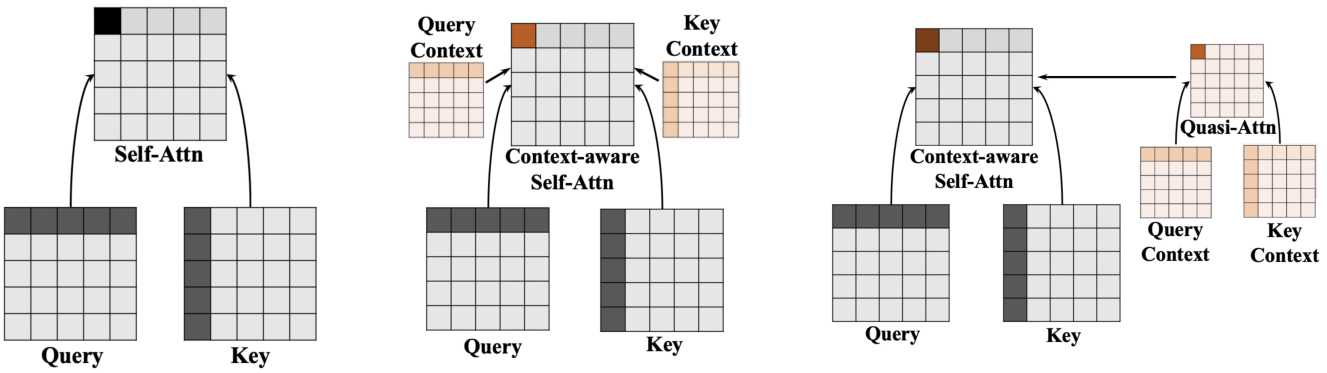
2.3 Aspect-based Sentiment Analysis

Early works in ABSA introduced benchmark datasets and proposed baseline methods including lexicon-based analyses and pre-neural classifiers (Pontiki et al. 2014; Kiritchenko et al. 2014; Pontiki et al. 2015, 2016). Since the debut of recurrent neural networks, various RNNs have been developed to generate aspect-aware sentence embeddings and sentiment labels (Tang et al. 2016; Chen et al. 2017; Li et al. 2018). Likewise, researchers have also adapted CNNs (Xue and Li 2018; Huang and Carley 2018), recursive neural networks (Nguyen and Shirai 2015), aspect-aware end-to-end memory networks (Tang, Qin, and Liu 2016) and cognitively inspired deep neural networks (Lei et al. 2019) to generate aspect-aware sentence embeddings.

Motivated by attention mechanisms in deep learning models, many recent works have integrated attention into neural models such as RNNs (Wang et al. 2016; Chen et al. 2017; Liu and Zhang 2017; He et al. 2018), CNNs (Zhang, Li, and Song 2019), and memory networks (Ma et al. 2017; Majumder et al. 2018; Liu, Cohn, and Baldwin 2018) to learn different attention distributions for aspects and generate attention-based sentence embeddings. Lately, self-attention-based models such as BERT (Devlin et al. 2019) have been applied to ABSA, by using BERT as the embedding layer (Song et al. 2019; Yu and Jiang 2019; Lin, Yang, and Lai 2019), or fine-tuning BERT-based models with an ABSA classification output layer (Xu et al. 2019). These works show BERT brings significant performance gains.

2.4 Targeted Aspect-based Sentiment Analysis

Building on ABSA, Saeidi et al. (2016) proposed a generalized TABSA task (with multiple potential targets) with a new benchmark dataset and LSTM-based baseline models. Various neural models have been proposed for TABSA such as a memory network with delayed context-aware updates (Liu, Cohn, and Baldwin 2018) and interaction-based embedding layers to generate context-aware embeddings (Liang et al. 2019). Researchers have also tried to integrate attention mechanism with LSTMs to predict sentiment for target-aspect pairs (Ma, Peng, and Cambria 2018). With the recent success of BERT-based models, various works have used BERT to generate contextualized embeddings for input sentences and using these as input into deeper layers for sentiment classifications with aspects and targets (Huang and Carley 2019; Hu et al. 2019). More recent works fine-tune BERT for TABSA either with auxiliary sentence constructed for different pairs of targets and aspects or with a classification layer at the top that takes in targets and aspects (Rietzler et al. 2020; Sun, Huang, and Qiu 2019; Li et al. 2019). To the best of our knowledge, no work has been published on modifying BERT architecture for TABSA tasks. Instead of keeping BERT as a blackbox, we enable BERT to be context-aware by modifying its neural architecture to account for context in its attention distributions, and further substantiate our changes by comparing performance results with current best performing models.



(a) BERT (b) CG-BERT (c) QACG-BERT

Figure 2: Illustration of the proposed models. (a) The vanilla self-attention network (e.g., BERT) calculates attention weights using the query and key matrices without considering context. (b) The CG-BERT model modifies query and key matrices using context, and then calculate attention weights as in (a). (c) The QACG-BERT model calculates attention weights by combining vanilla attention weights as in (a) with quasi-attention weights calculated using a separate pair of query and key matrices for context. Colors in the grids illustrate matrix operations.

3 Approach

We start by defining both TABSA and ABSA tasks, then we introduce our context-guided BERT models (Fig. 2). First, we describe the context-guided BERT model CG-BERT that uses *softmax*-attention, originally proposed by Yang et al. (2019), and the modifications we made to tailor it to the (T)ABSA task. Second, we propose a new neural architecture QACG-BERT that uses quasi-attention calculations. Lastly, we describe our methods to formulate our context matrices, and to integrate with pretrained BERT weights.

3.1 TABSA Task

We formulate the Sentihood dataset as a TABSA task. Given a sentence s with a sequence of words $\{w_1, w_2, ...w_n\}$[1], where some words are locations pronouns $\{w_i, ..., w_j\}$ from a fixed set T of k predefined targets $\{t_1, ..., t_k\}$, the goal is to predict sentiment labels for each aspect associated with each unique location mentioned in the sentence. Following the setup in the original Sentihood paper (Saeidi et al. 2016), given a sentence s, a predefined target list T and a predefined aspect list $A = \{$*general*, *price*, *transit-location*, *safety*$\}$, the model predicts a sentiment label $y \in \{$*none*, *negative*, *positive*$\}$ for a given pair of $\{(t, a) : (t \in T, a \in A)\}$. Note that the model predicts a single sentiment label for each unique location-aspect pair in a sentence. We show an example of TABSA in Fig. 1.

3.2 ABSA Task

We use the SemEval-2014 Task 4 dataset (Pontiki et al. 2014) to formulate an ABSA task: Given a sentence, we predict a sentiment label $y \in \{$*none*, *negative*, *neutral*, *positive*, *conflict*$\}$ for each aspect $\{a : (a \in A)\}$ with a predefined aspect list $A = \{$*price*, *anecdotes*, *food*, *ambience*, *service*$\}$.

[1]We append a classifier token (i.e., [CLS]) in the beginning of each input sentence as in the BERT model (Devlin et al. 2019).

3.3 Context-Guided *Softmax*-attention

Our Context-Guided BERT (CG-BERT) model is based on the context-aware Transformer model proposed by Yang et al. (2019), which we adapted to the (T)ABSA task. Multi-headed self-attention (Vaswani et al. 2017) is formulated as:

$$\mathbf{A}^h_{\text{Self-Attn}} = \text{softmax}\left(\frac{\mathbf{Q}^h {\mathbf{K}^h}^T}{\sqrt{d_h}}\right) \quad (1)$$

where $\mathbf{Q}^h \in \mathbb{R}^{n \times d}$ and $\mathbf{K}^h \in \mathbb{R}^{n \times d}$ are query and key matrices indexed by head h, and $\sqrt{d_h}$ is a scaling factor. We integrate context into BERT by modifying $\mathbf{Q}$ and $\mathbf{K}$ matrices of the original BERT model (Devlin et al. 2019):

$$\begin{bmatrix}\hat{\mathbf{Q}}^h \\ \hat{\mathbf{K}}^h\end{bmatrix} = \left(1 - \begin{bmatrix}\lambda^h_Q \\ \lambda^h_K\end{bmatrix}\right) \begin{bmatrix}\mathbf{Q}^h \\ \mathbf{K}^h\end{bmatrix} + \begin{bmatrix}\lambda^h_Q \\ \lambda^h_K\end{bmatrix} \left(\mathbf{C}^h \begin{bmatrix}\mathbf{U}_Q \\ \mathbf{U}_K\end{bmatrix}\right) \quad (2)$$

where $\mathbf{C}^h \in \mathbb{R}^{n \times d_c}$ is the context matrix for each head and defined in Sec. 3.6, $\{\lambda^h_Q, \lambda^h_K\} \in \mathbb{R}^{n \times 1}$ are learnt context weights, and $\{\mathbf{U}_Q, \mathbf{U}_K\} \in \mathbb{R}^{d_c \times d_h}$ are weights of linear layers used to transform input context matrix $\mathbf{C}_h$. The modified $\hat{\mathbf{Q}}$ and $\hat{\mathbf{K}}$ are then used to calculate context-aware attention weights using the dot-product of both matrices. In contrast to the original implementation (Yang et al. 2019), here we allow both λ^h_Q and λ^h_K to differ across heads, which allows variance in how context is integrated in each head.

We use a zero-symmetric gating unit to learn context gating factors $\{\lambda_Q, \lambda_K\}$:

$$\begin{bmatrix}\lambda^h_Q \\ \lambda^h_K\end{bmatrix} = \tanh\left(\begin{bmatrix}\mathbf{Q}^h \\ \mathbf{K}^h\end{bmatrix} \begin{bmatrix}\mathbf{V}^h_Q \\ \mathbf{V}^h_K\end{bmatrix} + \mathbf{C}^h \begin{bmatrix}\mathbf{U}_Q \\ \mathbf{U}_K\end{bmatrix} \begin{bmatrix}\mathbf{V}^C_Q \\ \mathbf{V}^C_K\end{bmatrix}\right) \quad (3)$$

where $\{\mathbf{V}^h_Q, \mathbf{V}^h_K, \mathbf{V}^C_Q, \mathbf{V}^C_K\} \in \mathbb{R}^{d_h \times 1}$ are weights of linear layers to transform corresponding matrices. We chose to use *tanh* as our activation function as this allows the context to contribute to $\hat{\mathbf{Q}}^h, \hat{\mathbf{K}}^h$ both positively and negatively[2]. This

[2]The original implementation (Yang et al. 2019) used *sigmoid*.

enriches the representation space of both matrices, and the resulting attention distribution. We note that *tanh* may increase the magnitude of $\mathbf{Q}$, $\mathbf{K}$, and large magnitude of $\mathbf{Q}$, $\mathbf{K}$ may push gradients to excessively small values, which may prevent model learning, as noted by Vaswani et al. (2017) and Britz et al. (2017). Our results suggest that this formulation does not negatively affect our model performance.

3.4 Context-Guided *Quasi*-Attention

Our second neural network model (QACG-BERT) architecture proposes using a Quasi Attention function for (T)ABSA. The value of self-attention weights $\mathbf{A}^h_{\text{Self-Attn}}$ in a vanilla implementation (using *softmax*), is bounded between $\{0, 1\}$. In other words, it only allows a convex weighted combination of hidden vectors at each position. This allows hidden states to contribute only additively, but not subtractively, to the attended vector. We include a *quasi*-attention calculation to enable learning of both additive as well as subtractive attention (Tay et al. 2019). Formally, we formulate our new attention matrix as a linear combination of a regular *softmax*-attention matrix and a *quasi*-attention matrix:

$$\hat{\mathbf{A}}^h = \mathbf{A}^h_{\text{Self-Attn}} + \lambda^h_A \mathbf{A}^h_{\textit{Quasi}\text{-Attn}} \tag{4}$$

where λ^h_A is a scalar to represent the compositional factor to control the effect of context on attention calculation. $\mathbf{A}^h_{\text{Self-Attn}}$ is defined as in Eqn. 1. To derive the quasi-attention matrix, we first define two terms quasi-context query $\mathbf{C}^h_Q$ and quasi-context key $\mathbf{C}^h_K$:

$$\begin{bmatrix} \mathbf{C}^h_Q \\ \mathbf{C}^h_K \end{bmatrix} = \mathbf{C}^h \begin{bmatrix} \mathbf{Z}_Q \\ \mathbf{Z}_K \end{bmatrix} \tag{5}$$

where $\{\mathbf{Z}_Q, \mathbf{Z}_K\} \in \mathbb{R}^{d_e \times d_h}$ are weights of linear layers to transform the raw context matrix, and $\mathbf{C}^h$ is the same context matrix in Eqn. 2 (Defined in Sec. 3.6). Next, we define the quasi-attention matrix as:

$$\mathbf{A}^h_{\textit{Quasi}\text{-Attn}} = \alpha \cdot \text{sigmoid}\left(\frac{f_\psi(\mathbf{C}^h_Q, \mathbf{C}^h_K)}{\sqrt{d_h}}\right) \tag{6}$$

where α is a scaling factor and $f_\psi(\cdot)$ is a similarity measurement to capture similarities between $\mathbf{C}^h_Q$ and $\mathbf{C}^h_K$. For simplicity, we use dot-product to parameterize f_ψ, and set α to be 1.0. Other f_ψ that have been used include negative L-1 distance (Tay et al. 2019). As a result, our $\mathbf{A}_{\text{Quasi-Attn}}$ is bounded between $\{0, 1\}$. We then derive our *bidirectional* gating factor λ_A as:

$$\begin{bmatrix} \lambda^h_Q \\ \lambda^h_K \end{bmatrix} = \text{sigmoid}\left(\begin{bmatrix} \mathbf{Q}^h \\ \mathbf{K}^h \end{bmatrix} \begin{bmatrix} \mathbf{V}^h_Q \\ \mathbf{V}^h_K \end{bmatrix} + \begin{bmatrix} \mathbf{C}^h_Q \\ \mathbf{C}^h_K \end{bmatrix} \begin{bmatrix} \mathbf{V}^C_Q \\ \mathbf{V}^C_K \end{bmatrix}\right) \tag{7}$$

$$\lambda^h_A = 1 - (\beta \cdot \lambda^h_Q + \gamma \cdot \lambda^h_K) \tag{8}$$

where $\{\beta, \gamma\}$ are scalars that control the composition weightings. For simplicity, we set $\{\beta, \gamma\}$ = 1.0. We formulate the gating factor to be *bidirectional*, meaning it takes on both positive and negative values, and the output is bounded between $\{-1, 1\}$. Our intuition is that the context-based *quasi*-attention may contribute either positively or negatively to the final attention weights. Consider Eqn. 4: as the first term $\mathbf{A}^h_{\text{Self-Attn}}$ is in $\{0, 1\}$, and the second term is made up of a term (λ_A) that is in $\{-1, 1\}$ and another ($\mathbf{A}_{\text{Quasi-Attn}}$) that is in $\{0, 1\}$, we can conclude that the final attention $\hat{\mathbf{A}}$ lies in $\{-1, 2\}$. That is to say, the final attention weights can take values representing compositional operations including subtraction (-1), deletion ($\times 0$), inclusion/addition ($+1$/$+2$) among hidden vectors across positions. We hypothesize that the *quasi*-attention provides a richer method to integrate context into the calculation of attention.

3.5 Classification

We use the final hidden state (the output of the final layer of the BERT model) of the first classifier token (i.e., [CLS]) as the input to the final classification layer for a *C*-class classification. This is similar to previous work (Sun, Huang, and Qiu 2019). For a given input sentence, we denote this vector as $e_{\text{CLS}} \in \mathbb{R}^{1 \times d}$. Then, the probability of each sentiment class y is given by $y = \text{softmax}(e_{\text{CLS}} \mathbf{W}^T_{\text{CLS}})$ where $\mathbf{W}_{\text{CLS}} \in \mathbb{R}^{C \times d}$ are the weights of the classification layer, and $y \in \mathbb{R}^{1 \times C}$. The label with highest probability will be selected as the final prediction.

3.6 Context Matrix

We use a single integer to represent a context associated with a aspect and a target in any (T)ABSA task, and only an aspect in the ABSA task. We transform these integers into embeddings via a trainable embedding layer. For example, given $|t|$ targets and $|a|$ aspects for any (T)ABSA task, the total number of possible embeddings is $|t| \cdot |a|$. We then concatenate the context embedding with the hidden vector for each position $\mathbf{E} \in \mathbb{R}^{n \times d}$, and pass them into a feed-forward linear layer with a residual connection to formulate the context matrix $\mathbf{C}^h = [\mathbf{E}_c, \mathbf{E}]\mathbf{W}^T_c$ where $\mathbf{E}_c \in \mathbb{R}^{n \times d}$ is the context embedding and $\mathbf{W}_c \in \mathbb{R}^{d \times 2d}$ are the learnt weights for this feed-forward layer.

3.7 Integration with Pretrained BERT

Previous works show that fine-tuning pretrained BERT models increases performance significantly in many NLP tasks (Sun, Huang, and Qiu 2019; Rietzler et al. 2020). Since most of our layers are the same as in BERT model, weights from pretrained BERT models are imported for the overlapped layers between our models and the pretrained model. The weights of newly added layers are initialized to be small[3] random numbers drawn from a normal distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = e^{-3}$. As a result, the gating factors in Eqn. 2 and Eqn. 1 start at values close to zero. This initialization enables the task-specific weights to start from the pretrained weights and slowly diverge during training.

Model	Aspect Categorization			Sentiment	
	Strict Acc. (%)	Macro-F1 (%)	AUC (%)	Acc. (%)	AUC (%)
LR (Saeidi et al. 2016)	-	39.3	92.4	87.5	90.5
LSTM-Final (Saeidi et al. 2016)	-	68.9	89.8	82.0	85.4
LSTM-Loc (Saeidi et al. 2016)	-	69.3	89.7	81.9	83.9
SenticLSTM (Ma, Peng, and Cambria 2018)	67.4	78.2	-	89.3	-
Dmu-Entnet (Liu, Cohn, and Baldwin 2018)	73.5	78.5	94.4	91.0	94.8
BERT-single (Sun, Huang, and Qiu 2019)	73.7	81.0	96.4	85.5	84.2
BERT-pair (Sun, Huang, and Qiu 2019)	79.8	87.9	97.5	93.6	97.0
CG-BERT	69.2	84.2	93.0	90.6	93.3
CG-BERT-*pretrain*	80.2	89.1	**97.8**	93.7	97.2
QACG-BERT	70.4	83.3	94.8	91.3	94.8
QACG-BERT-*pretrain*	**81.0**	**90.4**	97.7	**94.0**	**97.5**

Table 1: Model performance on SentiHood dataset. Best performances are bolded. We include the best results reported for previous models. "-" means not reported in the original paper.

Model	Aspect Categorization			Sentiment		
	Precision (%)	Recall (%)	F1 (%)	Binary (%)	3-class (%)	4-class (%)
XRCE (Brun, Popa, and Roux 2014)	83.23	81.37	82.29	-	-	78.1
NRC-Canada (Kiritchenko et al. 2014)	91.04	86.24	88.58	-	-	82.9
BERT-single (Sun, Huang, and Qiu 2019)	92.78	89.07	90.89	93.3	86.9	83.7
BERT-pair (Sun, Huang, and Qiu 2019)	93.57	90.83	92.18	95.6	89.9	85.9
CG-BERT	91.23	87.33	89.24	87.6	81.2	82.4
CG-BERT-*pretrain*	95.01	**91.86**	93.41	**96.2**	**91.2**	84.3
QACG-BERT	92.21	84.44	89.23	88.6	80.9	81.8
QACG-BERT-*pretrain*	**95.88**	91.12	**93.44**	95.8	90.4	**90.2**

Table 2: Model performance on the Semeval-2014 Task 4 dataset. Aspect categorization corresponds to subtask 3 in the original challenge, and sentiment classification corresponds to subtask 4. Best performances are bolded. We include the best results reported for previous models. "-" means not reported in the original paper.

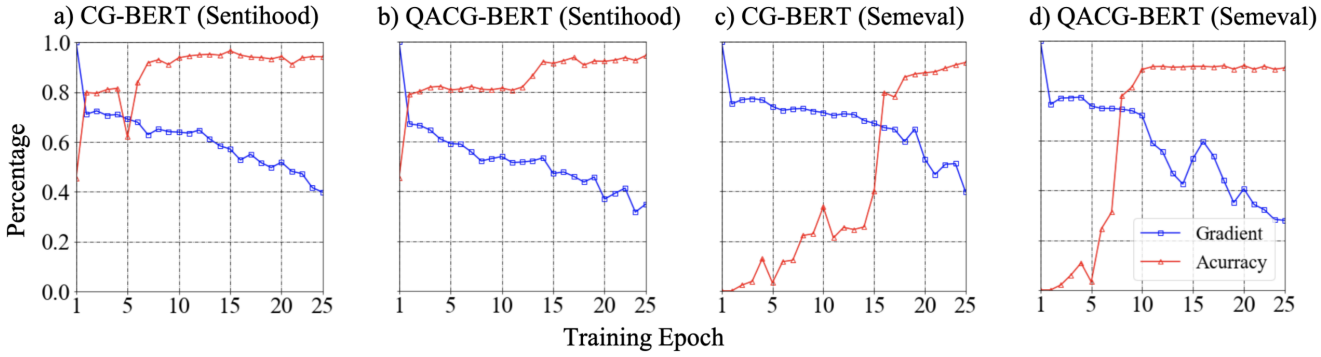


Figure 3: Visualization of the divergence of input relevant scores for the test sets via gradient sensitivity analysis, and the accuracy changes through out the training process. All models are initialized with pretrained BERT weights.

4 Experiments

4.1 Datasets

We evaluate our models with two datasets in English. For the TABSA task, we used the Sentihood dataset [4] which was built by questions and answers from Yahoo! with location names of London, UK. It consists of 5,215 sentences, with 3,862 sentences containing a single target and 1,353 sentences containing multiple targets. For each sentence, we predict sentiment label y for each target-aspect pair (t, a). For the ABSA task, we used the dataset from SemEval 2014,

[3] We also tried initializing the weights in the newly added layers with larger variance, and found similar performance.

[4] https://github.com/uclnlp/jack/tree/master/data/sentihood

Task 4 [5], which contains 3,044 sentences from restaurant reviews. For each sentence, we predict the sentiment label y for each aspect a. Each dataset is partitioned to train, development and test sets as in its original paper.

As in previous works (Pontiki et al. 2014; Saeidi et al. 2016) we define two subtasks for each dataset: (1) aspect categorization and 2) aspect-based sentiment classification. For aspect categorization, the problem is to detect whether a aspect a is mentioned (i.e., *none* means not mentioned) in the input sentence for a target t if it is a TABSA task. For aspect-based sentiment classification, we give the model aspects that present (i.e., ignoring *none*'s) and have the model predict the valence of the sentiment (i.e., potential labels include *negative* and *positive* for Sentihood, and *negative*, *neutral*, *positive*, *conflicting sentiment* for Semeval Task 4).

4.2 Experiment Settings

As in the original BERT-base model (Devlin et al. 2019), our models consists of 12 heads and 12 layers, with hidden layer size 768. The total number of parameters for both of our models increased slightly due to the additional linear layers added comparing to previous BERT-based models for ABSA tasks (Sun, Huang, and Qiu 2019) which consists of about 110M parameters. The CG-BERT and QACG-BERT consists of about 124M parameters. We trained for 60 epochs with a dropout probability of 0.1. The initial learning rate is $2e^{-5}$, with a batch size of 24. We used the pretrained weights from the uncased BERT-base model [6].

We used a single Standard NC6 instance on Microsoft Azure, which is equipped with a single NVIDIA Tesla K80 GPU with 12G Memory. Training both models across two datasets took approximately 11 hours.

4.3 Exp-I: TABSA

For the TABSA task, we compared the performance of our models with previous models in Table 1. Following Ma, Peng, and Cambria (2018) and Sun, Huang, and Qiu (2019), for aspect categorization (is a given aspect present in the sentence? If aspect is not present, the label is by definition *none*), we report strict accuracy (model needs to correctly identify all aspects for a given target in the sentence to be counted as accurate), Macro-F1 (the harmonic mean of the Macro-precision and Macro-recall of all targets.) and AUC. For sentiment classification (given an aspect present in the sentence, is the valence *negative* or *positive*?), we report accuracy and AUC.

Results The results of our non-pretrained models showed slight drops in strict accuracy, and comparable performance in other evaluation metrics when we compare with previous SOTA models such as Dmu-Entnet, BERT-single and BERT-pair. After using pretrained BERT weights, the performance of our models were boosted significantly, and surpassed the state-of-the-art models by a large margin. Across multiple evaluation metrics, our proposed quasi-attention model with pre-trained BERT weights (QACG-BERT-*pretrain*) brings promising performance gain over vanilla context-based BERT models (CG-BERT).

4.4 Exp-II: ABSA

For the ABSA task, we compared our models with multiple best performing models for the SemEval-2014 Task 4 dataset in Table 2. Following Pontiki et al. (2016), for aspect categorization, we report Precision, Recall, and F1. For aspect-based sentiment classification, we report accuracies for three different evaluations: binary classification (i.e., *negative* or *positive*), 3-class classification (*negative*, *neutral* or *positive*) and 4-class classification (*negative*, *neutral*, *positive* or *conflict*).

Results Consistent with Exp-I, our models using pretrained BERT weights improved over previous state-of-the-art models on the SemEval-2014 Task 4 dataset. For aspect catogorization, our CG-BERT-*pretrain* performed better than all previous best performing models while QACG-BERT-*pretrain* perform even better on Precision and F1 scores. For aspect sentiment classification, for binary and 3-class classification, CG-BERT-*pretrain* performs better than QACG-BERT-*pretrain* and both models surpass previous SOTA performance. For 4-class classification, QACG-BERT-*pretrain* performs the best, achieving more than 90% accuracy with the next closest model being 85.9%.

4.5 Feature Importance Analysis

Next, we wanted to visualize how our models learn over time, and importantly whether their gradient weights were diverging, indicating context-aware fine-tuning. As shown in Fig. 3, we plot how the gradient scores (Li et al. 2016; Arras et al. 2017) over test set sentences change as a function of training epochs. Specifically, at each epoch, we measure the cosine similarity of the gradient scores with the pre-trained BERT gradients. We also overlay the accuracy (i.e., strict accuracy for Sentihood, and F1 score for Semeval-2014) as a function of epoch. These learning curves show consistent increments in accuracy and reduction in similarities, which is expected given our performance results, and show that our models shift from pretrained models to being more context-aware.

In Fig. 5, we visualize the gradient scores of our QACG-BERT(-*pretrain*) model over two input sentences for which our model predicts sentiment labels correctly. For the first example, words like **rent** and **cheaper** are most relevant to the *price* aspect while words like **live**, **location**, **london** are more relevant to the *transit-location* aspect. Thus, the model learns how to pick out words that are important under different contexts. We then test on a second example with two targets (i.e., locations) for the same aspect, *price*, where the sentiment label for *LOC1* is *negative* while for *LOC2* is *positive*. For *LOC1*, the model correctly identifies **costs** (which in context refers to *LOC1*) and **posh**[7]. By contrast, the model identifies **cheap** when given *LOC2* as context, and assigns a

[5] http://alt.qcri.org/semeval2014/task4/

[6] https://storage.googleapis.com/bert_models/2020_02_20/uncased_L-12_H-768_A-12.zip

[7] The BERT Tokenizer breaks up 'poshest' into po-sh-est. We

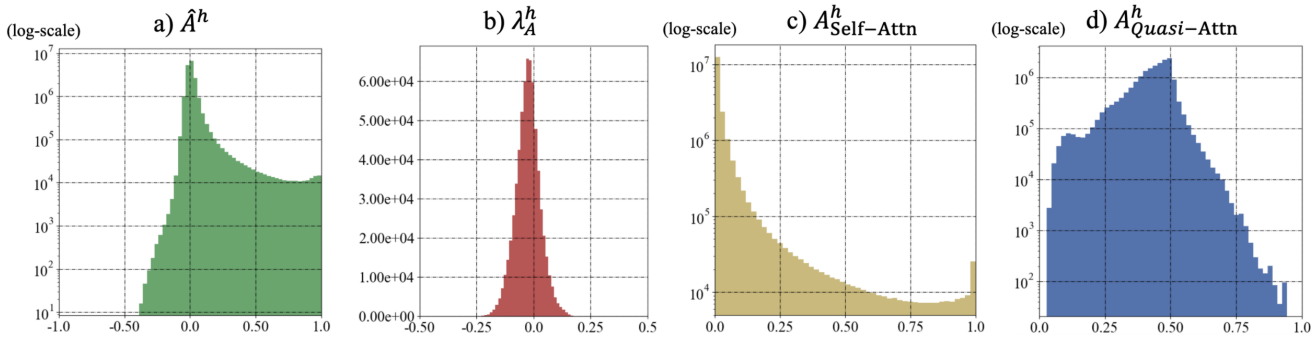


Figure 4: Histogram of QACG-BERT weights on n =200 randomly selected examples from SentiHood test set. For the matrices, each value in the matrix is a data point in the histogram. We note that the vertical axes of (a), (c) and (d) are in log-scale.

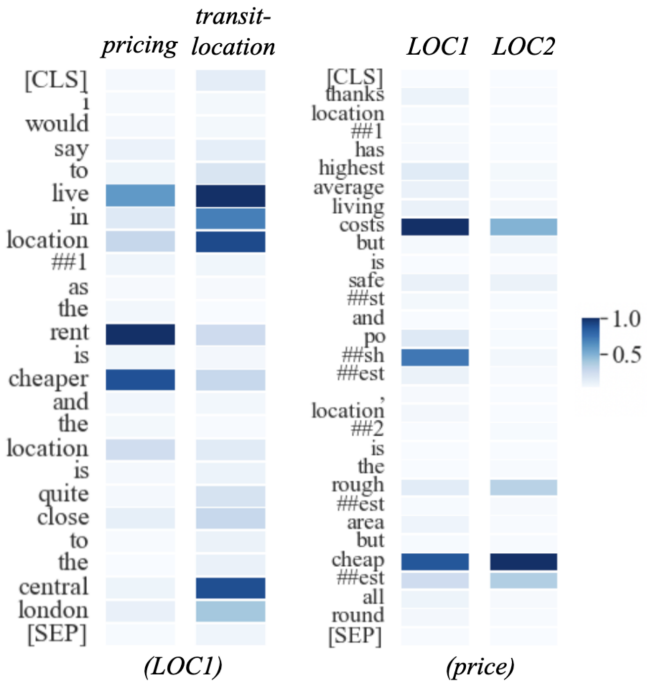


Figure 5: Example of relevance scores via gradient sensitivity analysis for different aspects and locations, from the SentiHood dataset. Left, gradients associated with two aspects (*pricing*, *transit-location*) for target *LOC1*. Right, gradients associated with a *general* aspect, but for *LOC1* and *LOC2*. Values are normalized with respect to the maximum value within a sentence.

greater gradient value, compared to *LOC1*. As a result, the model is able to identify words corresponding to different targets and different aspects.

4.6 Quasi-attention Visualization

We also inspected the quasi-attention parameters learnt by our QACG-BERT model. Specifically, we took the QACG-BERT-*pretrain* model trained on the Sentihood dataset and extracted weights in the following four variables: the final attention weights matrix $\hat{\mathbf{A}}^h$, the *bidirectional* gating factor matrix λ_A^h, the vanilla self-attention matrix $\mathbf{A}^h_{\text{Self-Attn}}$ and the context-guided quasi-attention matrix $\mathbf{A}^h_{Quasi\text{-Attn}}$. Fig. 4 illustrates the histogram of weights drawn from 200 examples from the test set.

We made several key observations. First, the behaviour of λ_A^h follows our intuition; it acts as a bidirectional control gate, with slightly more negative values, and determines whether context contributes to attention positively or negatively. Second, the learnt weights in $\mathbf{A}^h_{Quasi\text{-Attn}}$ are not near zero, in fact with the mass of the distribution between 0.25 and 0.50, so it does contribute to attention. Lastly, the non-zero weights in the final matrix $\hat{\mathbf{A}}^h$ are mainly positive, but some of the weights take on negative values due to the bidirectional gating factor. This is important as it enables the model to both attend to and "de-attend from" different parts of the input.

5 Conclusion

We proposed two BERT-based models for ABSA and TASBA, which outperformed state-of-the-art results on two datasets. Our first CG-BERT model introduce a new way of integrating context into pretrained BERT model and demonstrate promising results. The second QACG-BERT model formulate a new context-guided quasi-attention mechanism to enable compositional alignment score calculations including subtraction (-1), deletion ($\times 0$), inclusion/addition ($+1/+2$). Our results and analyses show strong performance results, especially for our QACG-BERT model, in solving (T)ABSA tasks, and suggest potential success in other context-based tasks in NLP.

References

Arras, L.; Montavon, G.; Müller, K.-R.; and Samek, W. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 159–168.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In

note that the BERT vocabulary does not have 'posh' in it, but after fine-tuning, appears to learn that the word is price-relevant.

Proceedings of the 4th International Conference on Learning Representations (ICLR).

Britz, D.; Goldie, A.; Luong, M.-T.; and Le, Q. 2017. Massive Exploration of Neural Machine Translation Architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1442–1451.

Brun, C.; Popa, D. N.; and Roux, C. 2014. XRCE: Hybrid Classification for Aspect-based Sentiment Analysis. In *SemEval@ COLING*, 838–842. Citeseer.

Chen, P.; Sun, Z.; Bing, L.; and Yang, W. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 452–461.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J. G.; Le, Q.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2978–2988.

Devlin, J.; Chang, M.-W.; Lee, K.; and 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*.

Fan, C.; Gao, Q.; Du, J.; Gui, L.; Xu, R.; and Wong, K.-F. 2018. Convolution-based memory network for aspect-based sentiment analysis. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1161–1164.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2018. Exploiting Document Knowledge for Aspect-level Sentiment Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 579–585.

Hu, M.; Zhao, S.; Guo, H.; Cheng, R.; and Su, Z. 2019. Learning to Detect Opinion Snippet for Aspect-Based Sentiment Analysis. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 970–979.

Huang, B.; and Carley, K. M. 2018. Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1091–1096.

Huang, B.; and Carley, K. M. 2019. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5472–5480.

Kang, H.; Yoo, S. J.; and Han, D. 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications* 39(5): 6000–6010.

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.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*. Citeseer.

Lei, Z.; Yang, Y.; Yang, M.; Zhao, W.; Guo, J.; and Liu, Y. 2019. A human-like semantic cognition network for aspect-level sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6650–6657.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Li, N.; and Wu, D. D. 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems* 48(2): 354–368.

Li, X.; Bing, L.; Lam, W.; and Shi, B. 2018. Transformation Networks for Target-Oriented Sentiment Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 946–956.

Li, X.; Bing, L.; Zhang, W.; and Lam, W. 2019. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. *W-NUT 2019* 34.

Liang, B.; Du, J.; Xu, R.; Li, B.; and Huang, H. 2019. Context-aware Embedding for Targeted Aspect-based Sentiment Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4678–4683.

Lin, P.; Yang, M.; and Lai, J. 2019. Deep Mask Memory Network with Semantic Dependency and Context Moment for Aspect Level Sentiment Classification. In *IJCAI*, 5088–5094.

Liu, F.; Cohn, T.; and Baldwin, T. 2018. Recurrent Entity Networks with Delayed Memory Update for Targeted Aspect-Based Sentiment Analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 278–283.

Liu, J.; and Zhang, Y. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 572–577.

Liu, Y.; and Lapata, M. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics* 6: 63–75.

Luong, M.-T.; Pham, H.; and 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*, 1412–1421.

Lwin, M. O.; Lu, J.; Sheldenkar, A.; Schulz, P. J.; Shin, W.; Gupta, R.; and Yang, Y. 2020. Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. *JMIR Public Health and Surveillance* 6(2): e19447.

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

Ma, Y.; Peng, H.; and Cambria, E. 2018. Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. In *Aaai*, 5876–5883.

Majumder, N.; Poria, S.; Gelbukh, A.; Akhtar, M. S.; Cambria, E.; and Ekbal, A. 2018. IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3402–3411.

Nguyen, T. H.; and Shirai, K. 2015. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2509–2514.

Ong, D.; Wu, Z.; Zhi-Xuan, T.; Reddan, M.; Kahhale, I.; Mattek, A.; and Zaki, J. 2019. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* .

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 486–495.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35.

Poria, S.; Cambria, E.; Hazarika, D.; and Vij, P. 2016. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1601–1612.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7: 249–266.

Rietzler, A.; Stabinger, S.; Opitz, P.; and Engl, S. 2020. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 4933–4941.

Saeidi, M.; Bouchard, G.; Liakata, M.; and Riedel, S. 2016. SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1546–1556.

Severyn, A.; and Moschitti, A. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 959–962.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. DiSAN: Directional Self-Attention Network for RNN/CNN-free language understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Song, Y.; Wang, J.; Jiang, T.; Liu, Z.; and Rao, Y. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314* .

Sun, C.; Huang, L.; and Qiu, X. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 380–385.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2016. Effective LSTMs for Target-Dependent Sentiment Classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3298–3307. Osaka, Japan: The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1311.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 214–224.

Tay, Y.; Luu, A. T.; Zhang, A.; Wang, S.; and Hui, S. C. 2019. Compositional De-Attention Networks. In *Advances in Neural Information Processing Systems*, 6135–6145.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615.

Wu, Z.; Zhang, X.; Zhi-Xuan, T.; Zaki, J.; and Ong, D. C. 2019. Attending to emotional narratives. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 648–654. IEEE.

Xu, H.; Liu, B.; Shu, L.; and Philip, S. Y. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2324–2335.

Xue, W.; and Li, T. 2018. Aspect Based Sentiment Analysis with Gated Convolutional Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2514–2523.

Yang, B.; Li, J.; Wong, D. F.; Chao, L. S.; Wang, X.; and Tu, Z. 2019. Context-aware self-attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 387–394.

Yu, J.; and Jiang, J. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5408–5414. AAAI Press.

Zhang, C.; Li, Q.; and Song, D. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4560–4570.