

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Tourist Sentiment Mining Based on Deep Learning

Weijun Li, Qun Yang and Wencai Du

Abstract

Mining the sentiment of the user on the internet via the context plays a significant role in uncovering the human emotion and in determining the exactness of the underlying emotion in the context. An increasingly enormous number of user-generated content (UGC) in social media and online travel platforms lead to development of data-driven sentiment analysis (SA), and most extant SA in the domain of tourism is conducted using document-based SA (DBSA). However, DBSA cannot be used to examine what specific aspects need to be improved or disclose the unknown dimensions that affect the overall sentiment like aspect-based SA (ABSA). ABSA requires accurate identification of the aspects and sentiment orientation in the UGC. In this book chapter, we illustrate the contribution of data mining based on deep learning in sentiment and emotion detection.

Keywords: Deep learning, Aspect-based Sentiment Analysis, User-generated content, Gated Recurrent Neural Network

1. Introduction

Since the world has been inundated with the increasing amount of tourist data, tourism organizations and business should keep abreast about tourist experience and views about the business, product and service. Gaining insights into these fields can facilitate the development of the robust strategy that can enhance tourist experience and further boost tourist loyalty and recommendations. Traditionally, business rely on the structured quantitative approach, for example, rating tourist satisfaction level based on the Likert Scale. Although this approach is effective to prove or disprove existing hypothesis, the closed ended questions cannot reveal exact tourist experience and feelings of the products or services, which hampers obtaining insights from tourists. Actually, business have already applied sophisticated and advanced approaches, such as text mining and sentiment analysis, to disclose the patterns hidden behind the data and the main themes.

Sentiment analysis (SA) has been used to deal with the unstructured data in the domain of tourism, such as texts, images, and video to investigate decision-making process [1], service quality [2], destination image and reputation [3]. As for the level of sentiment analysis, it has been found that most extant sentiment analysis in the domain of tourism is conducted at document level [4–7]. Document-based sentiment analysis (DBSA) regards the individual whole review or each sentence as an

independent unit and assume there is only one topic in the review or in the sentence. However, this assumption is invalid as people normally express their semantic orientation on different aspects in a review or a sentence [8]. For example, in the sentence “we had impressive breakfast, comfortable bed and friendly and professional staff serving us”, the aspects discussed here are “breakfast”, “bed” and “staff” and the users give positive comments on these aspects (“impressive”, “comfortable” and “friendly and professional”). Since the sentiment obtained through DBSA is at coarse level, aspect-based sentiment analysis (ABSA) has been suggested to capture sentiment tendency of finer granularity.

To obtain the sentiment at the finer level, ABSA has been proposed and developed over the years. ABSA normally involves three tasks, the extraction of opinion target (also known as the “aspect term”), the detection of aspect category and the classification of sentiment polarity. Traditional methods to extract aspects rely on the word frequency or the linguistic patterns. Nevertheless, it cannot identify infrequent aspects and heavily depends on the grammatical accuracy to manipulate the rules [9]. As for the detection of sentiment polarity, supervised machine learning approaches, like Maximum Entropy (ME), Conditional Random Field (CRF) and Support Vector Machine (SVM). Although machine learning-based approaches have achieved desirable accuracy and precision, they require huge dataset and manual training data. In addition, the results cannot be duplicated in other fields [10]. To overcome these shortcomings, ABSA of deep learning (DL) approaches has the advantage of automatically extracting features from data [9]. Extant studies based on DL methods in tourism have investigated and explored tourist experiences in economy hotel [11], the identification of destination image [12], review classification [13]. Although DL methods have been applied in tourism, ABSA in tourism is scant. Therefore, this study reviewed sentiment analysis at aspect level conducted by DL approaches, compared the performance of DL models, and explored the model training process.

With the references of surveys about DL methods [9, 14], this study followed the framework of ABSA proposed by Liu (2011) [8] to achieve the following aims: (1) provide an overview of the studies using DL-based ABSA in tourism for researchers and practitioners; (2) provide practical guidelines including data annotation, pre-processing, as well as model training for potential application of ABSA in similar areas; (3) train the model to classify sentiments with the state-of-art DL methods and optimizers using datasets collected from TripAdvisor. This paper is organized as follows: Section 2 reviews the cutting-edge techniques for ABSA, studies using DL for NLP tasks in tourism, and research gap; Section 3 presents the annotation schema of the given corpus and DL methods used in this study; Section 4 describes the details of annotation results, model training, and the experiment results. Section 5 provides the conclusions and future extensions.

2. Literature review

An extensive literature review of the state-of-art techniques for ABSA and the studies using DL in tourism is provided in this section.

2.1 Input vectors

To convert the NLP problems into the form that computers can deal with, the texts are required to be transformed into a numerical value. In ML-based approaches, One-

hot and Counter Vectorizer are commonly used. One-hot encoding can realize a token-level representation of a sentence. However, the use of One-hot encoding usually results in high dimension issues, which is not computationally efficient [15]. Another issue is the difficulty of extracting meanings as this approach assumes that words in the sentence are independent, and the similarities cannot be measured by distance nor cosine-similarity. As for Counter Vectorizer, although it can convert the whole sentence into one vector, it cannot consider the sequence of the words and the context.

Nevertheless, in DL based approaches, pre-trained word embeddings have been proposed in [16, 17]. Word embedding, or word representation, refers to the learned representation of texts in which the words with identical meanings would have similar representation. It has been proved that the use of word embeddings as the input vectors can make a 6–9% increase in aspect extraction [18] and 2% in the identification of sentiment polarity [19]. Pre-trained word embeddings are favored as random initialization could result in stochastic gradient descent (SGD) in local minima [20]. Based on the network language model, a feedforward architecture, which combined a linear projection layer and a non-linear hidden layer, could learn the word vector representation and a statistical language model [21].

Word2Vec [16] proposed the skip-gram and continuous bag-of-words (CBOW) models. By setting the window size, skip-gram can predict the context based on the given words, while the CBOW can predict the word based on the context. Frequent words also are assigned binary codes in Huffman trees because Also, due to the fact that the word frequency is appropriate to acquire classes in neural net language models, frequent words are assigned binary codes in Huffman trees. This practice in Word2Vec helps reduce the number of output units that are required to be assessed. However, the window-based approaches of Word2Vec do not work on the co-occurrence of the text and do not harness the huge amount of repetition in the texts. Therefore, to capture the global representation of the words in all sentences, GloVe can take advantage of the nonzero elements in a word-word cooccurrence matrix [17]. Although the models discussed above performed well in similarity tasks and named entity recognition, they cannot cope with the polysemous words. In a more recent development, Embeddings from language model (ELMo) [22], Bi-directional Encoder Representations from Transformers (BERT) [23] can identify the context-sensitive features in the corpus. The main difference between these two architectures is that ELMo is feature-based, while BERT is deeply bidirectional. To be specific, the contextual representation of each token is obtained through the concatenation of the left-to-right and right-to-left representations. In contrast, BERT applies masked language models (MLM) to acquire the pre-trained deep bidirectional representations. MLM can randomly mask certain tokens from the input and predict the ID of the input depending only on the context. Additionally, BERT is capable of addressing the issues of long text dependence.

Nonetheless, researchers have combined certain features with word embedding to produce more pertinent results. These features include Part-Of-Speech (POS) and chunk tags, and commonsense knowledge. It has been observed that aspect terms are usually nouns or noun phrases [8]. The original word embeddings of the texts are concatenated with as k-dimensional binary vectors that represent the k POS, or k tags. The concatenated word embeddings are fed into the models (Do et al., Prasad, Maag, and Alsadoon, 2019 [9]). It has been proved that the use of POS tagging as input can improve the performance of aspect extraction, with gains from 1% [18, 20] to 4% [24]. Apart from the POS, concepts that are closely related to the affections are

suggested to be added as word embeddings [25, 26]. POS focused on the grammatical tagging of the words in a corpus, while concepts that are extracted from SenticNet emphasize the multi-word expressions and the dependency relation between clauses. For example, the multi-word expression “win lottery” could be related to the emotions “Arise-joy” and the single-word expression “dog” is associated with the property “Isa-pet” and the emotions “Arise-joy” [26]. After being parsed by SenticNet, the obtained concept-level information (property and the emotions) is embedded into the deep neural sequential models. The performance of the Long Short-Term Memory (LSTM) [27] combined with SenticNet exceeded the baseline LSTM [26].

2.2 DL methods for ABSA

This section reviews the DL methods used for ABSA, including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Attention-based RNN, and Memory Network.

2.2.1 CNN

CNN can learn to capture the fixed-length expressions based on the assumption that keywords usually include the aspect terms with few connections of the positions [28]. Besides, as CNN is a non-linear model, it usually outperforms the linear-model and rarely relies on language rules [29]. A local feature window of 5 words was firstly created for each word in the sentence to extract the aspects. Then, a seven-layer of CNN was tested and generated better results [29]. To capture the multi-word expressions, the model proposed [30] contained two separate convolutional layers with non-linear gates. N-gram features can be obtained by the convolutional layers with multiple filters. Li et al. [13] put position information between the aspect words and the context words into the input layer in CNN and introduced the aspect-aware transformation parts. Fan et al. [31] integrated the attention mechanism with a convolutional memory network. This proposed model can learn multi-word expressions in the sentence and identify long-distance dependency.

Apart from simply extracting the aspects alone, CNN can identify the sentiment polarity at the same time, which can be regarded as multi-label tasking classification or multitasking issues. As for researchers who considered ABSA multi-label tasking classification, a probability distribution threshold was applied to select the aspect category and the aspect vector was concatenated with the word embedding, which was then further performed using CNN. Xu et al. [32] combined the CNN with the non-linear CRF to extract the aspect, which was then concatenated with the word embeddings and fed into another CNN to identify the sentiment polarity. Gu et al. [33] proposed a CNN with two levels that integrated the aspect mapping and sentiment classification. Compared with conventional ML approaches, this approach can lessen the feature engineering work and elapsed time [9]. It should be noticed that the performance of multitasking CNN does not necessarily outperform multitasking methods [19].

2.2.2 RNN and attention-based RNN

RNN has been applied for the ABSA and SBSA in the UGC. RNN models use a fixed-size vector to represent one sequence, which could be a sentence or a document,

to feed each token into a recurrent unit. The main differences between CNN and RNN are: (1) the parameters of different layers in RNN are the same, making a fewer number of parameters required to be learned; (2) since the outputs from RNN relies on the prior steps, RNN can identify the context dependency and suitable for texts of different lengths [34–36].

However, the standard RNN has prominent shortcomings of gradient explosion and vanishing, causing difficulties to train and fine-tune the parameter during the process of prorogation [34]. LSTM and Gated Recurrent Unit (GRU) [37] have been proposed to tackle such issues. Also, Bi-directional RNN (Bi-RNN) models have been proposed in many studies [38, 39]. The principle behind Bi-RNN is the context-aware representation can be acquired by concatenating the backward and the forward vectors. Instead of the forward layer alone, a backward layer was combined to learn from both prior and future, enabling Bi-RNN to predict by using the following words. It has been proved that the Bi-RNN model achieved better results than LSTM in the highly skewed data in the task of aspect category detection [40]. Especially, Bi-directional GRU is capable of extracting aspects and identifying the sentiment in the meanwhile [23, 41] by using Bi-LSTM-CRF and CNN to extract the aspects in the sentence that has more than one sentiment targets.

Another drawback of RNN is that RNN encodes peripheral information, especially when it is fed with information-rich texts, which would further result in semantic mismatching problems. To tackle the issue, the attention mechanism is proposed to capture the weights from each lower level, which are further aggregated as the weighted vector for high-level representation [42]. In doing so, the attention mechanism can emphasize aspects and the sentiment in the sentence. Single attention-based LSTM with aspect embeddings [43], and position attention-based LSTM [44], syntactic-aware vectors [45] were used to capture the important aspects and the context words. The aspect and opinion terms can be extracted in the Coupled Multi-Layer Attention Model based on GRU [46] and the Bi-CNN with attention [47]. These frameworks require fewer engineering features compared with the use of CRF.

2.2.3 Memory network

The development of the deep memory network in ABSA was originated from the multi-hop attention mechanism that applies the exterior memory to compute the influence of context words on the given aspects [36]. A multi-hop attention mechanism was set over an external memory that can recognize the importance level of the context words and can infer the sentiment polarity based on the contexts. The tasks of aspect extraction and sentiment identification can be achieved simultaneously in the memory network in the model proposed by [13]. Li et al. [13] used the signals obtained in aspect extraction as the basis to predict the sentiment polarity, which would further be computed to identify the aspects.

Memory networks can tackle the problems that cannot be addressed by attention mechanism. To be specific, in certain sentences, the sentiment polarity is dependent on the aspects and cannot be inferred from the context alone. For example, “the price is high” and “the screen resolution is high”. Both sentences contain the word “high”. When “high” is related to “price”, it refers to negative sentiment, while it represents positive sentiment when “high” is related to “screen resolution”. Wang et al. [48] proposed a target-sensitive memory network proposed six techniques to design target-sensitive memory networks that can deal with the issues effectively.

2.3 Studies using DL methods in tourism and research gap

To obtain finer-grained sentiment of tourists' experiences in economy hotels in China, [11] used Word2Vec to obtain the word embeddings as the model input, and bidirectional LSTM with CRF model was used to train and predict the data. The whole model includes the text layer, POS layer, connection layer, and output layer, in which CRF was used for data output, reaching an accuracy of 84%. Chang et al. [49] applied GloVe to pre-train the word embedding. To improve the performance, feature vectors, like sentiment scores, temporal intervals, reviewer profiles, were added into CNN models. Their results proved that temporal intervals made a greater contribution than the sentiment score and review profile for the managers to respond to the reviews. Gao et al. [50] explored the model that built CNN on LSTM and proved that the combined model outperformed the single CNN or LSTM model, with an improvement of 3.13% and 1.71% respectively.

To summarize, DL methods have been extensively used to perform ABSA. However, ABSA in the domain of tourism is little in the literature. Therefore, this study aimed at conducting ABSA using a dataset collected from TripAdvisor for predicting sentiments. Based on the literature review, it can be observed that RNN models especially attention-based RNN models achieved better performance than CNN models in terms of accuracy. Therefore, attention-based gated RNN models including LSTM and GRU were used in this study, which is summarized in the following section. Zhou et al. [14] conducted a series of ABSA on Semeval datasets [51, 52] using various DL methods. The experimental results confirmed that RNN with an attention-based mechanism obtained higher accuracies but relatively low precisions and recalls. This is because the Semeval datasets are naturally unbalanced datasets in which the fraction of positive sentiment samples is significantly higher than the fractions of neutral and negative sentiment samples, which indicates the importance of fractions of sentiment samples in the datasets. Inspired by ABSA on Semeval datasets, four datasets with different fractions of sentiment samples were resampled from the dataset of TripAdvisor hotel reviews to investigate the effect of sample imbalance on the model performance. Also, optimizers to minimize loss play a key role in model training. Therefore, three optimizers including the state-of-art optimizer were used in this study to compare their performance.

3. Research design and experiment

3.1 Corpora design

Based on the consideration and the purpose of the study, the corpora in this study will be completely in English and will include reviews collected from casino resorts in Macao. A self-designed tool programmed in Python was implemented to acquire all the URLs, which were first stored and further used as the initial page to crawl all the UGC that belongs to the hotel. The corpus includes 61544 reviews of 66 hotels. The length of the reviews varied greatly, with a maximum of 15 sentences, compared to the minimum of one sentence.

In terms of the size of the corpora that requires annotation, as there is no clear instruction regarding the size of the corpora, this study referred to Liu's work and SemEval's task. In machine learning based studies, it is reasonable to consider that the corpus that has 800–1000 aspects would be sufficient, while for deep-learning based

approach, we think at least 5000 aspects in total would be acceptable. As the original data was annotated first to be further analyzed, 1% of the reviews were randomly sampled from the corpus. Therefore, 600 reviews that contain 5506 sentences were selected for ABSA in this study.

3.2 Annotation

Although previous works annotated the corpora and performed sentiment analysis, they did not reveal the annotation principles [51, 53] and the categories are rather coarse. For example, [53] used pre-defined categories to annotate the aspects of the restaurant. The categories involved “Food, Service, Price, Ambience, Anecdotes, and Miscellaneous”, which did not annotate the aspects of finer levels. In addition, the reliability and validity of the annotation scheme have not been proved.

As the training of the models discussed above requires the annotation of domain-specific corpora, this study referred to [54]. The design of the annotation schema calls for the identification of aspect-sentiment pairs. Specifically, A is the collection of aspects a_j (with $j = 1, \dots, s$). Then, sentiment polarity p_k (with $k = 1, \dots, t$) should be added to each aspect in the form of a tuple (a_j, p_k) .

To ensure the reliability and validity, Cohen’s *kappa*, Krippendorff’s *alpha*, and Inter-Annotator-Agreement (IAA) are introduced in this study, which are calculated by the agreement package in NLTK. Both indicators are used to measure (1) the agreement of the entire aspect-sentiment pair, (2) the agreement of each independent category.

3.3 Attention-based gated RNN

3.3.1 LSTM unit

The LSTM unit proposed by [25] overcomes the gradient vanishing or exploding issues in the standard RNN. The LSTM unit is consisted of forget, input, and output gates, as well as a cell memory state. The LSTM unit maintained a memory cell c_t at time t instead of the recurrent unit computing a weighted sum of the inputs and applying an activation function. Each LSTM unit can be computed as follows:

$$X = [h_{t-1} \ x_t] \quad (1)$$

$$f_t = \sigma(XW_f^T + b_f) \quad (2)$$

$$i_t = \sigma(XW_i^T + b_i) \quad (3)$$

$$o_t = \sigma(XW_o^T + b_o) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(XW_c^T + b_c) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

where $W_f, W_i, W_o, W_c \in \mathbb{R}^{d \times 2d}$ are the weighted matrices, and $b_f, b_i, b_o, b_c \in \mathbb{R}^d$ are the bias vectors to be learned, parameterizing the transformation of three gates; d is the dimension of the word embedding; σ is the sigmoid activation function, and \odot represents element-wise multiplication; x_t and h_t are the word embedding vectors and hidden layer at time t , respectively.

The forget gate decides the extent to which the existing memory is kept (Eq. (2)), while the extent to which the new memory is added to the memory cell is controlled by the input gate (Eq. (3)). The memory cell is updated by partially forgetting the existing memory and adding a new memory content (Eq. (5)). The output gate summarizes the memory content exposure in the unit (Eq. (4)). LSTM unit can decide whether to keep the existing memory with three gates. Intuitively, if the LSTM unit detects an important feature from an input sequence at an early stage, it easily carries this information (the existence of the feature) over a long distance, hence, capturing potential long-distance dependencies.

3.3.2 GRU

A Gated Recurrent Unit (GRU) that adaptively remembers and forgets was proposed by [37]. GRU has reset and update gates that modulate the flow of information inside the unit without having a memory cell compared with the LSTM unit. Each GRU can be computed as follows:

$$X = [h_{t-1} \quad x_t] \quad (7)$$

$$r_t = \sigma(XW_r^T + b_r) \quad (8)$$

$$z_t = \sigma(XW_z^T + b_z) \quad (9)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh([r_t \odot h_{t-1} \quad x_t]W^T + b) \quad (10)$$

The reset gate filters the information from the previous hidden layer as a forget gate does in the LSTM unit (Eq. (8)), which effectively allows the irrelevant information to be dropped, thus, allowing a more compact representation. On the other hand, the update gate decides how much the GRU updates its information (Eq. (9)). This is similar to LSTM. However, the GRU does not have the mechanism to control the degree to which its state is exposed instead of fully exposing the state each time.

3.3.3 Attention mechanism

The standard LSTM and GRU cannot detect the important part for aspect-level sentiment classification. To address this issue, [43] proposed an attention mechanism that allows the model to capture the key part of a sentence when different aspects are concerned. The architecture of a gated RNN model considering the attention mechanism which can produce an attention weight vector α , and a weighted hidden representation r .

$$M = \tanh \left(\begin{bmatrix} W_h H \\ W_v v_a \otimes e_N \end{bmatrix} \right) \quad (11)$$

$$\alpha = \text{softmax}(W_m M) \quad (12)$$

$$r = H\alpha^T \quad (13)$$

where $H \in \mathbb{R}^{d_h \times N}$ is the hidden matrix, d_h is the dimension of the hidden layer, N is the length of the given sentence; $v_a \in \mathbb{R}^{d_a}$ is the aspect embedding, and $e_N \in \mathbb{R}^N$ is a N -dimensional vector with an element of 1; \otimes represents element-wise

multiplication; $W_h \in \mathbb{R}^{d \times d}$, $W_v \in \mathbb{R}^{d_a \times d_a}$, $W_m \in \mathbb{R}^{d+d_a}$, and $\alpha \in \mathbb{R}^N$ are the parameters to be learned.

The feature representation of a sentence with an aspect h^* is given by:

$$h^* = \tanh(W_p r + W_x h_N) \tag{14}$$

where $h^* \in \mathbb{R}^d$, W_p and $W_x \in \mathbb{R}^{d \times d}$ are the parameters to be learned.

To better take advantage of aspect information, aspect embedding is appended into each word embedding to allow its contribution to the attention weight. Therefore, the hidden layer can gather information from the aspect and the interdependence of words and aspects can be modeled when computing the attention weights.

4. Experiments and results

4.1 Annotation results

In the first trial, Cohen's *kappa* and Krippendorff's *alpha* are obtained at 0.80 and 0.78 respectively. Which are highly acceptable in the study since the scores measured the overall attribute and polarity. To identify the category that has the largest variation between two coders, Cohen's *kappa* for each label was calculated separately. Results (**Table 1**) indicated that Polarity had the highest agreement, while attribute showed lower agreement among two annotators. At the end of the first trial, both coders discussed the issues they encountered when they were annotating the corpus and make efforts to improve the preliminary annotation schema. The problems include dealing with the sentence that is difficult to assign the aspects.

Based on the revisions of the annotation schema, the coders conducted the second trial. With the revised annotation schema, the Cohen's *kappa* for the attribute and polarity is obtained at 0.89 and 0.91 respectively. In addition, Cohen's *kappa* and Krippendorff's *alpha* for the aspect-sentiment pair is computed by the end of the second trial, with 0.82 and 0.81 respectively, which indicated that the annotation schema in this study is valid.

4.2 Model training

The experiment was conducted on the dataset of TripAdvisor hotel reviews which contains 5506 sentences, where the numbers of positive, neutral, and negative sentiment samples are 3032, 2986, and 2725, respectively. Given a dataset, maximizing the predictive performance and training efficiency of a model requires finding the optimal network architecture and tuning hyper-parameters. In addition, the samples can significantly affect the performance of the model. To investigate the effect of

	Attribute	Polarity
First trial	0.86	0.88
Second trial	0.89	0.91

Table 1.
Cohen's *kappa* for categories of aspect and polarity.

sentiment sample fractions on the model performance, four sub-datasets with 4000 sentiment samples subjected to different sentiment fractions were resampled from the TripAdvisor hotel dataset as the train sets, one is a balanced dataset and three are unbalanced datasets that the sample fraction of sentiment positive, neutral, and negative dominated, respectively. In addition, it is observed that the average number of the aspects in a sentence is about 1.4, and the average length of the aspects in a sentence is about 8.0, which indicates that one sentence normally contains more than one aspect and the aspect averagely contains eight characters. The number of aspects in train and test sets is more than 850 and 320, respectively, which confirms the diversity of aspects in the dataset of TripAdvisor hotel reviews. For each train set, 20% of reviews were selected as the validation set.

Attention-based gated RNN models including LSTM and GRU were used for ABSA. Attention-based GRU/LSTM without and with aspect embedding were referred to as AT-GRU/AT-LSTM and ATAЕ-GRU/ATAЕ-LSTM, respectively. The details of the configurations and used hyper-parameters are summarized in **Table 2**. In the experiments, all word embeddings with the dimension of 300 were initialized by GloVe [17]. The word embeddings were pre-trained on an unlabeled corpus of which size is about 840 billion. The dimension of hidden layer vectors and aspect embedding are 300 and 100 respectively. The weight matrices are initialized with the uniform distribution $U(-0.1, 0.1)$, and the bias vectors are initialized to zero. The learning rate and mini-batch size are 0.001 and 16 respectively. The best optimizer and number of epochs were obtained from {SGD, Adam, AdaBelief} and {100, 300, 500} respectively via grid search. The optimal parameters based on the best performance on the validation set were kept and the optimal model is used for evaluation in the test set.

The aim of the training is to minimize the cross-entropy error between the target sentiment distribution y and the predicted sentiment distribution \hat{y} . However, overfitting is a common issue during training. In order to avoid the over-fitting, regularization procedures including L2-regularization, early stopping as well as dropout were used in the experiment. L2-regularization adds “squared magnitude” of coefficient as a penalty term to the loss function.

Configuration	Hyper-parameter
Word embedding	GloVe
Dimension of word embedding	300
Dimension of hidden layer	30
Dimension of aspect embedding	100
Initializer of weight matrices	Uniform distribution $U(-0.1, 0.1)$
Initializer of bias vectors	Zero
Optimizer	Search from {SGD, Adam, AdaBelief}
Number of epochs	Search from {100, 300, 500}
Dropout	0.5
Learning rate	0.001
Multi-batch size	16

Table 2.
Details of configurations and used hyper-parameters.

$$loss = -\sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2 \quad (15)$$

where i is the index of review; j is the index of sentiment class, and the classification in this paper is three-way; λ is the L2-regularization term, which modified the learning rule to multiplicatively shrink the parameter set on each step before performing the usual gradient update; θ is the parameter set.

On the other hand, early stopping is a commonly used and effective way to avoid over-fitting. It reliably occurs that the training error decreases steadily over time, but validation set error begins to rise again. Therefore, early stopping terminates when no parameters have improved over the best-recorded validation error for a pre-specified number of iterations. Additionally, dropout is a simple way to prevent the neural network from overfitting, which refers to temporarily removing cells and their connections from a neural network [55]. In an RNN model, dropout can be implemented on input, output, and hidden layers. In this study, only the output layer with a dropout ratio of 0.5 was followed by a linear layer to transform the feature representation to the conditional probability distribution.

Optimizers are algorithms used to update the attributes of the neural network such as parameter set and learning rate to reduce the losses to provide the most accurate results possible. Three optimizers namely SGD [56], Adam [57], and AdaBelief [58] were used in the experiment to search for the best performance. The standard SGD uses a randomly selected batch of samples from the train set to compute derivate of loss, on which the update of the parameter set is dependent. The updates in the case of the standard SGD are much noisy because the derivative is not always toward minima. As result, the standard SGD may have a more time complexity to converge and get stuck at local minima. In order to overcome this issue, SGD with momentum is proposed by Polyak [56] (1964) to denoise derivative using the previous gradient information to the current update of the parameter set. Given a loss function $f(\theta)$ to be optimized, the SGD with momentum is given by:

$$v_{t+1} = \beta v_t - \alpha g_t \quad (16)$$

$$\theta_{t+1} = \theta_t + v_{t+1} \quad (17)$$

where $\alpha > 0$ is the learning rate; $\beta \in [0, 1]$ is the momentum coefficient, which decides the degree to which the previous gradient contributing to the updates of the parameter set, and $g_t = \nabla f(\theta_t)$ is the gradient at θ_t .

Both Adam and AdaBelief are adaptive learning rates optimizer. Adam records the first moment of gradient m_t which is similar to SGD with momentum and second moment of gradient v_t in the meanwhile. m_t and v_t are updated using the exponential moving average (EMA) of g_t and g_t^2 , respectively:

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1) g_t \quad (18)$$

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2) g_t^2 \quad (19)$$

where β_1 and β_2 are exponential decay rates.

The second moment of gradient s_t in AdaBelief is updated using the EMA of $(g_t - m_t)^2$, which is easily modified from Adam without extra parameters:

$$s_{t+1} = \beta_2 s_t + (1 - \beta_2) (g_t - m_t)^2 \quad (20)$$

The update rules for parameter set using Adam and AdaBelief are given by Eqs. (23) and (24), respectively:

$$\theta_{t+1} = \theta_t - \frac{\alpha m_t}{\sqrt{v_t} + \varepsilon} \quad (21)$$

$$\theta_{t+1} = \theta_t - \frac{\alpha m_t}{\sqrt{s_t} + \varepsilon} \quad (22)$$

where ε is a small number, typically set as 10^{-8} .

Specifically, the update direction in Adam is $m_t/\sqrt{v_t}$, while the update direction in AdaBelief is $m_t/\sqrt{s_t}$. Intuitively, $1/\sqrt{s_t}$ is the “belief” in the observation, viewing m_t as the prediction of g_t , AdaBelief takes a large step when observation g_t is close to prediction m_t , and a small step when the observation greatly deviates from the prediction.

It is noted that the best models in the validation set were obtained by returning to the parameter set at the point in time with the lowest validation set error.

4.3 Results and analysis

As for the confusion matrix for a multi-class classification task, accuracy is the most basic evaluation measure of classification. The evaluation measure accuracy represents the proportion of the correct predictions of the trained model, and it can be calculated as:

$$Accuracy = \frac{\sum_1^C TP_i}{N} \quad (23)$$

where C is the number of classes (C equals to 3 in this study); N is the sample number of the test set; TP_i is the number of true predictions for the samples of the i^{th} class, which is diagonally positioned in the confusion matrix. In addition to accuracy, classification effectiveness is usually evaluated in terms of macro precision and recall, which are aimed at a class with only local significance. As **Figure 1** illustrates, the class that is being measured is referred to as the positive class and the rest classes are uniformly referred to as the negative classes. The macro precision is the proportion of correct predictions among all predictions with the positive class, while macro recall is the proportion of correct predictions among all positive instances. The macro F1-score is the harmonic mean of macro precision and recall. The macro-average measures take evaluations of each class into consideration, which can be computed as:

$$MacroPrecision = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i} \quad (24)$$

$$MacroRecall = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i} \quad (25)$$

$$Macro - F1 = \frac{2 \times MacroPrecision \times MacroRecall}{MacroPrecision + MacroRecall} \quad (26)$$

where FP_i and FN_i are the number of false predictions for the positive and negative samples of the i^{th} class, respectively.

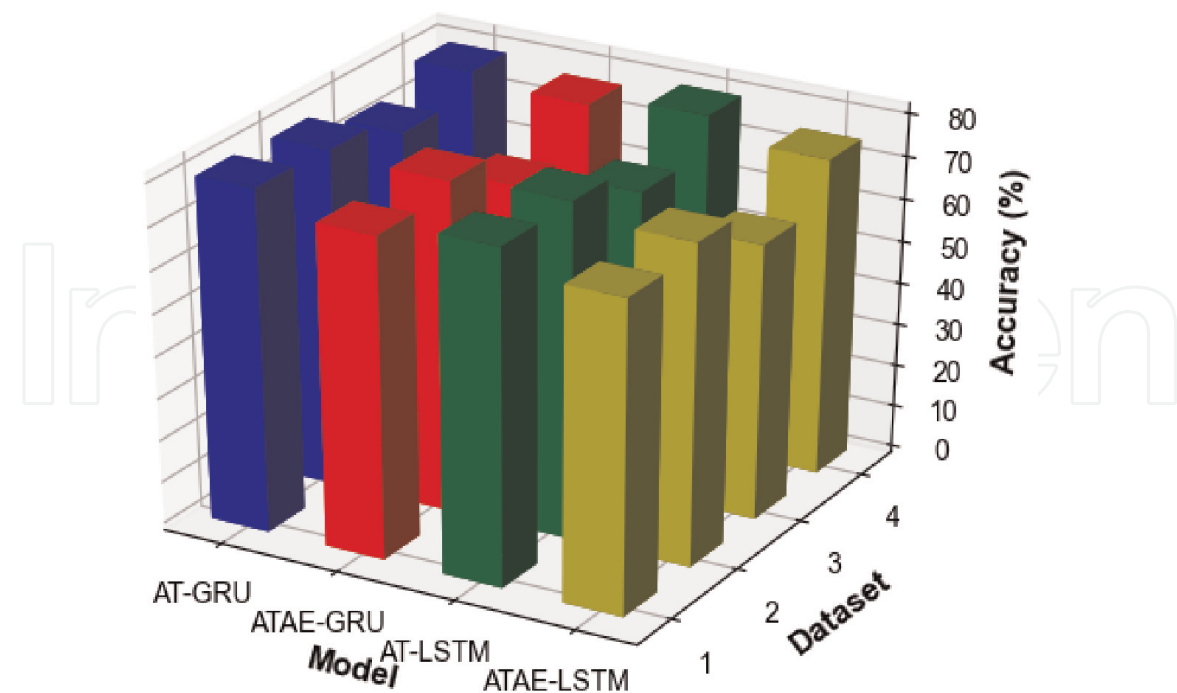


Figure 1.
Summary of model performance.

This study computed accuracy (A), macro precision (P), macro recall (R), and macro F1-score (F) of AT-GRU, ATAE-GRU, AT-LSTM, and ATAE-LSTM trained with various optimizers and epochs. The results show: (1) Attention-based models (AT-GRU and AT-LSTM) performed better than attention-based models with aspect embedding (ATAE-GRU and ATAE-LSTM). Taken Dataset 1 for example, the best accuracy in the test set using AT-GRU was 80.7%, while the best accuracy using ATAE-GRU was 75.3%; (2) Attention-based GRU performed better than attention-based LSTM. Taken AT-GRU and AT-LSTM for example, the accuracy and macro F1-score of AT-GRU for all datasets were higher than those of AT-LSTM; (3) The balanced dataset (Dataset 1) achieved the best predictive performance for all models. For the unbalanced datasets, the accuracy was exactly close to that of the balanced dataset. However, the macro precision, recall, and F1-score were significantly lower than those of the balanced dataset, which confirmed that the balanced dataset had the best generalization and stability in this study; (4) For Dataset 3 in which the neutral sentiment samples dominated, all of the models exhibited the worst predictive performance compared with other datasets. The candidate model for each dataset is illustrated in **Figure 1**. It is noted that the candidate model was selected according to accuracy. However, the model with a higher macro F1-score was selected as the candidate model instead when the accuracies of models were similar. Among 16 models, AT-GRU trained with the optimizer of AdaBelief and epoch of 300 in Dataset 1 achieved the highest accuracy of 80.7% and macro F1-score of 75.0% in the meanwhile. **Figure 2** illustrates the normalized confusion matrix of the best predictive model of which diagonal represented for the precisions. The precisions of positive and negative sentiment classification were about 20% higher than that of neutral sentiment classification, which confirmed that the need to boost the precision of neutral sentiment classification in order to globally improve the accuracy of the model in future work.

Early stopping was used in this research to avoid overfitting and save training time. **Figure 3** illustrates the learning history of AT-GRU using early stopping in four

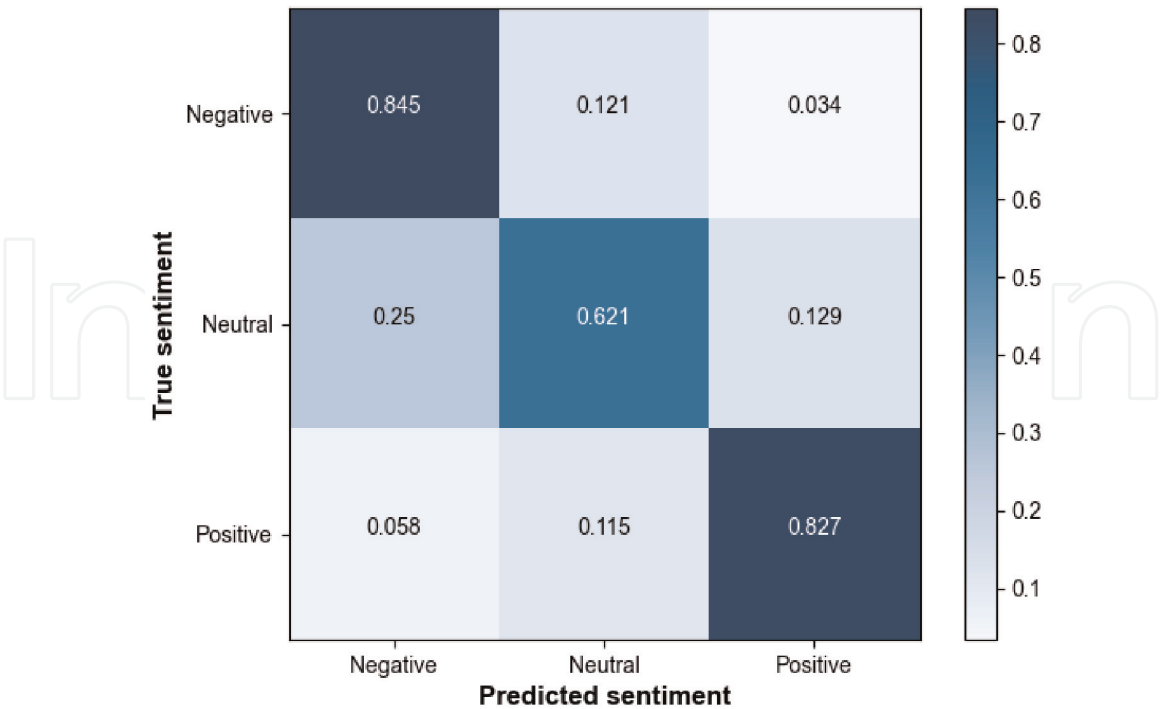


Figure 2. Normalized confusion matrix of model with best predictive performance.

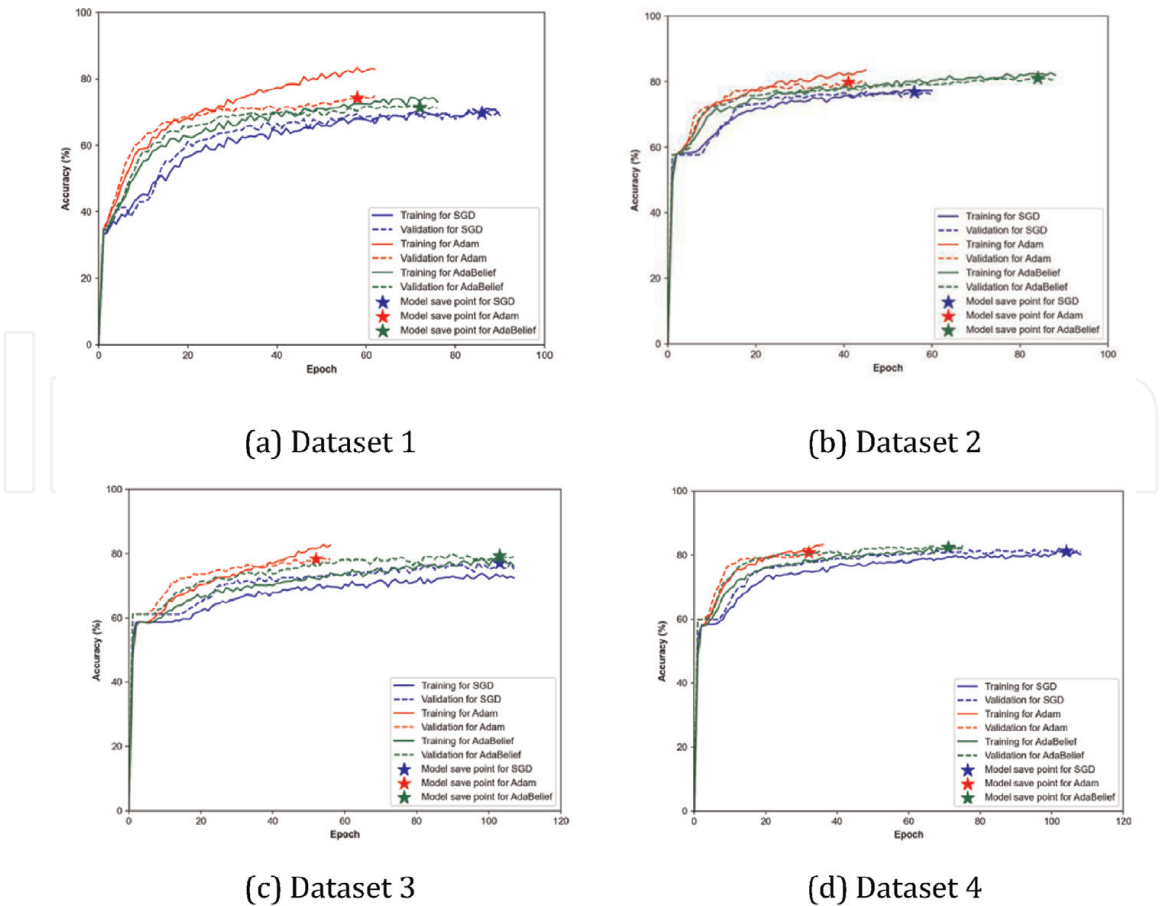


Figure 3. Learning history of AT-GRU using early stopping.

datasets, where the training stopped when the validation loss kept increasing for 5 epochs (i.e., “patience” equals to 5 in this study). For all datasets, the validation accuracy was exactly close to the training validation during the training procedure, which confirmed that early stopping was able to effectively avoid overfitting. Experimental results of A/P/R/F obtained based on training AT-GRU and AT-LSTM using early stopping. The accuracies obtained by AT-GRU and AT-LSTM were similar. For the balanced dataset, the accuracy and macro F1-score obtained by early stopping were significantly lower than that obtained by the corresponding model without early stopping. This is because the loss function probably found the local minima if the training stopped when the loss started to rise for 5 epochs. All of the optimizers used in this study were aimed at avoiding the loss function sticking at the local minima to find the global loss minima, therefore, using more epochs in the training was effective to obtain the best predictive performance model. On the other hand, for the unbalanced datasets, the accuracy and macro F1-score obtained by early stopping were similar to that obtained by the corresponding model without early stopping, which indicated that early stopping was effective to avoid overfitting as the loss converged fast in the unbalanced dataset. Although early stopping is a straightforward way of avoiding overfitting and improving training efficiency, the trade-off is that the model for test set possibly returns at the time point when reaching the local minima of loss function especially for the balanced dataset, and a new hyper-parameter of “patience” which is sensitive to the results is introduced.

Three optimizers were used in this study to find the best model. **Figure 4** illustrates the learning history of AT-GRU in four datasets. The gap between training and

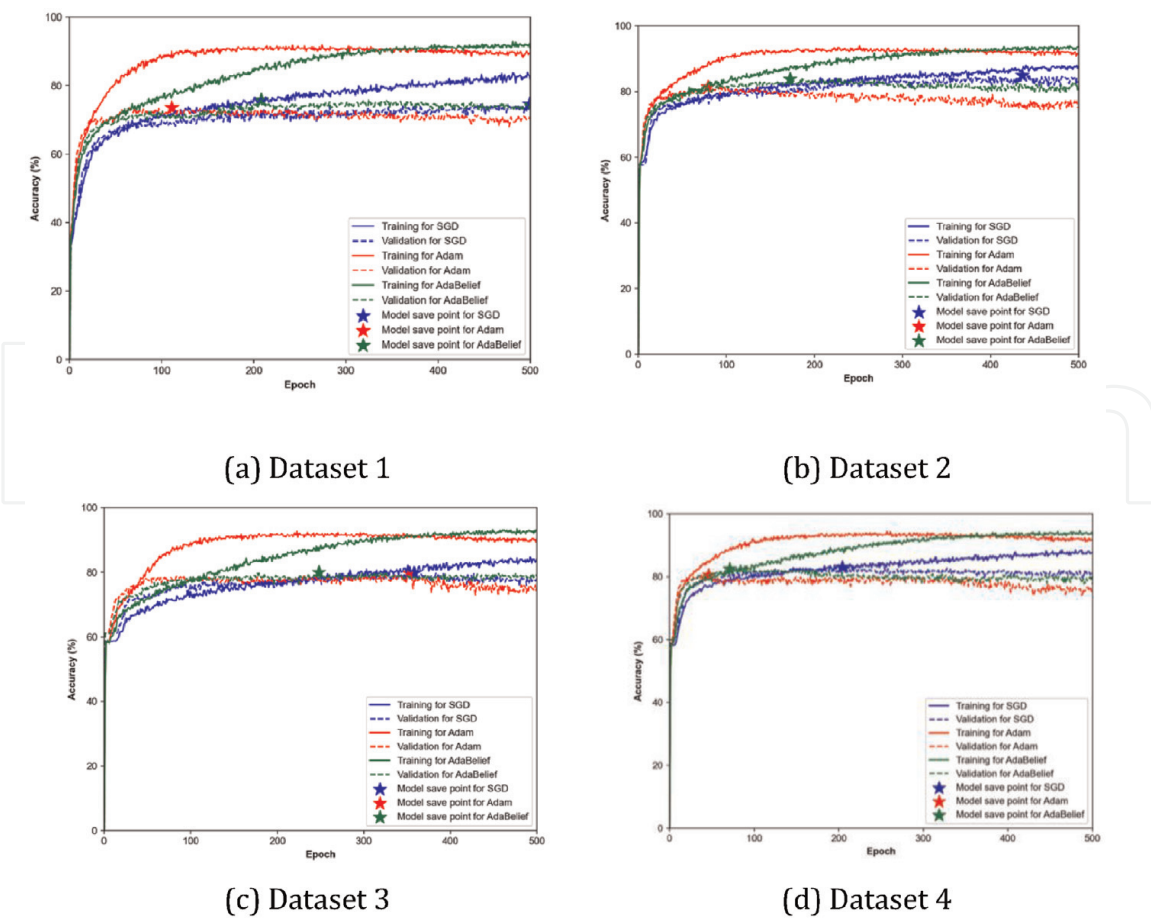


Figure 4.
Learning history of AT-GRU.

validation accuracy was the largest, which indicated that the worst generalization of Adam among three optimizers in this study although it converged quickly at the very beginning except for Dataset 3. Both SGD and AdaBelief can achieve good predictive performance with good generalization, however, AdaBelief converged faster than SGD, and the best results were achieved by AdaBelief.

5. Conclusions and future extensions

In this study, the hotel review dataset collected from TripAdvisor for aspect-level sentiment classification was first established. The dataset contains 5506 sentences in which the numbers of positive, neutral, and negative sentiment samples are 3032, 2986, and 2725, respectively. In order to study the effect of the fraction of sentiment samples on the model performance, four sub-datasets with a various fraction of sentiment samples were resampled from the TripAdvisor hotel review dataset as the train sets. The task in this study is to determine the aspect polarity of a given review with the corresponding aspects. To achieve a good predictive performance toward a multi-class classification task, attention-based GRU and LSTM (AT-GRU and AT-LSTM), as well as attention-based GRU and LSTM with aspect embedding (ATAE-GRU and ATAELSTM), were optimized with SGD, Adam, and AdaBelief and trained with epochs of 100, 300, and 500, respectively. Conclusions from these experiments are as follows:

1. AT-GRU and AT-LSTM performed better than ATAELSTM and ATAELSTM. Taken the balanced dataset as an example, the best accuracy in the test set using AT-GRU was 80.7%, while the best accuracy using ATAELSTM was 75.3%.
2. Attention-based GRU performed better than attention-based LSTM. Taken AT-GRU and AT-LSTM for example, the accuracy and macro F1-score of AT-GRU for all datasets were higher than those of AT-LSTM.
3. The balanced dataset achieved the best predictive performance. For the unbalanced datasets, the accuracy was exactly close to that of the balanced dataset, however, the macro precision, recall, and F1-score were significantly lower than those of the balanced dataset, which confirmed that the balanced dataset had the best generalization and stability in this study. For the dataset in which the neutral sentiment samples dominated, all of the models exhibited the worst predictive performance.
4. For the balanced dataset, the accuracy and macro F1-score obtained by early stopping was significantly lower than that obtained by the corresponding model without early stopping. However, for the unbalanced datasets, the accuracy and macro F1-score obtained by early stopping were similar to that obtained by the corresponding model without early stopping, which indicated that early stopping was effective to avoid overfitting as the loss converged fast in the unbalanced datasets.
5. For optimizers, both SGD and AdaBelief can achieve good predictive performance with good generalization, however, AdaBelief converged faster than SGD, and the best results were achieved by AdaBelief.

This work includes the application of natural language processing technologies on the aspect-level sentiment analysis of the TripAdvisor hotel dataset, and there are still several extensions to be explored as follows:

1. Enlargement of the dataset. This study focused on the hotel in Macau, collecting 5506 reviews from TripAdvisor. To improve the model performance, hotels from other countries and regions can be collected into the dataset.
2. Improvement of model performance, especially for the predictive capacity of the neutral sentiment samples. The sentence and aspect embeddings were initialized with GloVe, and BERT which is popular in recent research can be used. In addition, although the attention mechanism was used in this study to improve model performance, the state-of-art self-attention mechanism such as multi-head attention can be used in the future to further refine the model.
3. Development of a mobile application. Once the model with stable performance is achieved, RNN algorithms can be integrated into a portable device such as a smartphone to help with real-time aspect-level sentiment analysis in tourism.

Author details


Weijun Li¹, Qun Yang² and Wencai Du^{1*}

1 City University of Macau, Macao, SAR, China

2 Department of Civil and Environmental Engineering, The University of Auckland, Auckland, New Zealand

*Address all correspondence to: georgedu@cityu.mo

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Gretzel U, Yoo KH. Use and impact of online travel reviews. In: *Information and Communication Technologies in Tourism 2008*. Vienna: Springer Vienna; 2008. p. 35–46.
- [2] Li H, Ye Q, Law R. Determinants of customer satisfaction in the hotel industry: An application of online review analysis. *Asia Pac J. Tour Res.* 2013;18(7): 784–802.
- [3] Choi S, Lehto XY, Morrison AM. Destination image representation on the web: Content analysis of Macau travel related websites. *Tour Manag.* 2007;28(1):118–29.
- [4] García-Pablos A, Cuadros M, Linaza MT. OpeNER: Open tools to perform natural language processing on accommodation reviews. In: *Information and Communication Technologies in Tourism 2015*. Cham: Springer International Publishing; 2015. p. 125–37.
- [5] Kang H, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Syst Appl.* 2012;39(5): 6000–10.
- [6] Zheng W, Ye Q. Sentiment classification of Chinese traveler reviews by support vector machine algorithm. In: *2009 Third International Symposium on Intelligent Information Technology Application*. IEEE; 2009.
- [7] Zhang Z, Ye Q, Zhang Z, Li Y. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Syst Appl.* 2011;38(6): 7674–82.
- [8] Liu B. Opinion Mining and Sentiment Analysis. In: *Web Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 459–526.
- [9] Do HH, Prasad PWC, Maag A, Alsadoon A. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Syst Appl.* 2019;118:272–99.
- [10] Schouten K, Frasincar F. Survey on aspect-level sentiment analysis. *IEEE Trans Knowl Data Eng.* 2016;28(3): 813–30.
- [11] Luo J, Huang S (sam), Wang R. A fine-grained sentiment analysis of online guest reviews of economy hotels in China. *J Hosp Mark Manag.* 2021;30(1): 71–95.
- [12] Sheng F, Zhang Y, Shi C, Qiu M, Yao S. Xi'an tourism destination image analysis via deep learning. *J. Ambient Intell Humaniz Comput* [Internet]. 2020; Available from: <http://dx.doi.org/10.1007/s12652-020-02344-w>
- [13] Li X, Bing L, Lam W, Shi B. Transformation networks for target-oriented sentiment classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018.
- [14] Zhou J, Huang JX, Chen Q, Hu QV, Wang T, He L. Deep learning for aspect-level sentiment classification: Survey, vision, and challenges. *IEEE Access.* 2019; 7:78454–83.
- [15] Gu, S. Q., Zhang, L. P., Hou, Y. X., & Song, Y. A. A Position-aware Bidirectional Attention Network for Aspect-Level Sentiment Analysis. *Proceedings of the 27th International Conference on Computational Linguistics.* 2018;774–84.
- [16] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word

representations in vector space [Internet]. arXiv [cs.CL]. 2013. Available from: <http://arxiv.org/abs/1301.3781>

[17] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014.

[18] Poria S, Chaturvedi I, Cambria E, Bisio F. Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE; 2016.

[19] Wu H, Gu Y, Sun S, Gu X. Aspect-based opinion summarization with convolutional neural networks. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE; 2016.

[20] Liu P, Joty S, Meng H. Fine-grained opinion mining with recurrent neural networks and word embeddings. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2015.

[21] Bengio, Y., Schwenk, H., Senécal, J. S., Morin, F. M., & Gauvain, J. L. *Neural Probabilistic Language Models*. Heidelberg: Springer; 2006.

[22] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L, editor. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018. p. 2227–37.

[23] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language

understanding [Internet]. arXiv [cs.CL]. 2018. Available from: <http://arxiv.org/abs/1810.04805>

[24] Feng J, Cai S, Ma X. Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm. *Cluster Comput*. 2019;22(S3):5839–57.

[25] Jebbara S, Cimiano P. Aspect-based sentiment analysis using a two-step neural network architecture. In: *Semantic Web Challenges*. Cham: Springer International Publishing; 2016. p. 153–67.

[26] MA, Y., Peng, H. Y., & Cambria, E. Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. In: *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. 2018. p. 5876–83.

[27] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–80.

[28] Goldberg Y. Neural network methods for natural language processing. *Synth lect hum lang technol*. 2017;10(1):1–309.

[29] Poria S, Cambria E, Gelbukh A, Bisio F, Hussain A. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Comput Intell Mag*. 2015; 10(4):26–36.

[30] Xue W, Li T. 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)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2018.

- [31] Fan C, Gao Q, Du J, Gui L, Xu R, Wong K-F. Convolution-based memory network for aspect-based sentiment analysis. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*. New York, New York, USA: ACM Press; 2018.
- [32] Xu L, Lin J, Wang L, Yin C, Wang J. Deep convolutional neural network-based approach for aspect-based sentiment analysis. In *Science & Engineering Research Support Society*; 2017.
- [33] Gu X, Gu Y, Wu H. Cascaded convolutional neural networks for aspect-based opinion summary. *Neural Process Lett*. 2017;46(2):581–94.
- [34] Goldberg Y. A primer on neural network models for natural language processing. *J. Artif Intell Res*. 2016;57:345–420.
- [35] Bengio Y. *Deep Learning*. London, England: MIT Press; 2016.
- [36] Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2016.
- [37] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. 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)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014.
- [38] Graves A. Supervised sequence labelling. In: *Studies in Computational Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 5–13.
- [39] Fan, Yuchen. Qian, Yao. Xie, Feng-Long. Soong, Frank K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In: *INTERSPEECH-2014*. 2014. p. 1964–8.
- [40] Chaudhuri A, Ghosh SK. Sentiment analysis of customer reviews using robust hierarchical bidirectional recurrent neural network. In: *Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing; 2016. p. 249–61.
- [41] Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst Appl*. 2017;72:221–30.
- [42] Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2015.
- [43] Wang Y, Huang M, Zhu X, Zhao L. Attention-based LSTM for Aspect-level Sentiment Classification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2016.
- [44] Zeng J, Ma X, Zhou K. Enhancing attention-based LSTM with position context for aspect-level sentiment classification. *IEEE Access*. 2019;7:20462–71.
- [45] He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. Effective attention modeling for aspect-level sentiment classification. In: *Proceedings of the 27th International Conference on*

Computational Linguistics. 2018.
 p. 1121–31.

[46] Wang, W., Pan, S. J., & Dahlmeier, D. Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. 2017.

[47] Cheng J, Zhao S, Zhang J, King I, Zhang X, Wang H. Aspect-level sentiment classification with HEAT (HiErarchical ATtention) network. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*. New York, New York, USA: ACM Press; 2017.

[48] Wang S, Mazumder S, Liu B, Zhou M, Chang Y. Target-sensitive memory networks for aspect sentiment classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2018.

[49] Chang Y-C, Ku C-H, Chen C-H. Using deep learning and visual analytics to explore hotel reviews and responses. *Tour Manag.* 2020;80(104129):104129.

[50] Gao J, Yao R, Lai H, Chang T-C. Sentiment analysis with CNNs built on LSTM on tourists' comments. In: *2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*. IEEE; 2019.

[51] Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014.

[52] Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2015.

[53] Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the Stars: Improving Rating Predictions using Review Text Content [Internet]. Available from: Beyond the Stars: Improving Rating Predictions using Review Text Content. Twelfth International Workshop on the Web and Databases./<http://spidr-ursa.rutgers.edu/resources/WebDB.pdf>

[54] Moreno-Ortiz A, Salles-Bernal S, Orrequia-Barea A. Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector. *Inf Technol Tour.* 2019;21(4):535–57.

[55] Nitish Srivastava Geoffrey Hinton Alex Krizhevsky Ilya Sutskever Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research.* 2014;15:1929–58.

[56] Polyak BT. Some methods of speeding up the convergence of iteration methods. *USSR Comput Math Math Phys.* 1964;4(5):1–17.

[57] Kingma DP, Ba J. Adam: A method for stochastic optimization [Internet]. arXiv [cs.LG]. 2014. Available from: <http://arxiv.org/abs/1412.6980>

[58] Zhuang J, Tang T, Ding Y, Tatikonda S, Dvornek N, Papademetris X, et al. AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients [Internet]. arXiv [cs.LG]. 2020. Available from: <http://arxiv.org/abs/2010.07468>