Understanding Patient Reviews with Minimum Supervision

Lin Gui, Yulan He*

Department of Computer Science, University of Warwick, UK

Abstract

Understanding patient opinions expressed towards healthcare services in online platforms could allow healthcare professionals to respond to address patients' concerns in a timely manner. Extracting patient opinion towards various aspects of health services is closely related to aspect-based sentiment analysis (ABSA) in which we need to identify both opinion targets and target-specific opinion expressions. The lack of aspect-level annotations however makes it difficult to build such an ABSA system. This paper proposes a joint learning framework for simultaneous unsupervised aspect extraction at the sentence level and supervised sentiment classification at the document level. It achieves 98.2% sentiment classification accuracy when tested on the reviews about healthcare services collected from Yelp, outperforming several strong baselines. Moreover, our model can extract coherent aspects and can automatically estimate the distribution of aspects on different polarities without requiring aspect-level annotations for model learning.[1]

Keywords: Sentiment analysis, aspect extraction, patient reviews.

1. Introduction

A key tool to support the delivery of high quality healthcare is near real-time collection of patient feedback and timely response to address patients' concerns [1, 2]. Hospitals and healthcare providers have implemented various forms of mechanism for feedback collection, including feedback cards, hand-held electronic devices for collecting patients' opinions at the point of care, and dedicated mobile apps for capturing feedback. In addition, users can also post their comments to independent feedback platforms such as Yelp[2], Care

*Corresponding author.

Email addresses: `Lin.Gui@warwick.ac.uk` (Lin Gui), `Yulan.He@warwick.ac.uk` (Yulan He)

[1] Our code and dataset can be accessed at `https://github.com/GuiLinNLP/PatientReviewUnderstanding`.

[2] `https://www.yelp.com`

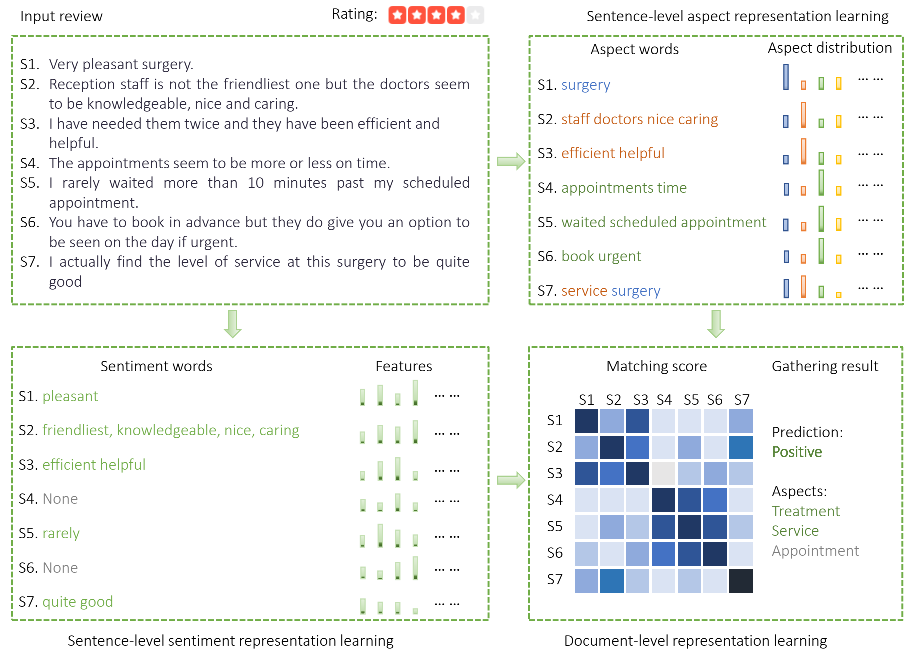
Opinion[3], Facebook pages of health service providers and Twitter [3, 4]. Such unstructured information, however, is not captured in a systematic way. Data on online platforms about patients' experiences largely go un-monitored. This represents a missed opportunity for understanding patients' experience in an increasingly "connected" world.

There is an urgent need of developing an automated tool to process large-scale online data in order to understand patient feedback so as to enable health-care professionals to address patients' concerns in a timely manner. Sentiment analysis methods have been applied to patient surveys [5] to classify patient comments as *positive* or *negative* polarities. It is however more desirable to go beyond sentiment classification to automatically map the extracted opinions into various aspects of healthcare services and discover connections between elements that result in a perception of low and high quality of services.

Extracting patient opinion towards various aspects of healthcare services is closely related to Aspect-Based Sentiment Analysis (ABSA) [6, 7, 8] in which both opinion targets and target-specific opinion expressions need to be identified. Existing sentence-level ABSA studies typically include three tasks: (1) *opinion target or aspect term extraction*, which aims to locate the target expression from input text towards which an opinion is expressed; (2) *aspect category classification*, which aims to identify aspect mentions from input sentences and classify them into pre-defined aspect categories; and (3) *aspect-term or aspect-category sentiment detection*, which aims to assign a polarity label to an aspect term/category identified in a sentence. Building an effective ABSA model, however, usually requires fine-grained annotations of both aspects and aspect-associated sentiments at the sentence level, involving heavy manual efforts. Furthermore, different from ABSA on product reviews in which the number of aspect/property categories is limited, patient reviews could discuss a wide variety of topics which makes it difficult to pre-define an aspect category schema for annotation. In this paper, we explore an alternative approach in which we aim to automatically extract aspects discussed in patient reviews without requiring fine-grained aspect-level annotations for training. The only supervision information provided for model training is document-level sentiment labels which can be easily obtained from rating scores of patient reviews. With the learned latent aspect distribution for each sentence, we are able to derive a better document representation taking into account the relative contributions of its constituent sentences calculated based on the similarity measured on their aspect representations.

Most existing sentiment analysis methods assume sentence- or document-level sentiment labels are largely determined by sentiment words occurred in text. This is, however, not always the case. The co-occurrence of a sentiment word with different aspect words may result in different polarities. For example, when the word '*high*' is used to describe the '*treatment cost*', it conveys a *negative* sentiment; but when it co-occurs with '*service quality*', it expresses a

[3] `https://www.careopinion.org.uk/`



S1. Very pleasant surgery.
S2. Reception staff is not the friendliest one but the doctors seem to be knowledgeable, nice and caring.
S3. I have needed them twice and they have been efficient and helpful.
S4. The appointments seem to be more or less on time.
S5. I rarely waited more than 10 minutes past my scheduled appointment.
S6. You have to book in advance but they do give you an option to be seen on the day if urgent.
S7. I actually find the level of service at this surgery to be quite good

Figure 1: Example outputs generated by our proposed approach on a patient review from Yelp. **Upper left**: a review document with its rating score; **Upper right**: aspect words are automatically identified and grouped into various aspect topics by the *Sentence-Level Aspect Representation Learning* module, which also infers the latent aspect distribution for the sentence; **Lower left**: sentiment words and identified and context feature representation of the sentence is learned by the *Sentence-Level Sentiment Representation Learning* module; **Lower right**: outputs from the aforementioned two modules are fed to the *Document-Level Representation Learning* module to generate the final sentiment classification result. The matching score matrix shows the pairwise similarity of sentences calculated based on their latent aspect distributions. Sentences are segmented into two groups: {S1, S2, S3} are about *treatment* and *service*, and {S4, S5, S6} are about *appointment*.

positive polarity. In this paper, therefore, we propose a joint learning framework for simultaneous aspect extraction and sentiment classification. For a better understanding of our proposed method, we illustrate an example in Figure 1 in which a review document consisting of seven sentences is fed to our model for the prediction of a sentiment label. The upper right box in Figure 1 shows the output of our *Sentence-Level Aspect Representation Learning* module, which automatically identifies aspect words from each sentence and learns a latent aspect distribution as shown by the colourful histograms. Here, different colours denote different aspects. Note that our model can gather relevant aspect words into an aspect topic in an unsupervised way. The lower left box in Figure 1 shows the output from the *Sentence-Level Sentiment Representation Learning Module*. Words listed for each sentence are extracted based on their attention weights.

We can observe that these words are mostly indicative of polarity. The lower right box in Figure 1 shows the output from the *Document-Level Representation Learning* module which processes the outputs generated from the previous two modules to produce the final sentiment classification result. The matching score matrix shows the pairwise similarity of two sentences measured based on their latent aspect distributions. It is clear that sentences are segmented into two groups, S1, S2 and S3 are about *treatment* and *service*, while S4, S5 and S6 are about *appointment*. The final predicted output is '*positive*'.

The rest of the paper is organised as follows. Section 2 presents a review of related work in patient experience understanding and aspect-based sentiment analysis. Section 3 describes our proposed joint-learning framework. Section 4 explains how to generate interpretations from the outputs of the proposed method. Section 5 discusses experimental setup and evaluation results. Finally, Section 6 concludes the paper and outlines the future research directions.

2. Related Work

Our work is related to three lines of research: patient experience understanding, sentiment classification, and aspect extraction for sentiment analysis.

2.1. Understanding Patient Experience

Despite the use of a wide variety of mechanisms for the collection of patient feedback and experience at both national and local levels, the translation of measures of patient experience into improvements in the quality of healthcare is largely under-exploited [9, 10, 11]. This is partly due to the limited scope of survey questions being used that the vast majority focus on specific services rather than patient journeys [12]. Ideally, survey instruments should be tailored to local care services and to each clinical topic. This is however not practical since it would require question sets tailored to every diagnostic step or decision point in the process and most healthcare organisations do not have resource or capacity to undertake this kind of work. Gibbons et al. [13] proposed to develop a standardised questionnaire of patient experience that could work across a range of services and pathways of care. Their approach however requires significant involvement and substantial co-design and collaboration with various healthcare sites and is thus expensive to deploy in scale.

Nevertheless, patient experience data are scattered across multiple information sources including patient surveys, text messages, emails, voice recordings (through landline survey), online review sites, news reports and social media sites. Such heterogeneous sources contain rich information about users' personal experiences and offer a great potential for automatic collection of patient feedback in real-time. Existing work on analysing unstructured data about quality of healthcare is very small in scale and mostly relies heavily on simple keyword frequency counting, sentiment analysis using pre-existing sentiment lexicons or manual processing. For example, Maramba et al. [14] studied the association of patients' experience scores with the occurrence of certain words from 3,426 free-text responses in the postal survey of patients about 25 English general practices. Rastegar-Mojarad et al. [15] created the Corpus of Patient Experience (COPE) which contains 6,914 patient reviews from the Yelp website and performed word statistics analysis and correlated sentiment scores derived based on a sentiment lexicon with user-generated review ratings. Greaves et al. [16] used sentiment analysis to categorise over 6,000 online comments by patients about hospitals on the UK's National Healthcare Services (NHS) website as either positive or negative. Hawkins et al. [17] crawled tweets directed to US hospitals over a one-year period and manually categorised just over 10k tweets into patient experience topics.

Existing approaches either focus on sentiment classification of patient reviews or rely on manual categorisation of patient experience topics, which are not feasible as generalised methodologies. There is thus an urgent need to develop a tool which is able to automatically extract and make sense of patient experience data from online sources with minimum supervision and at the same time is able to understand patient experience at a much more granular level.

2.2. Sentiment Classification

With the rapid development of deep learning, various architectures of neural networks including Convolution Neural Network (CNN) [18] and Recurrent Neural Network (RNN) [19] have been proposed for sentiment classification. In document-level sentiment classification, the hierarchical semantic composition of a document can be modeled by hierarchical models such as the Gated Recurrent Neural Network [20] and the Hierarchical Attention Network [21]. Various attention mechanisms have also been explored for sentiment analysis. Chen et al. [22] used the author and product information to obtain the attention signals with a hierarchical neural network for sentiment classification on product reviews. Tang et al. [23] deployed the memory network with attention for aspect-level sentiment classification. He et al. [24] stacked an auto-encoder layer in the memory network to obtain the aspect-specific attentions to improve the aspect-level sentiment classification. More recently, fine-tuning pre-trained language models such as BERT [25] on sentiment classification tasks gives superior performance compared to earlier approaches.

Besides the aforementioned methods, other complex neural modular networks [26, 27], prior knowledge obtained from a sentiment lexicon [28], and author profiles [29] have also been considered in sentiment classification. However, existing neural attention models largely derive the attention weights or signals based on lexical matching between the hidden state representation of the current lexical unit (word or sentence) and a common context vector shared across sentences or documents. They mostly ignore the latent topic information which captures the semantic information in the global context.

2.3. Aspect Extraction for Sentiment Analysis

In fine-grained sentiment analysis, the task of aspect extraction involves the identification of explicit product features in customer reviews. Nevertheless, most existing work casts aspect extraction as a sequence labelling problem that predicts an aspect label for each word token. In such a setup, data annotated with aspect labels at the word level are required for model training. Xu et al. [30] proposed a CNN-based method to tag each word token with as aspect label for input sentences. Wang et al. [31] deployed a method built on an adversarial network for cross-lingual aspect extraction. Li et al. [32] designed a Long-Short Terms Memory (LSTM) network with attentions to leverage the historical attention information to distill the knowledge about aspect terms. In real world applications, however, annotations of aspects at the word level are expensive to obtain. This motivates the development of unsupervised methods for aspect extraction from text. For example, He et al. [33] proposed an attention-based method to cluster word embeddings into different aspects. Chuanhan et al. [34] used a two-step hybrid unsupervised model with an attention mechanism for aspect extraction. Nevertheless, the aforementioned unsupervised methods require either linguistic knowledge or linguistic patterns to filter spurious aspect terms. Moreover, they cannot distinguish between positive and negative aspects. On the contrary, our proposed framework can not only extract aspects in an unsupervised way, but also separate them into positive and negative categories guided by document-level sentiment labels during the training.

3. Joint Aspect-Sentiment Extraction (JASE) Model

In this section, we will introduce our proposed learning framework for joint aspect-sentiment extraction, named as Jase. The overall architecture is illustrated in Figure 2. It contains the following main components: 1) A *sentence-level sentiment representation learning module* which aims to map words into a sentiment vector space based on attention signals; 2) A *sentence-level aspect representation learning module* which is built on an auto-encoder with a regulariser to cluster and map aspect words into a neutral sub-space; 3) A *transformer-based document-level representation learning module* which aims to match the two representations and extract the most relevant aspects. Finally, based on the matching scores, sentences are assigned with different weights in order to derive the document-level representations.

3.1. Sentence-Level Sentiment Representation Learning

In the *sentence-level sentiment representation learning module*, an attention-based Long-Short Term Memory (LSTM) network is used to capture the sentiment information conveyed in a sentence. In this model, the representation of each word is learned by its corresponding input embedding and the context through the LSTM unit. The attention mechanism is then employed to obtain the weight for each word and the sentence representation is generated based on a weighted sum of all the hidden representations of words.

Specifically, assuming that a document $\boldsymbol{w}_d$ contains M_d sentences, $\boldsymbol{w}_d = \{s_1, s_2, ...s_{M_d}\}$, and the word embedding of j-th word in i-th sentence is e_i^j.

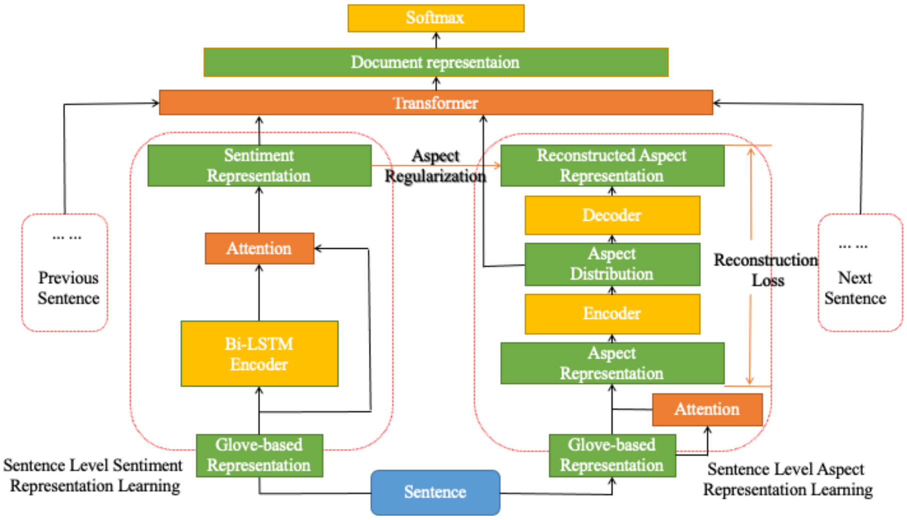


Figure 2: Jase: our proposed joint-learning architecture for sentiment classification and aspect extraction.

Then, the representation of sentence r_i, denoted as s_i, is obtained by:

$$\begin{aligned} x_i^j &= W \cdot e_i^j, \\ \overrightarrow{h}_i^j = \overrightarrow{\mathrm{LSTM}}(x_i^j), &\quad \overleftarrow{h}_i^j = \overleftarrow{\mathrm{LSTM}}(x_i^j), \quad h_i^j = \overrightarrow{h}_i^j \oplus \overleftarrow{h}_i^j, \\ u_i^j = \tanh(W_s \cdot h_i^j + b_s), &\quad \alpha_i^j = \frac{\exp(u^T \cdot u_i^j)}{\sum_t \exp(u^T \cdot u_i^t)}, \quad s_i = \sum_{j=1}^{n} \alpha_i^j \cdot h_i^j, \end{aligned} \tag{1}$$

where $e_i^j \in \mathbb{R}^{|V|}$ is the one-hot representation of the j-th word in i-th sentence, $|V|$ is the vocabulary size, $W \in \mathbb{R}^{|V| \times d}$ is the learnable word embedding matrix initialised by the GloVe [35] embedding, d is the dimension of word embeddings, $x_i^j \in \mathbb{R}^d$ is the learned word vector of the j-th word in i-th sentence, $\overrightarrow{\mathrm{LSTM}}$ and $\overleftarrow{\mathrm{LSTM}}$ are bi-directional LSTM, $W \in \mathbb{R}^{d \times d/2}$, $W_s \in \mathbb{R}^{d \times q}$, $b_s \in \mathbb{R}^q$ are learnable parameters, $h_i^j \in \mathbb{R}^{d/2}$ is the learned hidden representation of the j-th word in i-th sentence, $u_i^j \in \mathbb{R}^q$ is the attention vector of the j-th word in the i-th sentence, $u \in \mathbb{R}^q$ is a learned global vector to capture the attention signal[4] of α_i^j by u_i^j, and s_i is the learned sentiment representation for the i-th sentence in document $\boldsymbol{w}_d$.

[4] Essentially a two-layer perceptron is used for attention learning, where the first layer is parameterised by W_s with the bias term of b_s, and the second layer is parameterised by u without a bias term.

3.2. Sentence-Level Aspect Representation Learning

In the *sentence-level aspect representation learning module*, an attention-based autoencoder is used to generate the latent aspect-level representation of a sentence. A decoder is then used to reconstruct the sentence. We assume that the sentence-level sentiment representation learned by bi-LSTM as described in Section 3.1 captures the polarity information expressed in a sentence. On the contrary, we hope the sentence-level aspect representation learned by the autoencoder here does not carry any polarity information. As such, we want to make these two representations as different as possible. The aspect representation of a sentence r_i, denoted as z_i, can be obtained by the following steps:

$$v_i^j = \tanh(W_a \cdot w_i^j + b_a), \quad \beta_i^j = \frac{\exp(v^T \cdot v_i^j)}{\sum_t \exp(v^T \cdot v_i^t)}, \quad z_i = \sum_{j=1}^{n} \beta_i^j \cdot w_i^j, \tag{2}$$

where $w_i^j \in \mathbb{R}^d$ is the fixed word embedding from GloVe, $W_a \in \mathbb{R}^{d \times q}$, $b_a \in \mathbb{R}^{d/2}$, $v \in \mathbb{R}^q$ are learnable parameters, β_i^j is the attention signal of the i-th word in j-th sentence, $z_i \in \mathbb{R}^d$ is the aspect representation weighted by the attention signal of β_i^j. We then feed the aspect representation z_i for sentence r_i into an autoencoder to generate the reconstructed representation z_i':

$$h_i = \tanh(W_c \cdot z_i + b_c), \quad z_i' = \tanh(W_c' \cdot h_i + b_c'), \tag{3}$$

where $W_c \in \mathbb{R}^{d \times K}$, $b_c \in \mathbb{R}^K$, $W_c' \in \mathbb{R}^{K \times d}$, $b_c' \in \mathbb{R}^d$ are learnable parameters, K is the total number of aspects, and $h_i \in \mathbb{R}^K$ is the hidden aspect vector. Intuitively, the hidden aspect vector should capture the key semantic information from the input aspect representation, z_i, and can reconstruct the input through a decoder. Hence, in this paper, the learning objective is to minimise the hinge loss that maximises the inner product between z_i and z_i' (the input aspect representation should be close to its reconstructed aspect representation), and simultaneously minimises the inner product between the reconstructed aspect representation z_i' and the sentiment representation s_i (which is equivalent to maximising the margin between the aspect representation and the sentiment representation):

$$\mathcal{L}_a(\boldsymbol{w}_d) = \sum_i \max(0, 1 - z_i \cdot z_i' + z_i' \cdot s_i) \tag{4}$$

3.3. Regularisation Term

The learned aspects might be redundant, i.e., different aspects might contain many overlapping words. To ensure the diversity of the resulting aspects learned, inspired by the Tikhonov regularisation [36], which is widely used in many related works to guarantee the uniqueness of latent distributions [33], we add a regularisation term to the objective function to encourage the uniqueness of the aspect embeddings:

$$\mathcal{R}_a(\boldsymbol{w}_d) = \left\| W_c' \cdot W_c'^T - I \right\|_2, \tag{5}$$

where I is the identity matrix, and W_c' is the normalised decoder matrix, where each column can be considered as the representation of an aspect. $\mathcal{R}_a(\boldsymbol{w}_d)$ reaches its minimum value when the dot product between any two different aspect representations is zero. Thus the regularisation term encourages orthogonality among the columns of the aspect decoder matrix W_c' and penalises redundancy between different aspect vectors. The objective function for aspect extraction L_a is defined by:

$$\mathcal{L}_a(\boldsymbol{w}_d) = \sum_i \max(0, 1 - z_i \cdot z_i' + z_i' \cdot s_i) + \lambda \cdot \mathcal{R}_a(\boldsymbol{w}_d) \tag{6}$$

where λ is a hyperparameter that controls the weight of the regularisation term.

3.4. Transformer-Based Document-Level Representation Learning

In most existing research, the representation of a document $\boldsymbol{w}_d$ is learned with a similar architecture as in sentence representation learning, taking the input as a sequence of sentence representations. A softmax layer is then stacked at the top to predict the class label of a document by the cross entropy loss between the predicted label and the true label. We, however, argue that the weight of each sentence is not only determined by the sentiment information, but also by the relationship between sentiment words and the associated aspect words. For example, in a patient review, complaints about treatment or price might have a higher weight than parking services when contributing to the final sentiment classification results. Hence, in this work, we use a transformer-based model to capture the relation between the learned sentiment representations and the aspect representations. We assume that the learned sentiment representation and the aspect representation for i-th sentence are s_i and z_i', respectively. We first feed the sentiment representation to a bi-directional LSTM to obtain the hidden representation for each sentence:

$$\overrightarrow{g_i} = \overrightarrow{\text{LSTM}}(s_i), \quad \overleftarrow{g_i} = \overleftarrow{\text{LSTM}}(s_i), \quad g_i = \overrightarrow{g_i} \oplus \overleftarrow{g_i}, \tag{7}$$

Then, we use the matching score between s_i and h_i to learn the attention of the hidden state g_i by the transformer:

$$\gamma_i = \text{softmax}\big(\frac{s_i \cdot f(h_i)}{\sqrt{d}}\big), \tag{8}$$

where $\gamma_i \in \mathbb{R}$ is the attention signal of i-th sentence, d is the dimension of the sentiment representation and $f(\cdot)$ is a linear mapping function which maps the hidden state of an aspect into the vector space of sentiment representations for matching. Next, we obtain the document representation by a weighted sum:

$$f_d = \sum_{i=1}^{n} \gamma_i \cdot g_i \tag{9}$$

The learned document representation is then fed into a softmax classifier for sentiment classification as follow:

$$\hat{y} = \mathrm{softmax}(W_d \cdot f_d + b_d), \quad \mathcal{L}_c(\boldsymbol{w}_d) = -\sum_{t=1}^{T} y_t \cdot \log(\hat{y}_t), \tag{10}$$

where $\hat{y}$ is the distribution of the predicted label for a document and y is the distribution of the true label, $W_d \in \mathbb{R}^{d \times T}$ and $b_d \in \mathbb{R}^T$ are learnable parameters in the sentiment classifier, T is the number of class labels. The final loss function for joint learning is a weighted sum of the cross entropy loss for sentiment classification and the aspect extraction loss:

$$\mathcal{L}_{final}(\boldsymbol{w}_d) = \eta_a \cdot \mathcal{L}_a + \eta_c \cdot \mathcal{L}_c \tag{11}$$

where η_a and η_c are the weights to control the contribution of the aspect extraction loss and the sentiment classification loss to the final objective function, respectively.

Input:	A document $\boldsymbol{w}_d$ consisting of M_d sentences, with each sentence consisting of L words: $\boldsymbol{w}_d = \{r_i\}_{i=1}^{M_d}$, $s_i = \{e_i^j\}_{j=1}^{L}$	
Word Emb:	GloVe [35] Init.	$x_i^j = W \cdot e_i^j$, $e_i^j \in \mathbb{R}^{\lvert V \rvert}$, $W \in \mathbb{R}^{\lvert V \rvert \times d}$, $\lvert V \rvert = 15,000$, $d = 300$
Senti. learn:	Word-level Bi-LSTM	$\{x_i^j\}_{j=1}^{L} \rightarrow \{\text{Bi-LSTM}_1\}_{300\rightarrow300} \rightarrow \{\mathrm{h_i^j}\}_{\mathrm{j=1}}^{\mathrm{L}} \in \mathbb{R}^{\mathrm{d\times L}}$, $d = 300$
	Attention layer	$\{h_i^j\}_{j=1}^{L} \rightarrow \{\text{Linear}_1\}_{300\rightarrow100} \rightarrow \{\text{Linear}_2\}_{100\rightarrow1} \rightarrow \{\alpha_\mathrm{i}^\mathrm{j}\}^\mathrm{L} \in \mathbb{R}^\mathrm{L}$
	Context aggregate	$\sum_{j=1}^{L} \alpha_i^j \cdot h_i^j \rightarrow s_i \in \mathbb{R}^d$, $d = 300$.
Aspect learn:	GloVe[35] init.	$w_i^j \in \mathbb{R}^d$, $d = 300$
	Autorenoder	$\{w_i^j\}_{j=1}^{L} \rightarrow \{\text{Linear}_3\}_{300\rightarrow100} \rightarrow \{\text{Linear}_4\}_{100\rightarrow1} \rightarrow \{\beta_\mathrm{i}^\mathrm{j}\}^\mathrm{L} \in \mathbb{R}^\mathrm{L}$ $\sum_{j=1}^{L} \beta_i^j \cdot w_i^j \rightarrow z_i \in \mathbb{R}^d$, $d = 300$ $z_i \rightarrow \text{Tanh}(\text{Linear}_5)_{300\rightarrow50} \rightarrow \{\mathrm{h_i} \in \mathbb{R}^\mathrm{K}\} \rightarrow \text{Tanh}(\text{Linear}_6)_{50\rightarrow300} \rightarrow \mathrm{z_i'} \in \mathbb{R}^\mathrm{d}$, $d = 300, K = 50$
Doc Modeling	Sent. level Bi-LSTM	$\{s_i\}_{i=1}^{M_d} \rightarrow \{\text{Bi-LSTM}_2\}_{300\rightarrow300} \rightarrow \{\mathrm{g_i}\}_{\mathrm{i=1}}^{\mathrm{M_d}} \in \mathbb{R}^{\mathrm{d\times M_d}}$, $d = 300$
	Self att. layer	$\text{softmax}\big(\frac{s_i \cdot f(h_i)}{\sqrt{d}}\big) \rightarrow \gamma_i$
	Doc. aggregate.	$\sum_{i=1}^{M_d} \gamma_i \cdot g_i \rightarrow f_d \in \mathbb{R}^d$, $d = 300$
Classification		$y = \text{Softmax}\Big(\{\text{Linear}_7\}_{300\rightarrow2}(\mathrm{f_d})\Big)$

Table 1: Model architecture and parameter setup. (For each unit of the neural network, we denote its input dimensions and output dimensions as the subscript.)

Parameter setup: In our implementation, the dimension of the aspect distribution, i.e., the number of aspect topics, is set to 50; the dimension of the Bi-LSTM hidden layers is 300 (150 for the forward and the backward directions, respectively); the dimension of both sentence-level sentiment representations and aspect-representations is also set to 300, making it possible to compute the dot product of these two representations in the regularisation term. We choose 300-dimensional pre-trained GloVe word embeddings [35] as input for both sentiment representation learning and aspect representation learning. The value of λ in the regularisation term is set to 0.1, η_a and η_c, which control the contribution of the aspect extraction loss and the sentiment classification loss, are set to 0.1 and 0.9, respectively, the learning rate is $1e-4$, and the dropout rate is 0.4. More details of the model architecture and the parameter dimensions can be found in Table 1.

4. Interpretation Generation

Section 3 describes the Jase model which learns sentence-level sentiment and aspect representations separately and aggregates the sentence-level sentiment representations weighted by the similarities between the sentence-level aspect representations in order to derive the document-level representation for sentiment classification. In this section, we explain how to generate a list of words representing each aspect; how to derive the aspect-level sentiment label; and how to infer the weight of each aspect under different sentiment categories.

Aspect Topic Description. In sentence-level aspect representation learning, each sentence s_i in a document is mapped to a hidden aspect distribution, $h_i = \{h_{i1}, h_{i2}, ..., h_{iK}\}$, where K denotes the total number of aspects. Each of its elements essentially represents the probability that the input sentence contains the k-th aspect A_k, i.e., $P(A_k|s_i) \propto h_{ik}$. Let W_{ck} be the k-th column of the decoder matrix $W_c = \{W_{c1}, W_{c2}, ..., W_{cK}\}$, and $b'_c = \{b_{c1}, b_{c2}, ..., b_{cK}\}$ the bias term, we have:

$$h_{ik} = \frac{\exp(W_{ck}^{\intercal} \cdot \sum_j (x_j \cdot P(x_j|s_i)) + b_{ck})}{\sum_m \exp(W_{cm}^{\intercal} \cdot \sum_j (w_j \cdot P(x_j|s_i) + b_{cm})}, \tag{12}$$

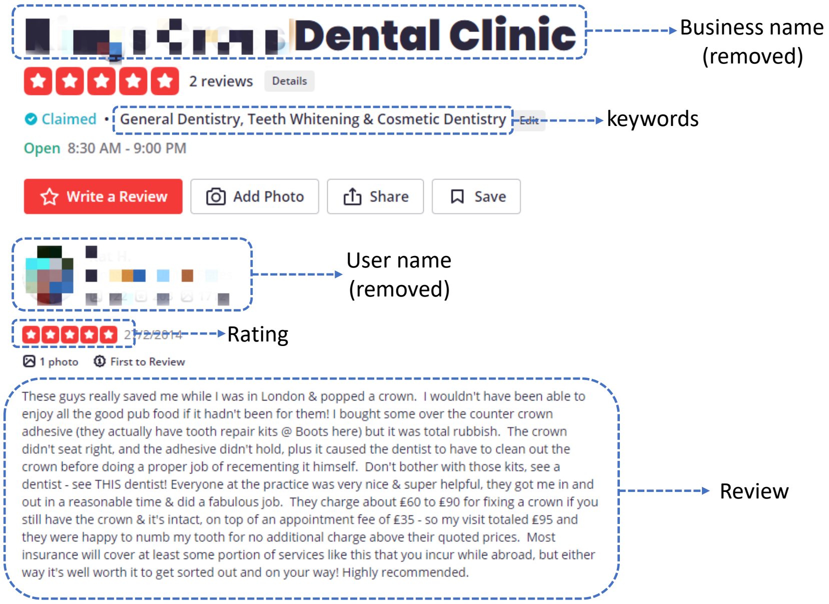
where x_j is the embedding of the j-th word in the i-th sentence, and $P(x_j|s_i) \propto \beta_i^j$, since the activation function in our network is a bijection. Hence, we can simply take $P(A_k|x_j) \propto W_{ck}^{\intercal} \cdot x_j$. That is, we can use the corresponding column in the encoder matrix to search the whole vocabulary in order to retrieve the top-n words associated with each aspect topic.

Aspect-Level Sentiment Label Generation. According to the central limit theorem, the inner product of $W_{ck}^{\intercal} \cdot x_j$ approximates to a cosine similarity between these two vectors. Hence, we can treat W_{ck} as the vector of aspect k, which is the centroid of its associated words grouped under the aspect k in the word embedding space. To obtain the sentiment label of aspect k, we can feed its representation, W_{ck}, into the final sentiment classification layer to get the result.

Aspect Weight Derivation under Different Sentiment Categories. In patient reviews, aspects may weigh differently in the positive and negative reviews. For example, in positive reviews, patients may express their gratitude to the services and treatments they received and care less about high costs. To measure the importance of each aspect, we first feed the one-hot representation of the k-th aspect where the k-th dimension is 1 and other dimensions are set to zero, into the document-level transformer to search for the best-matched sentence representation. Such a sentence representation is then fed into the softmax layer to obtain the importance score, which measures how important the aspect is for the final document-level sentiment classification.

5. Experiments

5.1. Data Collection



[redacted] Dental Clinic

2 reviews

Claimed • General Dentistry, Teeth Whitening & Cosmetic Dentistry

Open 8:30 AM - 9:00 PM

[redacted]

2/2/2014

These guys really saved me while I was in London & popped a crown. I wouldn't have been able to enjoy all the good pub food if it hadn't been for them! I bought some over the counter crown adhesive (they actually have tooth repair kits @ Boots here) but it was total rubbish. The crown didn't seat right, and the adhesive didn't hold, plus it caused the dentist to have to clean out the crown before doing a proper job of recementing it himself. Don't bother with those kits, see a dentist - see THIS dentist! Everyone at the practice was very nice & super helpful, they got me in and out in a reasonable time & did a fabulous job. They charge about £60 to £90 for fixing a crown if you still have the crown & it's intact, on top of an appointment fee of £35 - so my visit totaled £95 and they were happy to numb my tooth for no additional charge above their quoted prices. Most insurance will cover at least some portion of services like this that you incur while abroad, but either way it's well worth it to get sorted out and on your way! Highly recommended.

Figure 3: An example of patient review.

Reviews from Yelp[5] relevant to healthcare services were collected for evaluation. As shown in Figure 3, each review is accompanied with a user rating ranging from 1 to 5 stars, which can be used as its sentiment class label, and several keywords highlighting the healthcare categories of the review[6]. Yelp has an Application Programming Interface (API) for accessing Yelp reviews by query terms. We used this API to retrieve patient reviews based on a set of predefined

[5] `https://www.yelp.com/dataset`

[6] Sensitive information such as the name of the business and username have been removed from the collected data.

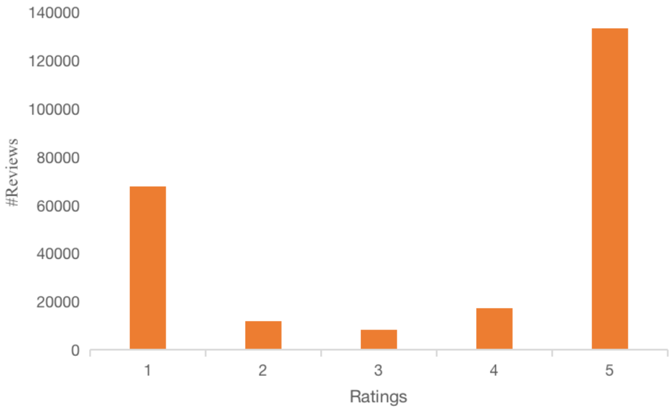


Figure 4: The distribution of ratings in the collected data. **Horizontal axis**: the rating stars of the collected reviews. **Vertical axis**: the number of reviews under the corresponding rating stars.

keywords which are claimed by the business owners[7]. We have retrieved a total of 238,796 reviews. Their rating distributions are shown in Figure 4. Since the majority of reviews have ratings of either 1 or 5 stars, we only keep the reviews with 1 and 5 stars as negative and positive instances, respectively, and discard reviews with other ratings. This results in a total of 201,246 reviews.[8] The statistics of our final dataset is shown in Table 2.

Total number of reviews	No. of pos reviews	No. of neg reviews	No. of classes	keyword categories	Vocab. size
201,246	133,306	67,940	2	58	476K

Table 2: The statistics of the collected data.

5.2. *Sentiment Classification Results*

For sentiment classification, we compare our approach with a number of baselines as discussed below. Whenever the original implementation is available, we provide the link of the source code in the footnote. Methods that were re-implemented by us are marked with †.

- CNN[9]: A strong baseline and has been widely used in sentiment classification [37]. In our experiments, we use one-dimensional CNN kernels with

[7] All the keywords are listed in Appendix A.

[8] We also evaluate the sentiment classification performance on the whole dataset by merging reviews with the rating scores of 2 to 4 as "Others". Details can be found in Appendix B.

[9] `https://github.com/yoonkim/CNN_sentence`

the kernel size set to 2, 3 and 4, and the number of kernels of each size set to 50. That is, the dimension of CNN features is 150.

- RNN†: RNN can be used to model text sequences [19]. In our experiments, we set the size of hidden states to 150.

- RCNN[10]: A hierarchical neural network stacked with CNN and RNN layers for sentiment classification [20]. The CNN layer uses 50 one-dimension convolutional kernels with sizes of 2, 3 and 4 each, while the RNN layer has 150 hidden states, which is the same as the size of CNN features.

- LSTM†: LSTM was previously used for sentiment classification [38]. For each document, words are fed into LSTM sequentially to obtain the hidden representations which are composed by mean pooling to derive the document representation. We also report the results with two variants of LSTM, LSTM with an attention mechanism (LSTM+Att) and LSTM with a self-attention mechanism (LSTM+SAtt). The implementation of LSTM+Att is modified from [22] by removing the hierarchical structure, and LSTM+SAtt is implemented by using the self-attention module [39] instead of the original attention block from LSTM+Att.

- HAN[11]: A neural sentiment classification method [22] with hierarchical RNNs and attentions mechanism [21]. The implementation we used is from [22] and we follow the default parameter setting in the source code.

- BERT†: Bidirectional Encoder Representations from Transformers [40]. We feed each review document as a long sentence with sentences separated by the `[SEP]` token into BERT, which is fine-tuned on our data. We truncate documents with length over 512 tokens and use the representation of `[CLS]` token for classification.

- JASE: Our proposed weakly-supervised joint sentiment classification and aspect extraction method.

We perform 5-fold cross validation and report the averaged sentiment classification results in Table 3. Among the baselines, RNN performs significantly worse than the others. LSTM and CNN outperform RNN. Adding the attention mechanism, LSTM+Att only gives a marginal improvement compared to LSTM. Models built on a hierarchical structure such as RCNN, HAN and our proposed JASE, i.e., deriving sentence representations first from word sequences and then generating document representations from sentence sequences before feeding them into a softmax layer for sentiment classification, clearly outperform other non-hierarchical models, including BERT. JASE achieves superior performance compared to all baseline models.

[10] `http://ir.hit.edu.cn/~dytang/paper/emnlp2015/codes.zip`
[11] `https://github.com/thunlp/NSC`

Method	Hierarchical	Transformer	Attention	Accuracy
CNN	No	No	No	0.941
RNN	No	No	No	0.850
RCNN	Yes	No	No	0.974
LSTM	No	No	No	0.971
LSTM+Att	No	No	Yes	0.973
LSTM+SAtt	No	Yes	Yes	0.967
HAN	Yes	No	Yes	0.977
BERT	No	Yes	Yes	0.953
JASE	Yes	Yes	Yes	**0.982**

Table 3: Sentiment classification results on the Yelp review dataset.

As we have imbalanced data with the number of positive reviews almost doubling that of negative reviews, we also report the per-class sentiment classification results in Table 4. Models which give better performance on the positive class are CNN, RNN, LSTM and BERT. The combination of RNN and CNN (RCNN) or the incorporation of attentions (LSTM+Att and LSTM+SAtt) achieve balanced results on both positive and negative classes. With the hierarchical representation learning structure and the incorporation of attention mechanisms, both HAN and JASE also achieve balanced results on both classes with marginally better F-measure on the negative class. Overall, JASE gives superior performance on both polarity classes in comparison to all baselines, showing the effectiveness of our proposed architecture.

Method	Positive			Negative			Macro-F
	P	R	F	P	R	F	
CNN	0.947	0.948	0.948	0.929	0.927	0.928	0.938
RNN	0.858	0.861	0.859	0.834	0.829	0.831	0.845
RCNN	0.971	0.975	0.973	0.980	0.971	0.976	0.974
LSTM	0.974	0.975	0.975	0.965	0.963	0.964	0.969
LSTM+Att	0.972	0.974	0.973	0.974	0.971	0.973	0.973
LSTM+SAtt	0.967	0.971	0.969	0.967	0.959	0.963	0.966
HAN	0.973	0.977	0.975	0.985	0.977	0.981	0.978
BERT	0.954	0.957	0.955	0.951	0.945	0.948	0.952
JASE	**0.980**	**0.978**	**0.979**	**0.987**	**0.991**	**0.989**	**0.984**

Table 4: The precision, recall, and macro F-measure of per-class sentiment classification results.

5.3. Aspect Extraction Results

The *sentence-level aspect representation learning module* in the JASE framework generates latent aspect vectors which allows us to extract top associated words for each latent aspect dimension by the weights connecting between the latent aspect vector with the reconstruction layer, as described in Section 4. We can interpret the top associated words for each latent aspect dimension as aspect topic words.

In this subsection, we evaluate aspect extraction results based on two measures: coherence calculated on the extracted aspect-associated words and aspect extraction results evaluated on a small dataset consisting of 300 sentences relating to three aspects, '*Price*', '*Service*', and '*Treatment*', with 100 sentences for each aspect. First, we report the coherence results using four different topic coherence measures, including the normalised Pointwise Mutual Information (NPMI) [41], a lexicon-based method (UCI) [42], and context-vector-based coherence measures (CV and CA) [43]. We compare the results with the following topic models:

- JST[12]: Joint Sentiment-Topic Model [44, 45], a Bayesian topic model which jointly models sentiments and topics. The model is initialised with the word prior polarity derived from the MPQA sentiment lexicon [46].
- NVDM[13]: Neural Variational Document model [47] which is based on Variational Autoencoder (VAE) [48] for topic extraction from text.
- NGTM†: Neural Generative Topic model [49], also based on VAE, but additionally incorporating the log background frequency of words to better deal with sparsity.
- Scholar: A neural topic model with metadata [50] and initialised with GloVe word embeddings [35].

Method	CA	NPMI	UCI	CV
JST [44]	0.131	**-0.039**	-1.495	0.377
NVDM [47]	0.138	-0.093	-2.775	0.408
NGTM [49]	0.136	-0.097	-2.894	0.425
Scholar [50]	0.153	-0.061	-2.188	0.424
Jase	**0.194**	**-0.039**	**-1.385**	**0.445**

Table 5: Topic coherence results of different models (the higher the better).

It can be observed from Table 5 that our proposed Jase clearly outperforms all the other topics models, including the traditional LDA-based model such as JST and the neural topic models built on VAE such as NVDM, NGTM and Scholar. One possible reason is that Jase is able to account for the full vocabulary while other topics models require pre-processing to filter out stopwords

[12] `https://github.com/linron84/JST`

[13] `https://github.com/ysmiao/nvdm`

and low-frequency words both for training efficiency and for generating better topic results. We argue that some low-frequency words may be semantically important for certain aspects and hence cannot be simply filtered out based on word occurrence frequencies. As opposed to other topic models, our proposed JASE is able to deal with the full vocabulary and identify aspect-associated words based on the aspect-driven attention weights.

To evaluate the aspect extraction result, we manually annotated 300 sentences from the collected Yelp data. For each sentence, we followed the annotation scheme from the previous research [51, 33] and selected sentences which only contain exactly one aspect. To assign an aspect label to a sentence in the test set, we compare the learned sentence-level aspect representation with the word embedding of each of the three target aspect labels, '*Price*', '*Service*', and '*Treatment*', and choose the nearest aspect label. For the topic results generated by baselines such as NVDM, NGTM, and SCHOLAR, we perform aspect label assignment in the same way. In addition, we also compare our method with two unsupervised aspect extraction methods:

- ABAE[14]: an attention-based aspect extraction model which uses an autoencoder to map an input sentence to latent aspects and subsequently reconstructs the input sentence from the latent aspects [33].
- CAT[15]: A contrastive-attention-based neural aspect extraction model, which use a single-head attention mechanism built on the Radial Basis Function (RBF) kernel to calculate the attention weight of each input word. This unsupervised method only requires word embeddings and a POS tagger to apply to new domains and languages [51].

Method	Precision	Recall	F-measure
NVDM	0.863	0.863	0.863
NGTM	0.871	0.870	0.870
SCHOLAR	0.874	0.873	0.873
ABAE	0.880	0.880	0.879
CAT	0.886	0.880	0.882
JASE	**0.901**	**0.900**	**0.900**

Table 6: Overall performance comparison of aspect extraction results on three aspects, *Price*', '*Service*', and '*Treatment*'.

It can be observed from Table 6 that neural topic models built on VAE perform relatively worse compared to unsupervised aspect extraction models such as ABAE and CAT. Our proposed JASE further improves upon ABAE and CAT by nearly 2% in F-measure. For more detailed comparison, we shown the

[14] `https://github.com/ruidan/Unsupervised-Aspect-Extraction`

[15] `https://github.com/clips/cat`

precision, recall and F-measure on each of the three aspects in Table 7. Jase outperforms baselines on the aspects of '*Service*' and '*Treatment*', but performs slightly worse than ABAE and CAT on the aspect of '*Price*'. One possible reason is that there are abundant instances relating to the aspect of '*Price*' in the training data. Hence, models are well trained on this aspect, as evidenced by the better performance achieved by all the models in comparison to the other two aspects. Nevertheless, as opposed to existing word-level models, our proposed Jase is able to weigh different aspects by the sentence-level attentions (§4) and achieves better sentiment classification at the document level.

Method	Price			Service			Treatment		
	P	R	F	P	R	F	P	R	F
NVDM	0.875	0.910	0.892	0.849	0.790	0.819	0.864	0.890	0.877
NGTM	0.873	0.890	0.881	0.890	0.810	0.848	0.850	**0.910**	0.879
Scholar	0.875	0.910	0.892	0.891	0.820	0.854	0.856	0.890	0.873
ABAE	0.874	**0.970**	**0.919**	0.868	0.790	0.827	0.898	0.880	0.889
CAT	**0.881**	0.960	**0.919**	0.853	0.810	0.831	0.915	0.860	0.887
JASE	0.879	0.940	0.908	**0.908**	**0.890**	**0.899**	**0.916**	0.870	**0.892**

Table 7: Aspect extraction results for each of the three aspects.

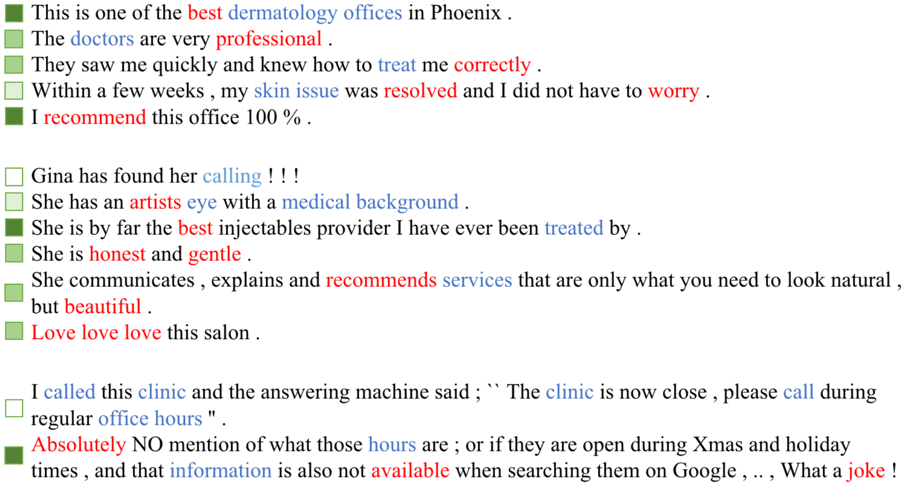
Polarity	**Score**	**Aspect**	**Aspect words**
Positive	0.679	Environment	comfortable conveniently situated cleanest comfortably
	0.632	Service	truly elegant caring atmosphere refreshing
	0.378	Cost	discounts free purchase deposit insurance
	0.351	Appreciation	thank thanks appreciate enjoy thx
	0.319	Waiting time	don't waited waiting didn't wait
Negative	0.439	Complaint	**** blood offensive dirty ethic
	0.335	Attitude	arrogant stale racist hateful insulting
	0.229	Equipment	accessories classroom models palate quality
	0.220	Service	disgusted unprofessional disappoint disappointed nack
	0.210	Cancer	carcinoma melanoma tumors chemotherapy prognosis

Table 8: Aspect extraction results under different sentiment categories. For each aspect, the score value is obtained by averaging the attention weights from the transformer which are associated with a specific polarity category. We select the top 5 aspects for each polarity category. The aspect labels listed under the "Aspect" column are assigned manually for easy inspection. We mask swear words by '*'.

In Table 8, we list some example aspects under different polarity categories. The score is obtained using the method described in Section 4. A higher score indicates that the corresponding aspect is associated with its related polarity category more closely. It can be observed that patients are positive towards the environment and services they received. They also appreciate free or discounted services and shorter waiting time. On the contrary, we can find that patients complain the staff attitude, equipment and also express negative feelings towards cancer-related issues. These results show that our proposed framework can indeed extract aspects discussed under different polarity categories despite using no aspect annotations for training.

5.4. Visualisation of Attention Weights

In Figure 5, we present the visualisation of both the word-level and the sentence-level attention weights generated by our JASE model for three patient reviews. For each sentence, we highlight the words with the highest sentiment-associated attention weights with the red colour, and those with the highest aspect-associated attention weights with the blue colour. It can be observed that most red-coloured words are indeed polarity-bearing words such as '*best*', '*honest*' and '*joke*', while most blue-coloured words are aspect-associated such as '*dermatology offices*' and '*services*'. Figure 5 also shows the sentence-level attention weights with green-shaded boxes. It is clear that the sentence with the highest weight is a good summary in their respective review (such as Sentence 3 in the second review and Sentence 2 in the third review).



🟩 This is one of the best dermatology offices in Phoenix .
▨ The doctors are very professional .
▨ They saw me quickly and knew how to treat me correctly .
░ Within a few weeks , my skin issue was resolved and I did not have to worry .
🟩 I recommend this office 100 % .

□ Gina has found her calling ! ! !
░ She has an artists eye with a medical background .
🟩 She is by far the best injectables provider I have ever been treated by .
▨ She is honest and gentle .
▨ She communicates , explains and recommends services that are only what you need to look natural , but beautiful .
▨ Love love love this salon .

□ I called this clinic and the answering machine said ; `` The clinic is now close , please call during regular office hours " .
🟩 Absolutely NO mention of what those hours are ; or if they are open during Xmas and holiday times , and that information is also not available when searching them on Google , .. , What a joke !

Figure 5: Visualisation of attention weights of three example reviews. Words in red are those with higher sentiment-associated attention weights, while words in blue are those with high aspect-associated attention weights. Each sentence is also preceded with a green-shaded box with varying intensities indicating different sentence-level attention weights.

5.5. Visualisation of Intermediate Training Results

To further understand if JASE is able to learn good aspect representations that separate words with different sentiment categories and aspects, we present in Figure 6 the training loss curves and the visualisation of aspect word vectors learned by Jase at various training stages. We select top 10 words from top 10 aspects based on polarity score. For each word, we derive the sentiment label from its polarity score obtained by our aspect modelling module. We use the blue colour for positive words and the red colour for negative words. In addition, we use different marker shapes to denote the words under different aspects. It can be observed that at the initial phase of the training, aspect words with different polarities and aspects are mixed. However, after training for 30,000 mini-batches, we see a better separation pattern. When training converges at 60,000 mini-batches, aspect words with different polarities are well separated in the learned embedding space. We can also observe that the aspect words under the same aspect category tend to be grouped together. This shows the effectiveness of our proposed Jase in learning words representations which capture both sentiment and aspect information well.

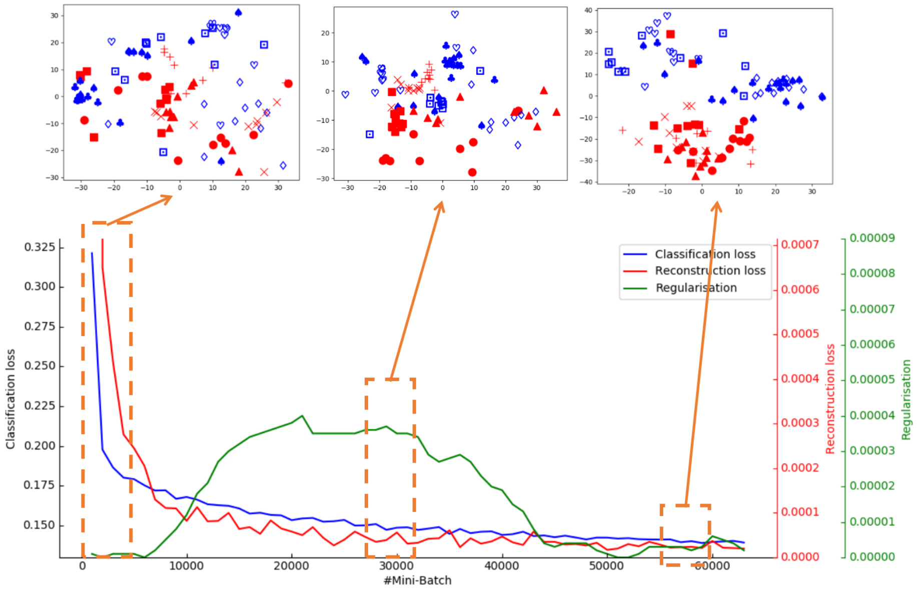


Figure 6: Visualisation of the classification loss, the aspect representation reconstruction loss, the regularisation and document vectors at various training stages.

To assess the risk of overfitting, we plot in Figure 7 the classification loss curves at different stages of the model learning process on the training, the validation, and the testing set, respectively. It can be observed that during the whole training process, the training loss keeps decreasing almost monotonously. However, after training for 25,000 epochs, the validation and testing losses are stabilised at the same level. Note that the horizontal axis is the same as that in Figure 6. This shows a very low risk of overfitting for the sentiment classification task and at the same time the high-quality separation pattern of aspect words can be obtained as shown in Figure 6.

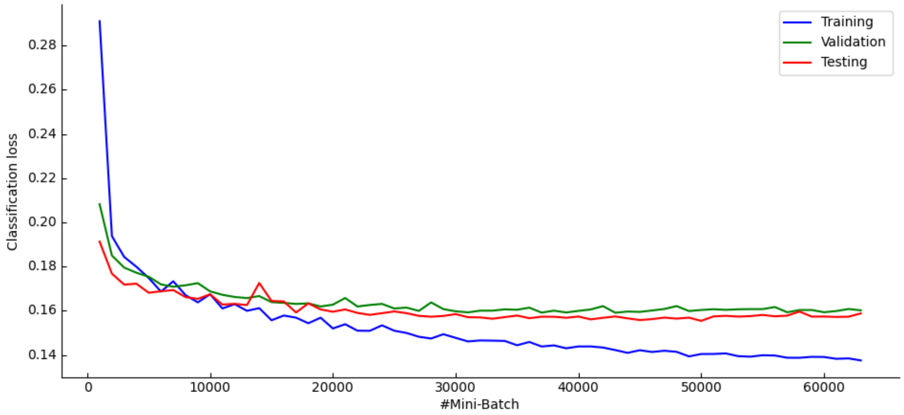


Figure 7: Visualisation of the classification loss on training, validation, and testing set at various training stages.

6. Conclusion

In this paper, we have presented Jase, a new neural hierarchical model for joint sentiment classification and aspect extraction. Existing aspect-based sentiment analysis methods require aspect-level annotations for model learning. In real-world scenarios, however, such fine-grained annotations are expensive to obtain. Also, supervised learning approaches cannot handle unseen aspects effectively. Furthermore, most existing ABSA approaches focus on sentence-level tasks due to the cost of annotation. The latent correlation between aspects and the hierarchical structure of text are under-explored in existing studies. Our proposed weakly supervised model is able to extract coherent aspect topics without requiring aspect-level annotations for training. The only supervision information required is document-level sentiment labels. In addition, Jase can learn good document representations which capture both sentiment and aspect information in the embedding space. Experimental results on the Yelp patient reviews show the superior performance of our proposed model over existing neural models for sentiment classification. Jase also extracts semantically more coherent aspect topics compared to both LDA-based and VAE-based topic models. In future work, we plan to explore the interpretabilities offered by Jase on other review data.

Funding

This work was funded by EU-H2020 (grant no. 794196) and the UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1, EP/T017112/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1).

References

[1] S. of State for Health, High Quality Care for All: NHS Next Stage Review Final Report, Vol. 7432, The Stationery Office, 2008.

[2] M. Sarrouti, S. O. E. Alaoui, Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions, Artif. Intell. Medicine 102 (2020) 101767.

[3] G. G. Gao, J. S. McCullough, R. Agarwal, A. K. Jha, A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period, Journal of medical Internet research 14 (1) (2012) e38.

[4] F. Greaves, C. Millett, Consistently increasing numbers of online ratings of healthcare in england, J Med Internet Res 14 (3) (2012) e94.

[5] F. Alemi, M. Torii, L. Clementz, D. C. Aron, Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments, Quality Management in Healthcare 21 (1) (2012) 9–19.

[6] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, Semeval-2014 task 4: Aspect based sentiment analysis, in: P. Nakov, T. Zesch (Eds.), Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014, The Association for Computer Linguistics, 2014, pp. 27–35.

[7] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, Semeval-2015 task 12: Aspect based sentiment analysis, in: D. M. Cer, D. Jurgens, P. Nakov, T. Zesch (Eds.), Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015, The Association for Computer Linguistics, 2015, pp. 486–495.

[8] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. D. Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. V. Loukachevitch, E. V. Kotelnikov, N. Bel, S. M. J. Zafra, G. Eryigit, Semeval-2016 task 5: Aspect based sentiment analysis, in: S. Bethard, D. M. Cer, M. Carpuat, D. Jurgens, P. Nakov, T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016, The Association for Computer Linguistics, 2016, pp. 19–30.

[9] A. DeCourcy, E. West, D. Barron, The national adult inpatient survey conducted in the english national health service from 2002 to 2009: how have the data been used and what do we know as a result?, BMC Health Services Research 12 (1) (2012) 71. `doi:10.1186/1472-6963-12-71`.

[10] A. Coulter, L. Locock, S. Ziebland, J. Calabrese, Collecting data on patient experience is not enough: they must be used to improve care, BMJ 348 (2014).

[11] S. M. J. Zafra, M. T. M. Valdivia, M. D. Molina-González, L. A. U. López, How do we talk about doctors and drugs? sentiment analysis in forums expressing opinions for medical domain, Artif. Intell. Medicine 93 (2019) 50–57.

[12] R. Fitzpatrick, C. Graham, E. Gibbons, J. King, K. Flott, C. Jenkinson, Development of new models for collection and use of patient experience information in the nhs–prp 070/0074 (2014).

[13] E. J. Gibbons, C. Graham, J. King, K. Flott, et al., Developing approaches to the collection and use of evidence of patient experience below the level of national surveys, Patient Experience Journal 3 (1) (2016) 92–100.

[14] I. D. Maramba, A. Davey, M. N. Elliott, M. Roberts, M. Roland, F. Brown, J. Burt, O. Boiko, J. Campbell, Web-based textual analysis of free-text patient experience comments from a survey in primary care, JMIR medical informatics 3 (2) (2015) e20.

[15] M. Rastegar-Mojarad, Z. Ye, D. Wall, N. Murali, S. Lin, Collecting and analyzing patient experiences of health care from social media, JMIR research protocols 4 (3) (2015) e78.

[16] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, L. Donaldson, Use of sentiment analysis for capturing patient experience from free-text comments posted online, Journal of medical Internet research 15 (11) (2013) e239.

[17] J. B. Hawkins, J. S. Brownstein, G. Tuli, T. Runels, K. Broecker, E. O. Nsoesie, D. J. McIver, R. Rozenblum, A. Wright, F. T. Bourgeois, et al., Measuring patient-perceived quality of care in us hospitals using twitter, BMJ quality & safety 25 (6) (2016) 404–413.

[18] Y. Kim, Convolutional neural networks for sentence classification, in: The 2014 Conference on Empirical Methods on Natural Language Processing, Association for Computational Linguistics, 2014, pp. 1746–1751.

[19] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850 (2013).

[20] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: The 2015 Conference on Empirical Methods on Natural Language Processing., 2015, pp. 1422–1432.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, E. H. Hovy, Hierarchical attention networks for document classification., in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[22] H. Chen, M. Sun, C. Tu, Y. Lin, Z. Liu, Neural sentiment classification with user and product attention, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 1650–1659.

[23] D. Tang, B. Qin, T. Liu, Aspect level sentiment classification with deep memory network, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 214–224.

[24] R. He, W. S. Lee, H. T. Ng, D. Dahlmeier, Effective attention modeling for aspect-level sentiment classification, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING, 2018, pp. 1121–1131.

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[26] C. Fan, Q. Gao, J. Du, L. Gui, R. Xu, K. Wong, Convolution-based memory network for aspect-based sentiment analysis, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, 2018, pp. 1161–1164.

[27] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, 2017, pp. 452–461.

[28] Y. Zou, T. Gui, Q. Zhang, X. Huang, A lexicon-based supervised attention model for neural sentiment analysis, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, 2018, pp. 868–877.

[29] J. Li, H. Yang, C. Zong, Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, 2018, pp. 925–936.

[30] H. Xu, B. Liu, L. Shu, P. S. Yu, Double embeddings and cnn-based sequence labeling for aspect extraction, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, 2018, pp. 592–598.

[31] W. Wang, S. J. Pan, Transition-based adversarial network for cross-lingual aspect extraction, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, 2018, pp. 4475–4481.

[32] X. Li, L. Bing, P. Li, W. Lam, Z. Yang, Aspect term extraction with history attention and selective transformation, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, 2018, pp. 4194–4200.

[33] R. He, An unsupervised neural attention model for aspect extraction, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, 2017, pp. 388–397.

[34] G. Singh Chauhan, Y. Kumar Meena, D. Gopalani, R. Nahta, A two-step hybrid unsupervised model with attention mechanism for aspect extraction, Expert Systems with Applications 161 (2020) 113673.

[35] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543.

[36] A. M. E. Saleh, M. Arashi, B. G. Kibria, Theory of ridge regression estimation with applications, Vol. 285, John Wiley & Sons, 2019.

[37] L. Gui, R. Xu, Y. He, Q. Lu, Z. Wei, Intersubjectivity and sentiment: From language to knowledge, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, 2016, pp. 2789–2795.

[38] X. Wang, Y. Liu, C. Sun, B. Wang, X. Wang, Predicting polarities of tweets by composing word embeddings with long short-term memory, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, 2015, pp. 1343–1353.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural

Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[40] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[41] N. Aletras, M. Stevenson, Evaluating topic coherence using distributional semantics, in: Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany, 2013, pp. 13–22.

[42] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA, 2010, pp. 100–108.

[43] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: The 8th ACM International Conference on Web Search and Data Mining, 2015, pp. 399–408.

[44] C. Lin, Y. He, Joint sentiment/topic model for sentiment analysis, in: The ACM Conference on Information and Knowledge Management, 2009, pp. 375–384.

[45] C. Lin, Y. He, R. Everson, S. Ruger, Weakly supervised joint sentiment-topic detection from text, IEEE Transactions on Knowledge and Data engineering 24 (6) (2012) 1134–1145.

[46] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada, 2005, pp. 347–354.

[47] Y. Miao, L. Yu, P. Blunsom, Neural variational inference for text processing, in: The International Conference on Machine Learning, 2016, pp. 1727–1736.

[48] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014.

[49] D. Card, C. Tan, N. A. Smith, A neural framework for generalized topic models, arXiv preprint arXiv:1705.09296 (2017).

[50] D. Card, C. Tan, N. A. Smith, Neural models for documents with metadata, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, 2018, pp. 2031–2040.

[51] S. Tulkens, A. van Cranenburgh, Embarrassingly simple unsupervised aspect extraction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, 2020, pp. 3182–3187.

Appendix A

List of Keywords
Walk-in Clinics, Surgeons, Oncologist, Cardiologists, Hospitals, Internal Medicine, Assisted Living Facilities, Cannabis Dispensaries, Doctors, Home Health Care, Health Coach, Emergency Pet Hospital, Pharmacy, Sleep Specialists, Professional Services, Addiction Medicine, Weight Loss Centers, Pediatric Dentists, Cosmetic Surgeons, Nephrologists, NaturopathicHolistic, Pediatricians, Nurse Practitioner, Urgent Care, Orthopedists, Drugstores, Optometrists, Rehabilitation Center, HypnosisHypnotherapy, Physical Therapy, Neurologist, Memory Care, Allergists, Counseling & Mental Health, Pet Groomers, Podiatrists, Dermatologists, Diagnostic Services, Radiologists, Medical Centers, Gastroenterologist, Obstetricians & Gynecologists, Pulmonologist, Ear Nose & Throat, Ophthalmologists, Sports Medicine, Nutritionists, Psychiatrists, Vascular Medicine, Cannabis Clinics, Hospice, First Aid Classes, Medical Spas, Spine Surgeons, Health Retreats, Medical Transportation, Dentists, Health & Medical, Speech Therapists, Emergency Medicine, Chiropractors, Medical Supplies, General Dentistry, Occupational Therapy, Urologists

Table A1: The list of keywords used for retrieving patient reviews from Yelp.

Appendix B

Table A2 shows the sentiment classification results on the full Yelp dataset with the reviews receiving the rating scores of 2-4 grouped under the “Others” category. Compared with the binary classification result shown in Table 4, all models perform worse in both the “Positive” and the “Negative” categories. The results in the category of “Others” are very low. This is due to the highly imbalanced class distributions since the number of reviews in the “Others” category is significantly less compared to the other two categories. Nevertheless, our proposed method still beats the baselines on the Marco-F measure.

Method	Positive			Negative			Others			Macro-F
	P	R	F	P	R	F	P	R	F	
CNN	0.906	0.927	0.917	0.872	0.983	0.924	0.471	0.259	0.335	0.725
RNN	0.871	0.882	0.876	0.824	0.921	0.870	0.384	0.228	0.286	0.677
RCNN	0.916	0.927	0.922	0.865	0.976	0.917	0.429	0.240	0.308	0.715
LSTM	0.897	0.922	0.909	0.862	0.975	0.915	0.444	0.235	0.307	0.710
LSTM+Att	0.897	0.917	0.907	0.863	0.974	0.915	0.470	0.255	0.331	0.718
LSTM+Satt	0.903	0.927	0.915	0.866	0.975	0.917	0.477	0.261	0.337	0.723
HAN	0.945	0.959	0.952	0.874	0.984	0.926	0.491	0.278	0.355	0.744
BERT	0.928	0.942	0.935	0.867	0.975	0.918	0.439	0.248	0.317	0.723
JASE	0.963	0.978	0.970	0.882	0.991	0.933	0.494	0.281	0.358	0.754

Table A2: The precision, recall, and macro F-measure of per-class sentiment classification results on the full Yelp dataset.