Document-Level Multi-Aspect Sentiment Classification for Online Reviews of Medical Experts

Tian Shi
Virginia Tech
tshi@vt.edu

Vineeth Rakesh
Interdigital
vineeth.mohan@interdigital.com

Suhang Wang
Pennsylvania State University
szw494@psu.edu

Chandan K. Reddy
Virginia Tech
reddy@cs.vt.edu

ABSTRACT

In the era of big data, online doctor review platforms, which enable patients to give feedback to their doctors, have become one of the most important components in healthcare systems. On one hand, they help patients to choose their doctors based on the experience of others. On the other hand, they help doctors to improve the quality of their service. Moreover, they provide important sources for us to discover common concerns of patients and existing problems in clinics, which potentially improve current healthcare systems. In this paper, we systematically investigate the dataset from one of such review platform, namely, ratemds.com, where each review for a doctor comes with an overall rating and ratings of four different aspects. A comprehensive statistical analysis is conducted first for reviews, ratings, and doctors. Then, we explore the content of reviews by extracting latent topics related to different aspects with unsupervised topic modeling techniques. As the core component of this paper, we propose a multi-task learning framework for the document-level multi-aspect sentiment classification. This task helps us to not only recover missing aspect-level ratings and detect inconsistent rating scores but also identify aspect-keywords for a given review based on ratings. The proposed model takes both features of doctors and aspect-keywords into consideration. Extensive experiments have been conducted on two subsets of ratemds dataset to demonstrate the effectiveness of the proposed model.

CCS CONCEPTS

• **Information systems → Sentiment analysis**; *Clustering and classification*; • **Applied computing → Health care information systems**; • **Computing methodologies →** *Topic modeling*.

KEYWORDS

Online reviews, sentiment classification, multi-aspect, multi-task learning, attention mechanism.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM ’19, November 3–7, 2019, Beijing, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357828

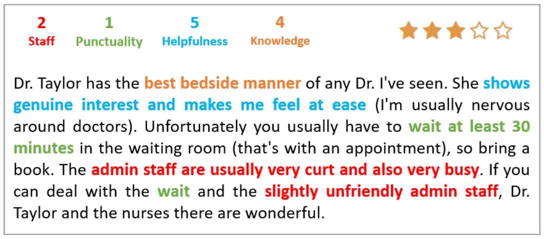
ACM Reference Format:
Tian Shi, Vineeth Rakesh, Suhang Wang, and Chandan K. Reddy. 2019. Document-Level Multi-Aspect Sentiment Classification for Online Reviews of Medical Experts. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM ’19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3357384.3357828

1 INTRODUCTION

Healthcare systems are evolving rapidly due to advancements in recent artificial intelligence techniques, especially deep learning frameworks [23, 30]. A number of automated tools and ML driven micro-services in healthcare, e.g., medical imaging diagnosis for diabetic eye disease [9] and cancer [18], have caught many attentions from both industry and academia. Online doctor review systems, such as ratemds[1] and zocdoc[2], establish a unique environment for patients to give feedback to their doctors on visiting experience. These reviews are evolving into an important source for evaluating performance of doctors in medical practices as a supplement to their professional knowledge. For example, ratemds is one of such review platforms for doctors and facilities (e.g., hospitals or clinics), which has more than 1.7 million healthcare providers (i.e., doctors and facilities) and 2.5 million reviews. On their website, a patient can anonymously post a review along with an overall rating and ratings from four different aspects i.e., staff, punctuality, helpfulness and knowledge, to their doctors. Similarly, patients can also review and rate facilities. Fig. 1 shows an example of doctor reviews. In this figure, there is a plain-text review with four aspect-level ratings, in which staff and punctuality refer to front-desks and appointments, respectively, while helpfulness and knowledge are about bedside manners of doctors and medical procedures. Generally speaking, these reviews sketch more detailed profiles of doctors in medical practices, so they can not only help other patients to find better options, but also help doctors to improve their service quality.

Nowadays, different knowledge discovery and opinion mining techniques allow us to find out general needs of patients and existing problems in clinics from large amount of online reviews, which helps to improve current healthcare systems. Many of these techniques, including graphical models [21], regression approaches [36] and deep learning methods [14, 41, 43], have been successfully applied to similar online review systems in other domains, such as BeerAdvocate[3] and TripAdvisor[4]. However, online doctor

[1] https://www.ratemds.com/
[2] https://www.zocdoc.com/
[3] https://www.beeradvocate.com/
[4] https://www.tripadvisor.com/



2 Staff 1 Punctuality 5 Helpfulness 4 Knowledge ★★★☆☆

Dr. Taylor has the best bedside manner of any Dr. I've seen. She shows genuine interest and makes me feel at ease (I'm usually nervous around doctors). Unfortunately you usually have to wait at least 30 minutes in the waiting room (that's with an appointment), so bring a book. The admin staff are usually very curt and also very busy. If you can deal with the wait and the slightly unfriendly admin staff, Dr. Taylor and the nurses there are wonderful.

Figure 1: An example of ratemds reviews. Keywords corresponding to different aspects are highlighted with different colors.

review systems, which are primary platforms for patients to give feedback, have not been sufficiently investigated before [3, 8, 11], especially for those systems which evaluate doctors' medical practices from different aspects. There are many tasks associated with this type of data. For example, many patients are less motivated to give aspect-level ratings and some ratings are inconsistent with reviews, can we predict rating scores based on plain-text reviews to recover missing values and correct inconsistent ratings? On the other hand, given aspect categories and aspect-level ratings, can we use these ratings as a form of weak supervision to obtain keywords corresponding to different categories? Alternately, can we use unsupervised methods to discover cluster structures in latent space for keywords in reviews and associate them with different aspects? Although sophisticated models have been proposed for these tasks, they have only been applied to other types of datasets [15] and some of them have only been tested on small-scale datasets [14, 41]. In this paper, we first thoroughly explore ratemds dataset, and then, due to the strong correlations of multi-aspect ratings, we formulate a multi-task learning model to predict ratings and detect aspect-keywords in each review with attention mechanism [1, 20]. Our contributions can be summarized as follows:

- Propose a multi-task learning framework, which takes features of doctors and aspect-keywords discovered by the topic model into consideration, for the document-level multi-aspect sentiment classification task and conduct extensive experiments on two subsets of ratemds dataset.
- Introduce a new dataset which consists of more than 2 million reviews with multi-aspect ratings. Different from datasets for commercial products and entertainment (like BeerAdvocate and TripAdvisor), this dataset is healthcare related and an important source for studying general concerns of patients and existing problems in clinics.
- Conduct a comprehensive statistical analysis on this dataset, including statistics of reviews, ratings and doctors. We also explore aspect-keywords of reviews with a topic model [2].

The rest of this paper is organized as follows: In Section 2, we introduce related work of document-level multi-aspect sentiment classification (DLMASC). In Section 3, we study statistics of ratemds dataset and explore aspect-keywords from plain-text reviews. In Section 4, we present details of our proposed multi-task learning framework for the DLMASC task. In Section 5, we introduce two subsets of ratemds dataset, baseline methods, implementation details and evaluation metrics, as well as analyze experimental results. This discussion concludes in Section 6.

2 RELATED WORK

In this section, we review related work of document-level multi-aspect sentiment classification. The sentiment analysis, also known as opinion mining [17], aims to determine the attitude of a person via analyzing polarity (e.g., positive, neutral, or negative) of given text [24, 34]. Document-level sentiment classification is a fundamental problem of sentiment analysis and opinion mining, which intends to determine the sentiment polarity of documents and online reviews. Many recent studies in this field are based on deep neural networks with hierarchical structures [4, 32, 40]. The document-level multi-aspect sentiment classification, which takes aspect categories and ratings into consideration, can be seen as an extension of document-level sentiment classification (single aspect). Early studies in this topic rely on feature engineering to extract features (e.g., *n*-gram features) corresponding to different aspects and use regression approaches (e.g., Support Vector Regression [31]) to predict multi-aspect ratings [19, 21, 36]. Recently, Yin et al. [41] proposed a multi-task learning framework where each aspect is viewed as a task. For each single task, a hierarchical attention module, which includes input encoders and iterative attention modules, has been used to encode documents for classification. This model requires pre-generated pseudo-questions to perform iterative attention and has only been tested on two small-scale datasets[5]. In [15], Li et al. proposed incorporating users' information, overall ratings and aspect keywords into their model, which is also based on a multi-task learning framework. However, it is not suitable for our problem, because, in ratemds dataset, reviews are written anonymously by patients due to privacy concerns. In other words, user information is not available. In addition, overall ratings are calculated by averaging aspect-level ratings, thus we cannot use overall ratings as the input. Another area, known as aspect-based sentiment classification [27, 28], is also related to our work. It consists of several fine-grained sentiment classification tasks, including aspect term extraction, aspect term polarity, aspect category detection, and aspect category polarity. There are many research works in this area [33, 35, 37]. For example, Tang et al. [33] introduced a deep memory network for aspect-level sentiment classification. These models usually focus on sentence-level sentiment classification. Moreover, aspect terms, categories, and entities in this problem need to be carefully annotated by human experts. Therefore, we will only work on document-level multi-aspect sentiment classification in this paper.

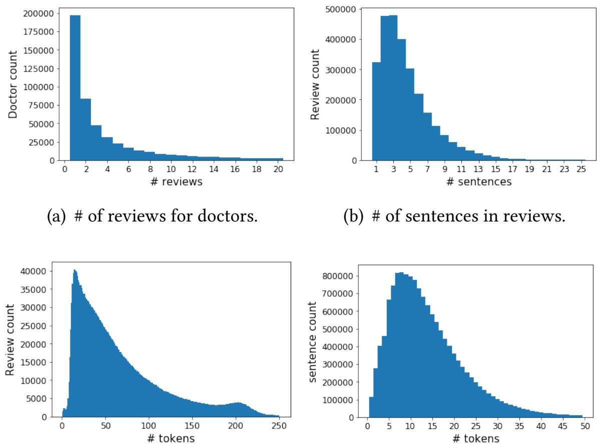
3 PRELIMINARY DATA ANALYSIS

In this section, we first conduct data analysis of reviews, ratings and doctors to get a comprehensive understanding of key features that can be useful for document-level multi-aspect sentiment classification. To gain deeper insights into content of reviews, we also use topic models to discover aspect-keywords from latent topics.

3.1 Overview

Ratemds dataset was obtained from ratemds.com website, which has records (e.g., specialties, insurance plans, etc.) of more than two million doctors world-wide, and over three million reviews along with numeric ratings of four aspects. The original ratemds dataset has many missing values for multi-aspect ratings which

[5] Both datasets only keep reviews with different aspect-level ratings [14].



(a) # of reviews for doctors.

(b) # of sentences in reviews.

(c) # of tokens in reviews.

(d) # of tokens in sentences.

Figure 2: Statistics of reviews in ratemds dataset.

reflects the fact that patients are less motivated to provide ratings from different aspects even if their comments are about multiple things. This problem shows the importance of the multi-aspect rating prediction/sentiment classification task (see Section 4). Due to the missing value problem, we first removed reviews with missing aspect-level ratings and eliminated records of doctors without reviews before investigating statistics of the dataset. Then, we obtain a refined ratemds dataset, in which distributions of doctors and reviews are shown in Fig. 2. In this dataset, there are more than 500K doctors and 2.7 million reviews, and the average number of reviews for each doctor is 4.6. From Fig. 2 (a), we observed that the distribution of doctors over review counts follows the power law distribution and almost 40% of doctors have only one review. Thus, it is difficult to apply collaborative filtering based methods to predict multi-aspect ratings.

Alternately, we can make use of textual reviews for the rating prediction task, which is the same as sentiment classification task in this paper. Therefore, we further studied the quality of textual reviews based on lengths of texts in order to make sure that they are not composed of short texts, since short texts may cause several problems in this task. First, short reviews cannot contain information of four aspects, which can probably confuse the classifier with respect to the aspect-keywords. Second, due to lack of semantic relationships [29, 39], it is difficult to use traditional knowledge discovery methods such as topic models [2] to automatically uncover the hidden thematic information from them. As a result, we cannot incorporate external knowledge discovered by these models into the sentiment classifiers for better classification performance. Fig. 2(b) and 2(c) show distribution of reviews over numbers of tokens and sentences, respectively. Fig. 2(d) is the distribution of sentences over the number of tokens. From these figures, we observed that most reviews have at least 2 sentences and over 12 tokens, and most of sentences have more than 10 tokens, which indicate that reviews in this dataset are not dominated by short texts. Moreover, the average length of reviews are more than 4 sentences and 72 tokens, which implies that there are a number of reviews whose content covers all four aspects in this dataset.

3.1.1 *Ratings.*

Each review comes with an overall rating and ratings for four different aspects, i.e., staff, punctuality, helpfulness and knowledge. The overall rating is the average of aspect-level ratings, which are integer numbers ranging from 1 to 5, where 1 and 5 represent extremely unsatisfied and satisfied, respectively. We show the distribution of reviews over rating scores in Fig. 3. From this figure, we observed that more than 60% of reviews have all aspect rating scores 5, which indicates that most patients are satisfied with their visits. About 17% of them are 1. It seems that patients with negative experience with their doctors tend to give extremely unsatisfied score to express their sentiment, especially when their doctors are not helpful. Many patients are slightly unsatisfied with staff and punctuality even if they are satisfied with their doctors, which may be because of their appointments and waiting time.

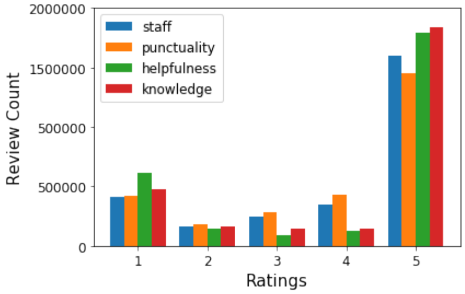


Figure 3: Statistics of reviews over aspect-level ratings.

3.1.2 *Doctors.*

Apart from the basic statistics of reviews and ratings, it is also important to investigate demographic features of doctors, since they may affect the visit experience of patients. For example, doctors who work in hospitals located in urban cities may receive lower punctuality scores in general than those who work in suburban clinics. Each doctor has a certain specialty (e.g., dentist). In Fig. 4, we show the average ratings for doctors with different specialties. It can be seen that dentists have much higher rating scores than other types of doctors. General practitioners and family practitioners (family-gp) have lower punctuality scores than others, due to the fact that patients with nearly any issue can visit them and get referrals when they have complicated health issues. Therefore, incorporating these demographic features into sentiment prediction models may increase the accuracy of results. In ratemds dataset, the key features of doctors includes gender, facility categories, specialties, locations and insurance plans.

We first observed that doctors in this dataset are from six different countries, i.e., United States (US), Canada (CA), Australia, India, United Kingdom and South Africa, where around 77% and 19% of them locate in the US and CA, respectively. There are three categories of facilities, i.e., hospital, clinic and urgent-care. About 90% of doctors work in clinics, 10% of them are in hospitals, and very few are in urgent-care. Many doctors work in different facilities and some of them work in two different countries. In this paper, we remove those doctors who work in more than one country, regarding health-care systems in different countries are different. For the feature gender, we observed that around 32% of doctors are female. For the feature specialty, which has been briefly mentioned in the beginning of this section, each doctor is assigned with one specialty and there are totally 57 different specialties. Almost 20% of doctors are family-gp. Dentists and obstetrician-gynecologists get relatively more reviews than doctors with other types of specialties.

Table 1: Aspect-keywords extracted with the topic model.

Specialty	Aspect	Keyword Examples
family-gp	staff	staff, office, rude, nurse, service, charge, call, visit, contact, insurance, follow, phone.
	punctuality	wait, hour, long, time, late, appointment, minute.
	helpfulness	care, see, listen, regard, consider, refer, show, understanding.
	knowledge	lab, symptom, treatment, professional, medicine, knowledge, drug, skill, prescription, diagnosis.
dentist	staff	insurance, charge, service, receive, nice, kind, smile, front-desk, polite, sweet, respect, assistant, staff.
	punctuality	rush, drive, late, time, appointment, wait, day, long.
	helpfulness	help, make, feel, comfortable, ease, care, ask, follow.
	knowledge	knowledgeable, procedure, explain, treatment, implant, review, replace, perform, extraction, experience, professional, tooth.
gynecologist-obgyn	staff	call, tell, ask, nurse, rude, staff, office, nice, friendly, service.
	punctuality	time, wait, appointment, hour, long, minute, day, week, rush.
	helpfulness	care, concern, understanding, warm, ease, helpful, think, save, offer, answer, consider, refuse, suggest.
	knowledge	knowledgeable, test, exam, review, explain, complication, pregnancy, deliver, experience, baby, surgery, pain, hysterectomy, surgeon, medication, bleed, cry, fibroid, treatment, diagnosis, scar.

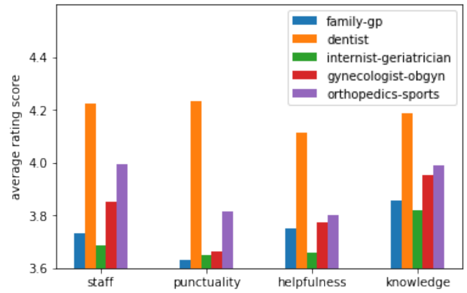


Figure 4: Ratings for doctors with different specialties.

3.2 Discover Aspect-Keywords

We further investigated reviews by extracting aspect-keywords using topic models [2]. Topic modeling approaches were considered because they can automatically uncover thematic information from a corpus in an unsupervised manner. In addition, keywords in each topic usually have strong semantic correlations and well-defined cluster structures. In ratemds dataset, reviews are assumed to be written from different aspects (different topics), whose keywords are expected to be less correlated.

3.2.1 Datasets. We first separated ratemds dataset based on countries and chose reviews for doctors in the US. It was followed by dividing selected reviews into sub-categories according to specialties. Then, we tokenized all reviews with SpaCy[6] package and removed stop-words, punctuation and rare words. Among all 57 specialties, we chose only three of them, i.e., family-gp, dentist, and gynecologist-obgyn, to illustrate our experiments and results.

3.2.2 Experiments and Results. For each dataset, we run Latent Dirichlet Allocation (LDA) [2, 10] model using package gensim[7]. The number of topics was set as 10 considering the fact that topics which are different from the four aspects may also be discovered. For each topic, we extracted top-20 keywords based on their weights. Finally, we empirically assigned these keywords to different aspects which has been shown in Table 1.

It can be seen from the table that *staff* usually represents *front-desk* or *nurse*. Their duties include receptions, contacting patients, managing insurance plans and bills, and so on. *Punctuality* is associated with *appointment* and *waiting time* in offices. From Fig. 3, we have found that fewer patients are satisfied with punctuality. This may be explained as it is hard to make appointment, waiting time is too long, or doctors rush to see other patients. *Helpfulness* can be understood as bedside manner of doctors. For example, a good doctor can carefully listen to complaints of patients, answer their questions and make them feel comfortable. Finally, *knowledge* in general is related with *diagnosis*, *exam*, *treatment*, and so on. From the table, we also observed that keywords of *staff*, *punctuality* or *helpfulness* are similar to each other for doctors with different specialties. However, since they are experts in different fields, the *knowledge* for different specialties has different keywords. For example, *surgery*, *hysterectomy*, *fibroid* and *pregnancy* are related with doctors specialized in gynecologist-obgyn.

4 DOCUMENT LEVEL MULTI-ASPECT SENTIMENT CLASSIFICATION

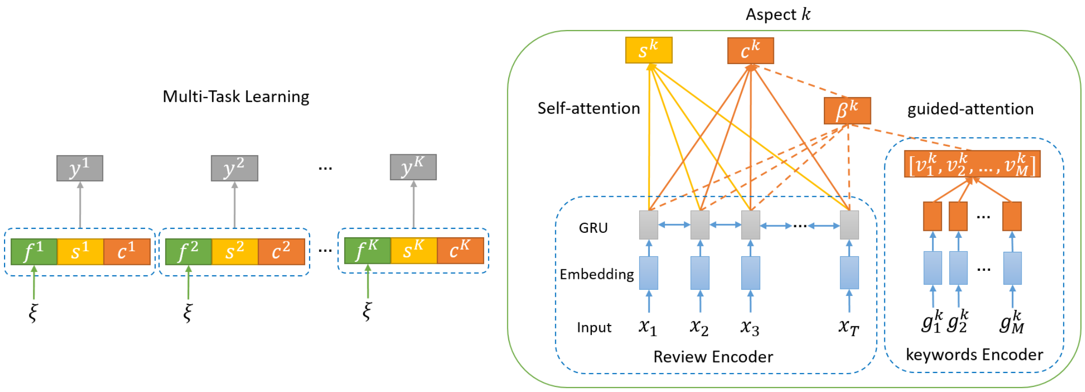
In Section 3, we had a comprehensive understanding of statistics and key features of ratemds dataset, and also extracted aspect-keywords with the topic model. In this section, we perform document-level multi-aspect sentiment classifications for reviews in ratemds dataset.

4.1 Preliminaries

In this problem, multi-aspect rating predictions can be viewed as tasks. Due to the strong correlations between different tasks, this problem can be naturally formulated as a multi-task learning problem. Hence, we propose a multi-task deep learning framework which takes plain-text reviews, aspect-keywords from topic models and features of doctors into consideration. Formally, this document-level multi-aspect sentiment classification problem can be described as follows: Given a textual review $X = (x_1, x_2, ..., x_T)$, keywords associated with different aspects $G = (G^1, G^2, ..., G^K)$ and a set of features ξ, our goal is to predict class labels, i.e., integer ratings, $y = (y^1, y^2, ..., y^K)$, where T and K are the number of tokens in the review and the number of aspects, respectively. x_t represents the one-hot encoding of word t. $G^K = (g_1^k, g_2^k, ..., g_M^k)$ is a list of keywords of aspect k, where g_m^k is the one-hot encoding of keyword m. y^k is an one-hot vector of the class label of aspect k. Specific to ratemds dataset, there are four aspects, so $K = 4$, and each aspect has 5 classes corresponding to rating scores from 1 to 5. The proposed framework (see Fig. 5) has a *review encoder* to encode textual reviews, a *multi-aspect self-attention* layer to selectively focus on parts of the review for a given aspect, an *aspect-keywords guided-attention* layer to focus on parts of the review that are related to aspect-keywords, and an *aspect-specific feature encoder* to incorporate features of doctors into the sentiment classification.

[6] https://spacy.io/

[7] https://radimrehurek.com/gensim/



(a) (b)

Figure 5: An illustration of the model architecture. (a) The proposed multi-task learning model. (b) Self-attention and guided attention for aspect k. Different aspects share the review encoder and word embedding.

We first use word embedding [22] to map one-hot representations of tokens to a continuous vector space, thus, a review is represented as $(E_{x_1}, E_{x_2}, ..., E_{x_T})$, where E_{x_t} is the word vector of x_t. Then, a bi-directional GRU [5] encoder takes these word vectors as input and turns the review into a sequence of hidden states $H = (h_1, h_2, ..., h_M)$.

4.2 Multi-Aspect Self-Attention

After encoding a review into a sequence of hidden states, our goal is to use these encoded vectors to predict rating scores (i.e., class labels) of different aspects. However, not all of them contribute equally to the predictions, especially for different aspects. Take the review in Fig. 1 as an example. A model might need to focus on "*wait at least 30 minutes*" to predict punctuality score, while for staff, we may put more attention to "*very curt and also very busy*". Therefore, we introduce a multi-aspect self-attention mechanism to capture important parts of each review.

Formally, for an aspect k, we first use a self-attention mechanism [40] to determine attention weights α_t^k of each token in the review

$$u_t^k = (r_{\text{self}}^k)^\top \tanh(W_{\text{self}}^k h_t + b_{\text{self}}^k), \quad \alpha_t^k = \frac{\exp(u_t^k)}{\sum_\tau \exp(u_\tau^k)} \qquad (1)$$

where W_{self}^k, r_{self}^k and b_{self}^k are learnable parameters. Then, the representation of the review under self-attention can be calculated by taking the weighted sum of all hidden states,

$$s^k = \sum_{t=1}^{T} \alpha_t^k h_t \qquad (2)$$

which will be used for the classification task.

4.3 Aspect-Keywords Guided-Attention

The multi-aspect self-attention mechanism relies on model itself to discover relationships between class labels and keywords in a review. However, due to strong correlations of rating scores of different aspects, the model will be confused on 'where to attend', when aspect-level rating scores are the same. In this case, the model may make mistakes, like placing the same class label for all aspects for new coming reviews. This problem can be alleviated by bringing in external knowledge of keywords associated with different aspects (see Section 3.2).

Given a list of aspect-keywords for aspect k, we first obtain the word embedding for them, i.e., $(E_{g_1^k}, E_{g_2^k}, ..., E_{g_M^k})$. Then, each word vector is transformed into a hidden state with[8]

$$v_m^k = (1 - \sigma(W_0^k E_{g_m^k} + b_0^k)) \tanh(W_1^k E_{g_m^k} + b_1^k + b_3^k \sigma(W_2^k E_{g_m^k} + b_2^k)) \qquad (3)$$

It is followed by concatenating all hidden states into a single vector $v^k = [v_1^k, v_2^k, ..., v_M^k]$. Here, the average of all vectors is not taken because we consider that averaging may neutralize some features. We then use the global attention mechanism [1, 20] to calculate alignment scores between encoded vectors of aspect-keywords and tokens in the review as

$$w_t^k = w(v^k, h_t) = (v^k)^\top W_{\text{guide}}^k h_t \qquad (4)$$

where W_{guide}^k are learnable parameters. Thus, the guided-attention weights and vector representation of the review are obtain by

$$\beta_t^k = \frac{\exp(w_t^k)}{\sum_\tau \exp(w_\tau^k)}, \quad c^k = \sum_{t=1}^{T} \beta_t^k h_t \qquad (5)$$

4.4 Aspect-Specific Feature Encoder

From the basic statistics given in Fig. 4, we can observe that some features of doctors (such as specialty and locations) also affect rating scores, therefore, we incorporate them into our model to improve the prediction accuracy. Formally, we embed one hot representations of features of doctors into a continuous vector space for each aspect k as

$$f^k = W_f^k \xi + b_f^k \qquad (6)$$

where W_f^k and b_f^k are model parameters.

4.5 Multi-Aspect Rating Prediction

So far, we have obtained aspect-specific representations of a review via self-attention and guided-attention mechanisms, and representations of features of doctors. These vectors will be concatenated and fed into a classifier, which is a single layer feed-forward network with a softmax activation function, to predict rating scores. The classifier outputs probability distribution of class labels of different aspects with

$$y^k = \text{softmax}(W_{\text{out}}^k [f^k, s^k, c^k] + b_{\text{out}}^k) \qquad (7)$$

where W_{out}^k and b_{out}^k are parameters.

[8] Here, we want to apply a single step GRU transformation for every keyword. But each aspect has only one GRU cell.

Given predicted labels y^k and ground-truth labels $\hat{y}^k$, we train our model in an end-to-end manner using back-propagation, where the loss function is defined as the cross-entropy loss. The goal of the training is to minimize average cross-entropy error between y^k and $\hat{y}^k$ for all aspects. Formally, it is given as

$$\mathcal{L}_\theta = -\sum_{k=1}^{K}\sum_{i=1}^{N}\hat{y}_i^k\log(y_i^k) + \lambda\Omega(\theta) \tag{8}$$

where $\Omega(\theta)$ and λ are a regularizer and a scalar, respectively. θ is a parameter set including all weight matrices and bias vectors. N represents the number of classes.

5 EXPERIMENTS

In this section, we conduct an extensive set of experiments on ratemds dataset for document-level multi-aspect sentiment classification and explain different experimental results. We start with introducing two subsets of ratemds dataset, baseline methods, and implementation details of the proposed model and evaluation metrics. Then, we will show the classification performance of different models along with some qualitative results.

5.1 Datasets Used

We created two subsets from ratemds dataset, i.e., ratemds-us and ratemds-ca, based on countries that doctors work in. We chose the US and CA, because 90% of reviews are from these two countries (see Fig. 4 (a)). The ratemds-us consists of 1,414,235 reviews for 385,407 doctors, while ratemds-ca has 1,252,941 reviews for 99,719 doctors. We first tokenized texts with SpaCy[9]. Since features of doctors are used as additional input, we also extracted attributes of doctors, including specialties, insurance plans, locations, genders and facilities, and transform them into one-hot representations. In addition, aspect-keywords were selected from latent topics. Finally, we randomly split each dataset into training, development and testing sets at the ratio of 0.8/0.1/0.1.

5.2 Compared Methods

We compare the proposed model with different baseline methods, including conventional classification and deep learning models.

- **MAJOR**. This method simply uses the majority label of each aspect in the training set as the prediction label.
- **GLVL**. In this model, we first calculate the vector representation of each review by taking the average of vectors of all keywords in the review. Word vectors were pre-trained on the twitter datasets with 2 billion tweets by GloVe [26]. Then, we use LIBLINEAR [7] package[10] for the classification task.
- **BOWL**. This model feeds Bag-Of-Words (BOW) representations of reviews into LIBLINEAR package for the sentiment classification. In the experiment, we have removed stop-words and punctuation in textual reviews to make the model capture keywords efficiently.
- **CNN**. We adopt convolutional neural network (CNN) structure proposed in [12, 42] for the rating prediction of reviews. In our experiments, 1-directional convolutions with different filter sizes along sequence time-step dimension are first applied to the word embedding of a review. Then, a max-over-time pooling operation [6] is built upon each feature map. By selecting the maximum value, we obtain the key-feature of each filter. Finally, the vector representation of the review is obtained by concatenating all features. This vector will be fed into a feed-forward network for classification (Similar for other deep learning models.).
- **GRU**. We use GRU to refer to a bi-directional GRU with multiple hidden layers [5]. In this model, we concatenate output vectors of the last hidden states of the top hidden layer in both forward and backward directions[11] to represent a review.
- **GRU-ATN**. GRU-ATN first builds a self-attention layer [16, 32] on top of a recurrent neural network. With attention weights, we can compute a context vector for a review by taking the weighted sum of all hidden states (see Eq. (1)).
- **MT-BASE and MT-FEAT**. MT-BASE is a multi-task learning framework with only a review encoder, self-attention layers and classifiers (see Fig. 5) [41]. In this model, different tasks (i.e., aspects) share the same review encoder. MT-FEAT also takes features of doctors into consideration.

5.3 Implementation Details

We implemented all deep learning models using Pytorch [25] and model parameters are selected based on the development set. For both ratemds-us and ratemds-ca, vocabulary sizes are set to 50,000. We do not use the pre-trained word embeddings [22, 26] and they are learned from scratch during the training. The dimension of word embeddings is set to 128. For CNN, filter sizes were chosen to be 3, 4, 5 and the number of filters are 100 for each size. For all GRU based models, the dimension of hidden states is set to 128 and the number of layers is 2. All parameters are trained with ADAM [13] optimizer with learning rate 0.0001. Gradient clipping has also been applied to prevent gradient explosion. Our codes are available at https://github.com/tshi04/dmsc_ratemds.

In this paper, we adopt 'macro' averaged F-score and mean squared error (MSE) to evaluate performance of different models. Accuracy has been used in [14, 41], however their models are only tested on reviews with different aspect-level ratings, since those with identical aspect-level ratings can make it difficult for their models to distinguish keywords of different aspects. In our experiments, these reviews are still kept, because we assume that aspect-keywords guided-attention mechanism can alleviate this problem. However, based on distributions of aspect-level ratings and their correlations (see Fig. 3), the data is highly imbalanced, therefore, accuracy is not a suitable evaluation metric and we adopt F-score instead. Both accuracy and F-score are based on exact match of class labels, however, for sentiment analysis, we only need predicted rating scores close to the ground-truth. For example, if the ground truth score is 5, a model still performs reasonably well by predicting 4. Therefore, MSE is also a promising metric.

5.4 Rating Prediction Performance

We first present quantitative results of different models in Tables 2 and 3, where we use bold font to show the best performance values and underline to highlight the second best values.

From these two tables, we can observe that MAJOR gets the lowest performance among all compared methods, since it simply

[9] https://spacy.io/

[10] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[11] Here, the first token in a sequence corresponding to the last hidden state in the backward direction.

Table 2: Performance comparison of different models on ratemds-us. For MSE, smaller is better.

	Staff		Punctuality		Helpfulness		Knowledge	
	F-score	MSE	F-score	MSE	F-score	MSE	F-score	MSE
MAJOR	0.1453	3.6394	0.1370	3.7749	0.1546	4.5445	0.1575	3.8039
GLVL	0.2893	1.9486	0.2777	2.0598	0.3341	1.4356	0.3140	1.6360
BOWL	0.3805	1.3691	0.3744	1.4440	0.4142	0.8564	0.4151	1.0056
CNN	0.3767	1.1588	0.3721	1.2375	0.4208	0.5355	0.4205	0.7079
GRU	0.4101	0.9717	0.3885	1.1000	0.4602	0.4617	0.4419	0.6326
GRU-ATN	0.4090	0.9638	0.3896	1.0938	0.4479	0.4817	0.4597	0.6078
MT-BASE	0.4093	0.9495	0.3997	1.0273	0.4554	0.4569	0.4528	0.5993
MT-FEAT	0.4187	0.9456	0.3976	1.0443	0.4684	0.4461	0.4721	0.5722
MT-FAKGA (our)	**0.4193**	**0.9061**	**0.4103**	**1.0018**	**0.4787**	**0.4437**	**0.4822**	**0.5681**

Table 3: Performance comparison of different models on ratemds-ca.

	Staff		Punctuality		Helpfulness		Knowledge	
	F-score	MSE	F-score	MSE	F-score	MSE	F-score	MSE
MAJOR	0.1466	3.1578	0.1377	3.3958	0.1590	3.8706	0.1613	3.2678
GLVL	0.2665	2.1426	0.2645	2.1774	0.3209	1.6168	0.3028	1.6960
BOWL	0.3663	1.4573	0.3651	1.5007	0.4239	0.8667	0.4179	0.9554
CNN	0.3480	1.3431	0.3568	1.3520	0.4267	0.5871	0.4197	0.7042
GRU	0.3778	1.1466	0.3958	1.1282	0.4714	0.4742	0.4519	0.5977
GRU-ATN	0.3907	1.0910	0.3891	1.1457	0.4827	0.4743	0.4739	0.5714
MT-BASE	0.3894	1.0730	0.3905	1.1205	0.4806	0.4686	0.4759	0.5568
MT-FEAT	0.3965	1.0838	0.3916	1.1020	0.4856	0.4556	0.4833	0.5362
MT-FAKGA (our)	**0.4013**	**1.0403**	**0.3965**	**1.0781**	**0.5051**	**0.4432**	**0.5025**	**0.5203**

classifies all reviews to the dominant labels without using textual reviews. GLVL achieves much better results than MAJOR, but is still not as good as other methods. Although it attempts to take advantage of semantic information of the word embedding, simply averaging all word vectors in a review can cause information offset, which results in poor review representation[12]. BOWL can also capture word-level semantic information via bag-of-word (BOW) representations of reviews. It performs significantly better than GLVL and as good as CNN. Compared to GLVL, BOW representation encodes each review into a high-dimensional space, thus, BOWL requires more parameters to classify reviews which avoids under-fitting. On the other hand, by removing stop-words and punctuation in reviews, we only keep keywords relevant to classification and frequency of keywords in a review can partially reflect their importance. Therefore, representations of reviews by BOWL are better than those obtained by GLVL.

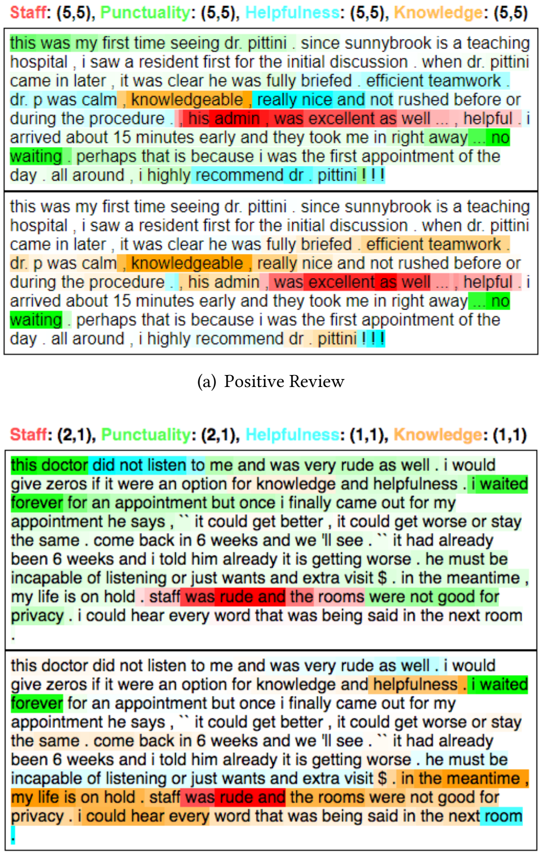
Compared to traditional methods and CNN, GRU based models have achieved significantly better results on both datasets. GRU and GRU-ATN are simple classification methods and trained separately for different aspects, while MT-BASE, MT-FEAT, MT-FAKGA are multi-task models and they share the word embedding and recurrent hidden layers. Since most model parameters are attributed to these layers, multi-task models require significantly fewer parameters. Moreover, GRU and GRU-ATN need K different training for K different aspects, while multi-task learning framework can simultaneously learn different aspects, thus, they require much lesser training time. As to the performance of rating predictions, we first observe that multi-task learning models can perform as good as or even better than GRU and attention-based GRU models. MT-FEAT performs slightly better than MT-BASE in most cases, since it considers features of doctors. By incorporating knowledge from aspect-keywords, we further improve the performance of MT-FEAT. The proposed MT-FAKGA achieves the best results in terms of F-score and MSE on both datasets.

5.5 Attention Visualization

As attention mechanism enables a model to selectively focus on important parts of reviews, visualization of attention weights has become a popular tool that helps to interpret models and analyze experimental results [38, 41]. Specific to our multi-aspect classification task, our goal is to investigate if models accurately attend keywords of different aspects or not.

In Fig. 6, we first show one example with positive ratings and one with negative ratings. In these examples, the proposed model makes correct predictions of sentiment, and reviews contain keywords of all four different aspects, therefore, we only need to check if the model can successfully detect these keywords. Take Fig. 6(a) as an example, both self-attention and guided-attention focus on "*excellent, helpful*" for staff. As to punctuality, both of them capture "*no waiting*". However, self-attention also highlights *"this was my first time ..."* which is not quite relevant. Helpfulness and knowledge are often difficult to be distinguished in many examples. Here, self-attention focuses on *"efficient teamwork, calm, really nice and not rush"* for helpfulness, while guided-attention does not successfully detect these keywords, which might be because the extracted aspect-keywords do not align well with *"calm, nice, rush"*. Finally, for knowledge, both mechanisms capture *"knowledgeable"*. The guided-attention also treat *"efficient teamwork"* as knowledge aspect keywords, which is reasonable. For the negative review (see Fig. 6 (b)), both self-attention and guided-attention highlight *"rude"* for staff, and *"i waited forever"* for punctuality. Therefore, the model predicts a rating score of 1 for both aspects, which is consistent with ground-truth in sentiment sense. As to helpfulness, guided-attention incorrectly attends *"room"*. However, it also focuses on *"he must be incapable of listening or just wants and extra visit"* which

[12] Before averaging word vectors, we have removed stop-words and punctuation from reviews. However, the performance has not improved significantly using this trick.



Staff: (5,5), Punctuality: (5,5), Helpfulness: (5,5), Knowledge: (5,5)

this was my first time seeing dr. pittini . since sunnybrook is a teaching hospital , i saw a resident first for the initial discussion . when dr. pittini came in later , it was clear he was fully briefed . efficient teamwork . dr. p was calm , knowledgeable , really nice and not rushed before or during the procedure . , his admin , was excellent as well ... , helpful . i arrived about 15 minutes early and they took me in right away ... no waiting . perhaps that is because i was the first appointment of the day . all around , i highly recommend dr . pittini ! ! !

this was my first time seeing dr. pittini . since sunnybrook is a teaching hospital , i saw a resident first for the initial discussion . when dr. pittini came in later , it was clear he was fully briefed . efficient teamwork . dr. p was calm , knowledgeable , really nice and not rushed before or during the procedure . , his admin , was excellent as well ... , helpful . i arrived about 15 minutes early and they took me in right away ... no waiting . perhaps that is because i was the first appointment of the day . all around , i highly recommend dr . pittini ! ! !

(a) Positive Review

Staff: (2,1), Punctuality: (2,1), Helpfulness: (1,1), Knowledge: (1,1)

this doctor did not listen to me and was very rude as well . i would give zeros if it were an option for knowledge and helpfulness . i waited forever for an appointment but once i finally came out for my appointment he says , `` it could get better , it could get worse or stay the same . come back in 6 weeks and we 'll see . `` it had already been 6 weeks and i told him already it is getting worse . he must be incapable of listening or just wants and extra visit $. in the meantime , my life is on hold . staff was rude and the rooms were not good for privacy . i could hear every word that was being said in the next room .

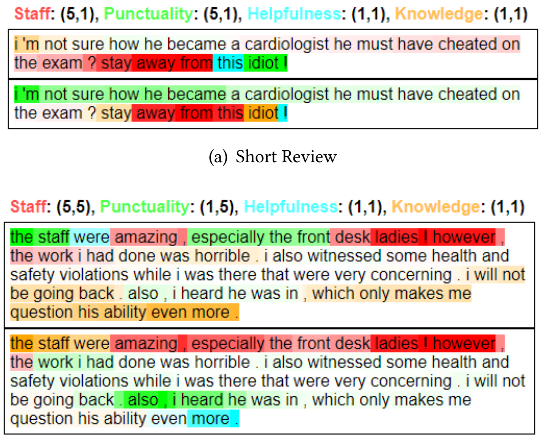
this doctor did not listen to me and was very rude as well . i would give zeros if it were an option for knowledge and helpfulness . i waited forever for an appointment but once i finally came out for my appointment he says , `` it could get better , it could get worse or stay the same . come back in 6 weeks and we 'll see . `` it had already been 6 weeks and i told him already it is getting worse . he must be incapable of listening or just wants and extra visit $. in the meantime , my life is on hold . staff was rude and the rooms were not good for privacy . i could hear every word that was being said in the next room .

(b) Negative Review

Figure 6: Visualization of attention weights. In parentheses, first and second numbers represent ground-truth and predicted ratings, respectively. For each sub-figure, the first and second rows represent self-attention and guided-attention weights, respectively. Different aspects are labeled with different colors, therefore, this figure is best viewed in color.

reflects the fact that the doctor does not help. On the other hand, self-attention focuses on *"did not listen"*, which is also good. Finally, we observe that self-attention fails to capture knowledge aspect keywords, while guided-attention highlights *"helpfulness, my life is on hold, rooms were not good for privacy"*, which can partially indicate that the patient is not happy with the knowledge of this doctor.

As we can see from above examples, attention mechanism cannot always build accurate connections between rating and keywords of the same aspect. In practice, we found that failure of attention may be caused by several reasons: 1) A review is very short and only discusses a certain issue. For example, in Fig. 7 (a), the patient first questioned the knowledge of the doctor and then suggested others to stay away from him/her. However, it does not mention anything about staff and punctuality. Therefore, both self-attention and guided-attention make mistakes in finding aspect-keywords, which will then result in incorrect predictions. 2) A review is long enough, but does not cover all aspects. Fig. 7 (b) shows an example in which the patient did not mention anything about punctuality. Thus, *"the staff, also, i heard"* are highlighted for this aspect, which lead to the opposite sentiment. 3) We may need some reasoning for a review to make predictions. For example, some reviews start with *"dr. started out being an excellent doctor for us."*, then the patients begin to complain about different issues. 4) Many keywords and phrases are ambiguous in different context, such as *"long"* in *"wait very long"* and *"he has been my doctor very long"*.



Staff: (5,1), Punctuality: (5,1), Helpfulness: (1,1), Knowledge: (1,1)

i 'm not sure how he became a cardiologist he must have cheated on the exam ? stay away from this idiot !

i 'm not sure how he became a cardiologist he must have cheated on the exam ? stay away from this idiot !

(a) Short Review

Staff: (5,5), Punctuality: (1,5), Helpfulness: (1,1), Knowledge: (1,1)

the staff were amazing , especially the front desk ladies ! however , the work i had done was horrible . i also witnessed some health and safety violations while i was there that were very concerning . i will not be going back . also , i heard he was in , which only makes me question his ability even more .

the staff were amazing , especially the front desk ladies ! however , the work i had done was horrible . i also witnessed some health and safety violations while i was there that were very concerning . i will not be going back . also , i heard he was in , which only makes me question his ability even more .

(b) Review does not cover punctuality.

Figure 7: Visualization of attention weights. This figure will be best viewed in color.

5.6 Practical Implications

In this section, we describe the practical applications of our tool. Similar to the example shown in Fig. 1, our tool can highlight keywords corresponding to different aspects, so that both patients and doctors can get the important information from these reviews more efficiently. For doctors, they can find out their problems by just visualizing keywords of the aspects with negative ratings. For example, if the punctuality is a problem in a clinic, then, "wait very long" may appear in many reviews. Coloring these keywords can help doctors to find out this problem in seconds. On the other hand, patients may need to read the reviews of many doctors, which takes a long time, before they can find their primary care physicians or specialists. However, if they are trying to find a doctor who is caring and helpful, they can use this tool, which can also highlight the keywords of positive and negative sentiment with different colors for aspect “helpfulness”, to see the experience of other patients instead of browsing all reviews.

6 CONCLUSION

Online doctor review systems provide a platform for patients to give feedback to their doctors. These reviews not only help other patients to learn more about a doctor before they visit, but also help doctors to improve their service quality. From these reviews, we can also discover common concerns of patients and existing problems in clinics. In this paper, we systematically investigated the dataset from one such review systems, i.e., ratemds.com, where each review comes with an overall rating and ratings for four different aspects. We first studied statistics of reviews, ratings and doctors. Then, we attempted to explore content of reviews by extracting aspect-keywords with the topic modeling. We proposed a multi-task learning framework for the document-level multi-aspect sentiment classification, which can help us to not only recover missing aspect-level ratings and detect inconsistent rating scores, but also identify aspect-keywords in a given review based on ratings. The proposed model takes both features of doctors and aspect-keywords into consideration. Extensive experiments have been conducted on two subsets of ratemds dataset to demonstrate effectiveness of the proposed model. Qualitative results show the power of attention mechanisms and reveal some linguistic problems in the textual reviews. In the future, we will work on solving these problems and applying fine-grained aspect-based sentiment classification techniques to study these reviews.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation grants IIS-1619028, IIS-1707498, and IIS-1838730.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Migena Ceyhan, Zeynep Orhan, and Elton Domnori. 2017. Health service quality measurement from patient reviews in Turkish by opinion mining. In *CMBEBIH 2017*. Springer, 649–653.

[4] Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1650–1659.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, Aug (2011), 2493–2537.

[7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research* 9, Aug (2008), 1871–1874.

[8] Sunir Gohil, Sabine Vuik, and Ara Darzi. 2018. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. *JMIR public health and surveillance* 4, 2 (2018).

[9] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

[10] Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*. 856–864.

[11] Anthony M Hopper and Maria Uriyo. 2015. Using sentiment analysis to review patient satisfaction data located on the internet. *Journal of health organization and management* 29, 2 (2015), 221–233.

[12] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[15] Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level Multi-aspect Sentiment Classification by Jointly Modeling Users, Aspects, and Overall Ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*. 925–936.

[16] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[17] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[18] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. 2017. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442* (2017).

[19] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 11th IEEE International Conference on Data Mining Workshops*. IEEE, 81–88.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[21] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 1020–1025.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[23] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. [n. d.]. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* ([n. d.]).

[24] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[27] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 19–30.

[28] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 27–35.

[29] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1105–1114.

[30] Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *arXiv preprint arXiv:1706.03446* (2017).

[31] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[32] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1422–1432.

[33] Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900* (2016).

[34] Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 417–424.

[35] Bo Wang and Min Liu. 2015. Deep learning for aspect-based sentiment analysis.

[36] Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 783–792.

[37] Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[39] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1445–1456.

[40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*. 1480–1489.

[41] Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-Level Multi-Aspect Sentiment Classification as Machine Comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2044–2054.

[42] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2017. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709* (2017).

[43] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2018), e1253.