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

-Tian Shi, Vineeth Rakesh, Suhang Wang, Chandan K. Reddy

Outline

- Contributions
- Part 1
  - Preliminary Data Analysis on RateMDs dataset
  - Topic Modelling and Aspect Keywords
- Part 2
  - Problem Statement
  - Architecture and Code
- Part 3
  - Baseline methods and Results
  - Attention visualisations and Discussion.

Contributions

- Introduce a new dataset with more than 2 million reviews with multi-aspect ratings.
- A comprehensive statistical analysis on this dataset.
- A framework for multi task learning that takes into account
  - Features of doctors and
  - Aspect keywords discovered by topic model.

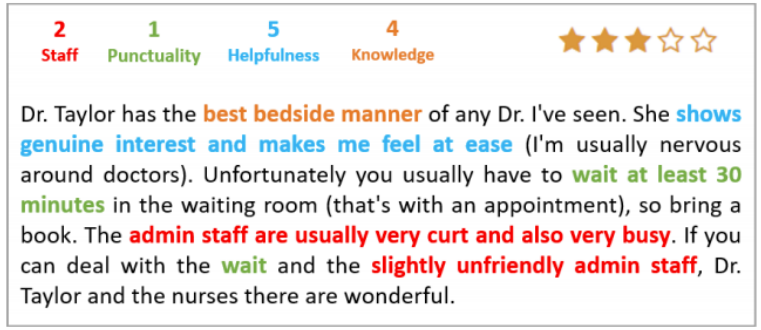
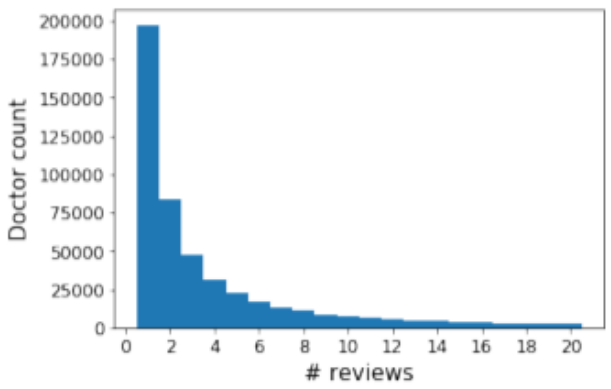


Figure 1: An example of ratemds reviews. Keywords corresponding to different aspects are highlighted with different colors.

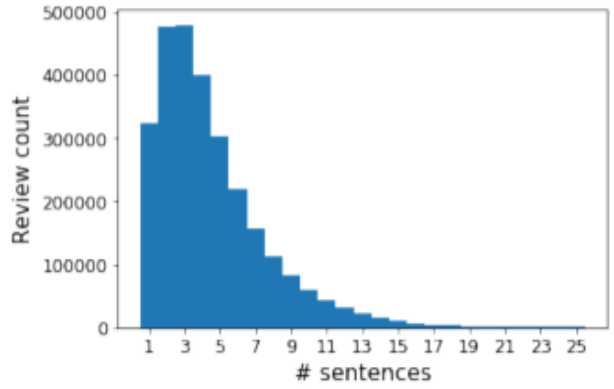
*Figure 1 taken from paper

Preliminary Data Analysis

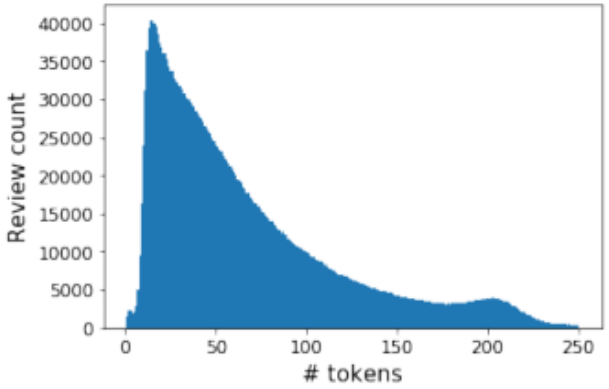
- Dropped records with missing ratings and missing reviews- 500K doctors, 2.7M reviews
- Average number of ratings per doctor is 4.6 and almost 40% doctors have single review.
- Short reviews are a problem because
  - Contain single aspect
  - Topic modelling is difficult
- Most reviews have more than 2 sentences and most sentences have over 12 tokens.
- Average length of reviews is more than 4 sentences and 72 tokens.
- This implies that there are a number of reviews whose content covers all four aspects in this dataset.



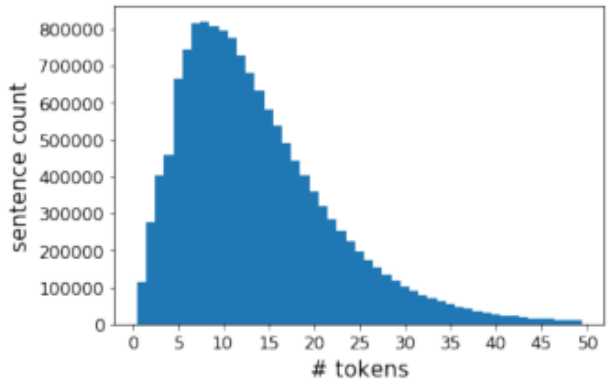
(a) # of reviews for doctors.



(b) # of sentences in reviews.



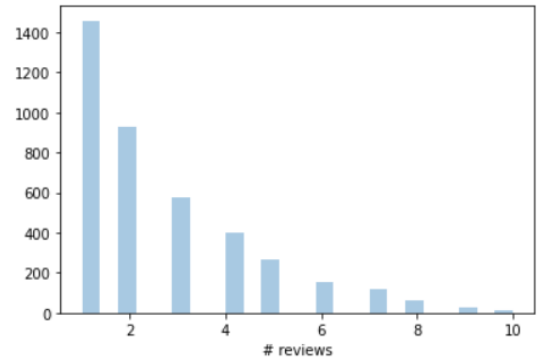
(c) # of tokens in reviews.

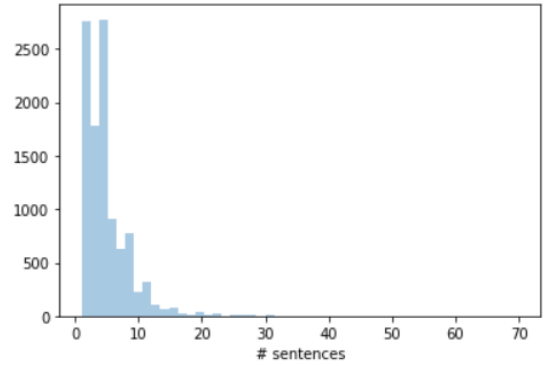


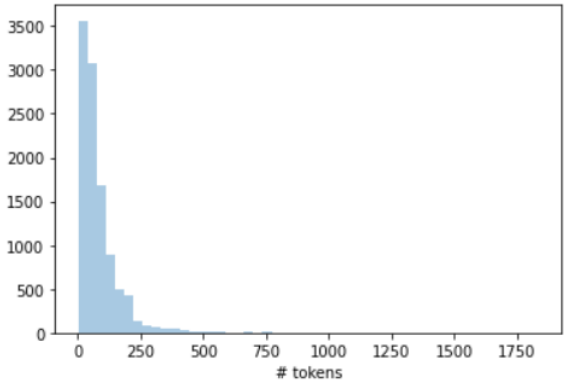
(d) # of tokens in sentences.

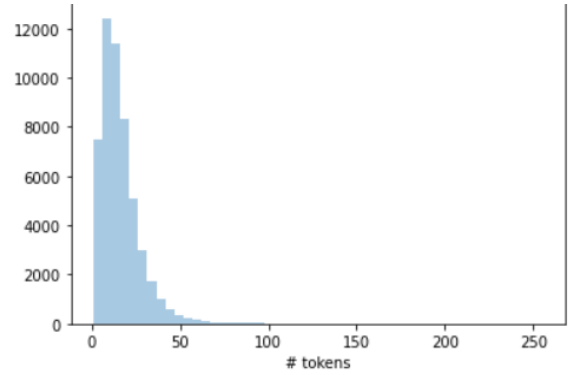
*Figures taken from paper

- Most doctors have single review.
- Most reviews have more than 2 sentences. However, a large number of reviews also have a single sentence.
- Most sentences have more than 10 tokens.
- Some outliers had ~1k tokens.









Ratings distribution

- More than 60% reviews have an aspect rating of 5 in all aspects. Most patients are satisfied with their visits.
- 17% of the ratings are 1.
- Many patients are slightly unsatisfied with staff and punctuality even if they are satisfied with their doctors. This may be because of appointments and waiting time.

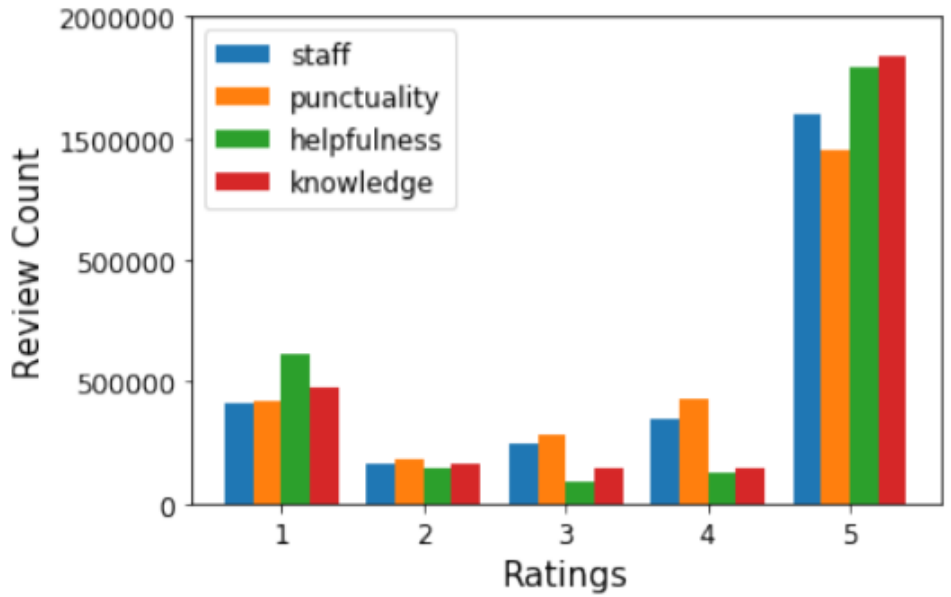
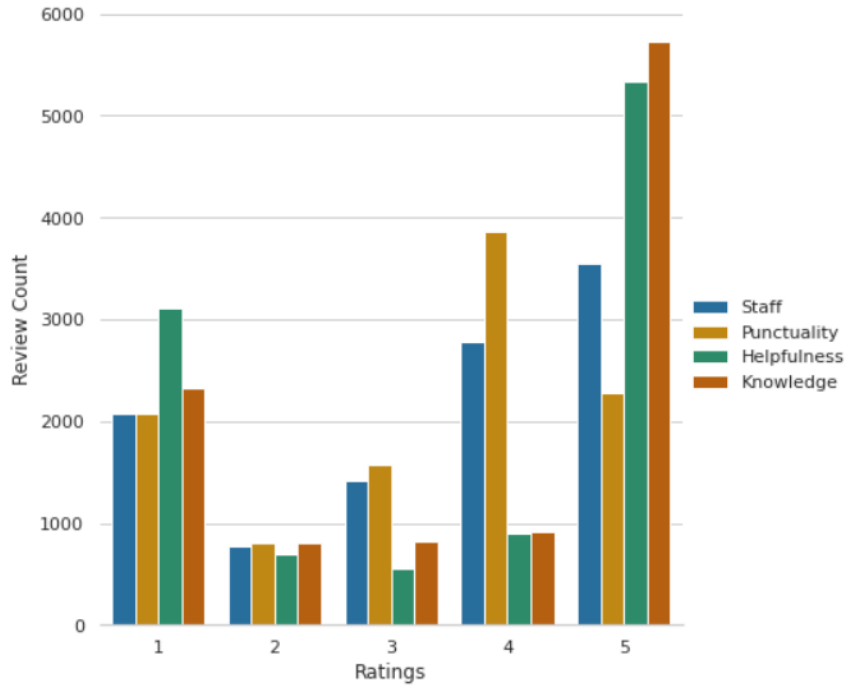


*Figure on left taken from paper

- Most reviews have rating 5 for helpfulness, staff and knowledge aspects.
- Punctuality remains the biggest inconvenience to patients as in the original dataset.
- Patients tend to give very low ratings if the doctor is not helpful. This trend is followed in both datasets.



Doctors' statistics

- Dentists have a much higher score on average than others.
- General practitioners have a lower punctuality score as patients with any issue can visit them and get referrals when they have complicated health issues.
- Using these features in the model might help with accuracy.
- Key features of doctors in the dataset are : gender, facility categories, specialties, locations and insurance plans.

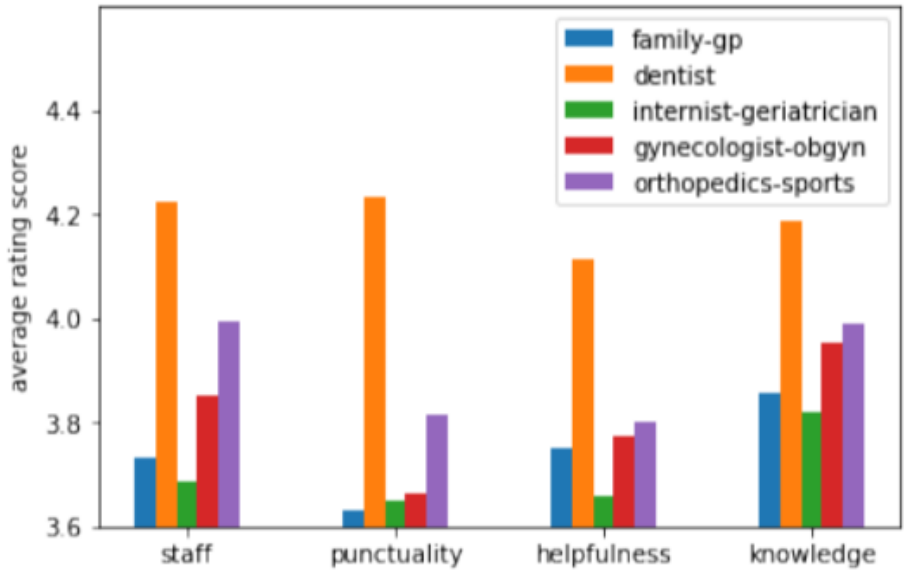


*Figure on left taken from paper

- Dentists have higher scores in most aspects.
- Key features of doctors in the dataset are : gender and specialties.
- There are 57 different specialities in both datasets.

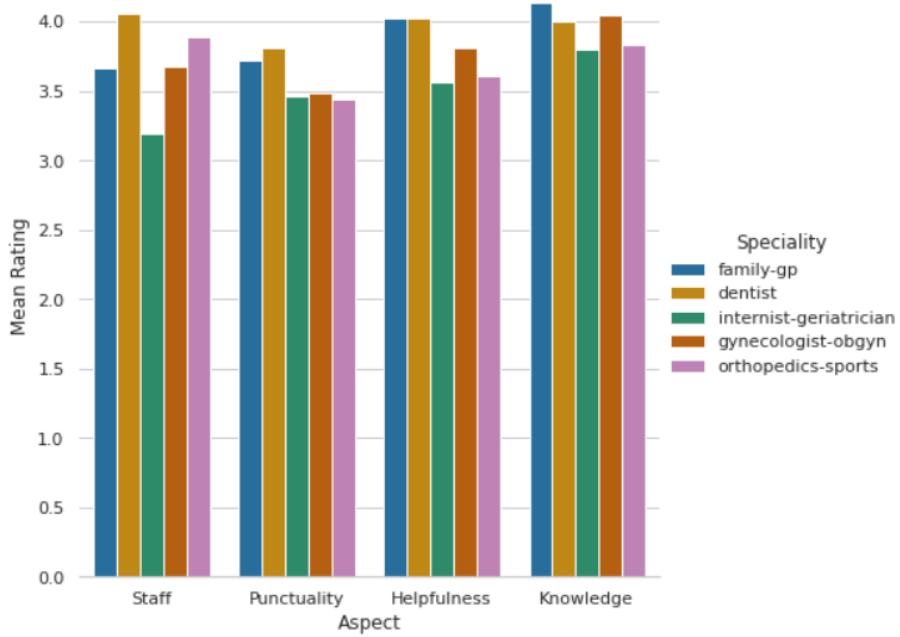


Table 1: Aspect-keywords extracted with the topic model.

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.

*Table taken from paper

- Preprocessing :Tokenization using spacy, removal of stopwords and rare words.
- LDA is run with 10 topics. Top 20 Keywords from each topic are empirically assigned to different aspects.
- Keywords of staff, punctuality and helpfulness are similar across specialities. However knowledge has different keywords for each field.

	Speciality	Aspect	Keyword Examples
0	family-gp	staff	clinic, staff, office, rude, nurse, service, charge, visit, contact, insurance, follow, phone
1	family-gp	punctuality	wait, hour, long, time, late, appointment, minute
2	family-gp	helpfulness	care, listen, regard, consider, refer, understanding, feel, thank, concern, love, help, helpfulness, helpful, like, answer
3	family-gp	knowledge	lab, symptom, treatment, professional, medicine, knowledge, drug, skill, prescription, diagnosis, physician
4	dentist	staff	insurance, charge, service, receive, nice, kind, smile, polite, sweet, respect, assistant, visit, office, friendly
5	dentist	punctuality	rush, drive, late, time, appointment, wait, day, long, year
6	dentist	helpfulness	help, feel, comfortable, ease, care, ask, follow, helpful, care, feel, love
7	dentist	knowledge	knowledgeable, procedure, explain, treatment, implant, review, replace, perform, extraction, experience, professional, tooth, cleaning, dental
8	gynecologist-obgyn	staff	tell, ask, nurse, rude, staff, office, nice, friendly, service
9	gynecologist-obgyn	punctuality	time, wait, appointment, hour, long, minute, day, week, rush
10	gynecologist-obgyn	helpfulness	care, concern, understanding, warm, ease, helpful, think, save, offer, answer, consider, refuse, suggest
11	gynecologist-obgyn	knowledge	knowledgeable, test, exam, review, explain, complication, pregnancy, deliver, experience, baby, surgery, pain, hysterectomy, surgeon, medication, bleed, fibroid, treatment, diagnosis, scar

Notations

- Problem Statement : Given
  - plain text review,
  - aspect keywords from topic models, and
  - features of doctors

predict rating for all 4 aspects (Staff, Punctuality, Helpfulness and Knowledge).

- Textual Review X = (x1,x2,…xT)
- Keywords for different aspects G = (G1,G2,…G_k)
  - G_i = (g1i,g2i,..g_mi) -> m keywords for ith aspect.
- Features ξ
- Labels y = (y1,y2,…yk)

Architecture

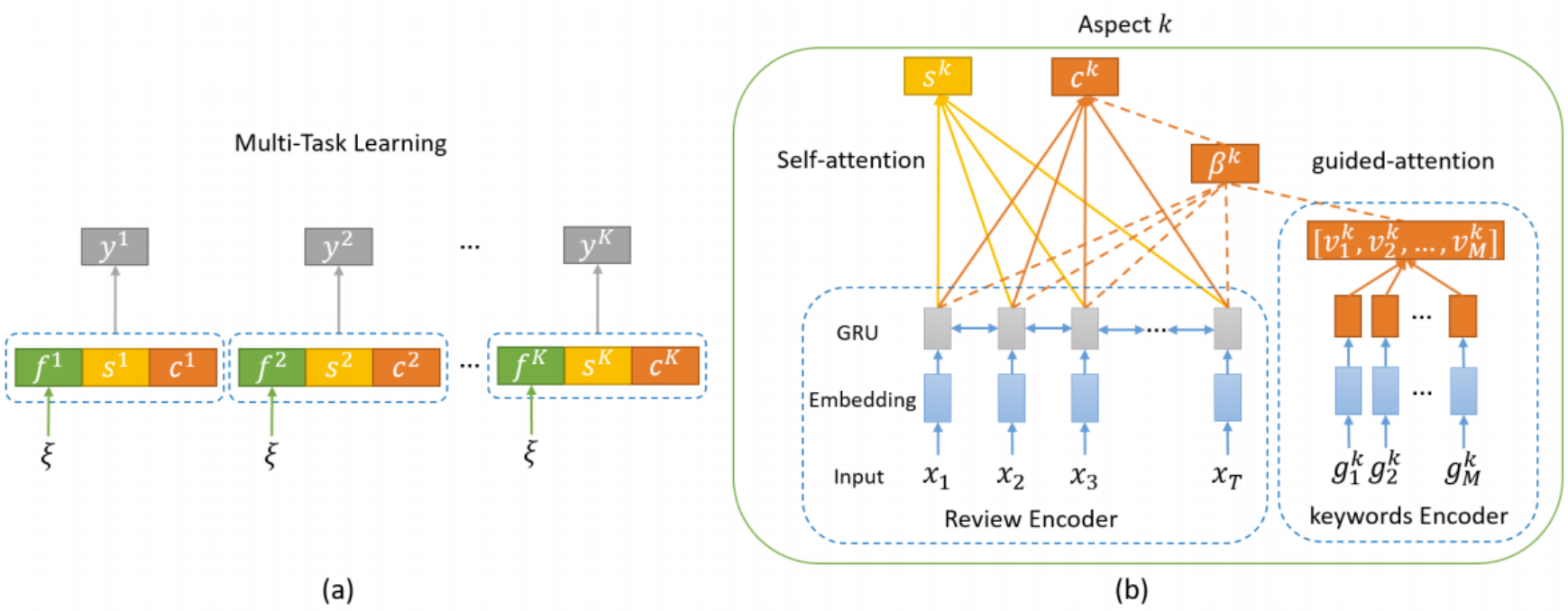


Figure 5: An illustration of the model architecture. (a) The proposed multi-task learning model. (b) Self-attention and guided attention for aspect k. Different aspects share the review encoder and word embedding.

*Figure taken from paper

- Review Encoder : Embedding layer followed by bi-directional GRU. Generates hidden states for each word in input text.

$$(E_{x_1}, E_{x_2}, \ldots, E_{x_T})$$

$$H = (h_1, h_2, \ldots, h_M)$$

- Multi Aspect Self Attention : To capture important parts of review.

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

```
class MultiAspectSelfAttention(nn.Module):
  def __init__(self,params):
    super(MultiAspectSelfAttention, self).__init__()
    self.r = torch.autograd.Variable(torch.randn(params.hidden_dim,device = params.device),requires_grad = True)
    self.linear_layer = nn.Linear(params.hidden_dim *2, params.hidden_dim)

  def forward(self,h,h_mask):
    reduced_h = torch.tanh(self.linear_layer(h))
    u = torch.bmm(reduced_h,self.r.unsqueeze(0).repeat(reduced_h.shape[0],1).unsqueeze(2))

    h_mask = h_mask[:,:u.shape[1]].unsqueeze(2)
    u = u.masked_fill(h_mask,-np.inf)

    alpha = torch.softmax(u,dim=1)
    final_rep  = torch.bmm(h.permute(0,2,1),alpha)

    return final_rep.squeeze(2),alpha
```

- Aspect Keywords Guided Attention : Brings in external knowledge of keywords associated with different aspects.

$$(E_{g_1^k}, E_{g_2^k}, \ldots, E_{g_M^k}).$$

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

$$v^k = [v_1^k, v_2^k, \ldots, v_M^k]$$

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

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

```
class AspectKeywordsGuidedAttention(nn.Module):
  def __init__(self,params,num_kwd):
    super(AspectKeywordsGuidedAttention,self).__init__()
    self.linear0 = nn.Linear(params.hidden_dim,params.hidden_dim)
    self.linear1 = nn.Linear(params.hidden_dim,params.hidden_dim)
    self.linear2 = nn.Linear(params.hidden_dim,params.hidden_dim)
    self.bias3 = torch.autograd.Variable(torch.randn(params.hidden_dim,device = params.device),requires_grad = True)
    self.W_guide = nn.Linear(params.hidden_dim*2,params.hidden_dim*num_kwd,bias = False)

  def forward(self,h,h_mask,aspect_kwd_embeddings):
    #aspect_kwd_embeddings: num_kwd X embed_dim
    aspect_kwd_embeddings2 = torch.sigmoid(self.linear2(aspect_kwd_embeddings))
    aspect_kwd_embeddings2 = torch.mul(self.bias3,aspect_kwd_embeddings2)

    aspect_kwd_embeddings1 = torch.tanh(self.linear1(aspect_kwd_embeddings) + aspect_kwd_embeddings2)
    update_gate = 1 - torch.sigmoid(self.linear0(aspect_kwd_embeddings))

    vk = torch.mul(update_gate,aspect_kwd_embeddings1)
    vk = vk.view(1,-1).squeeze(0).contiguous()

    reduced_h = self.W_guide(h)

    wkt = torch.bmm(reduced_h,vk.unsqueeze(0).repeat(reduced_h.shape[0],1).unsqueeze(2))
    h_mask = h_mask[:,:wkt.shape[1]].unsqueeze(2)

    wkt = wkt.masked_fill(h_mask,-np.inf)

    beta = torch.softmax(wkt,dim=1)
    final_rep = torch.bmm(h.permute(0,2,1),beta)

    return final_rep.squeeze(2),beta
```

- Aspect Specific Feature Encoder : Embeds one hot feature vector ξ into continuous feature space.

$$f^k = W_f^k \xi + b_f^k$$

- Classifiers and Loss Function

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

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

  - N is the number of classes.

Compared Models

- MAJOR : Majority label for each aspect.
- GLVL : Average of Glove vectors for review fed into linear classifier.
- BOWL : Stopwords and punctuations removed. Bag of words representation fed into linear classifier.
- CNN : Architecture from Yoon Kim’s paper trained separately for each aspect.
- GRU : Bi-directional GRU with 2 layers. Concatenated [fwd,bkwd] hidden state of top layer used as representation for review document.
- GRU-ATTN : Self attention layer on top of GRU hidden states.
- MT-BASE : Architecture from paper with review encoder, self attention layer and classifier.
- MT-FEAT : MT-BASE + Doctor Features
- MT-FAKGA : MT-FEAT + Aspect Keyword Guided Attention

Last three are multi-task learning frameworks sharing same review encoding layer and single loss function .

  - Hence less parameters and less training time.

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**

*Table taken from paper

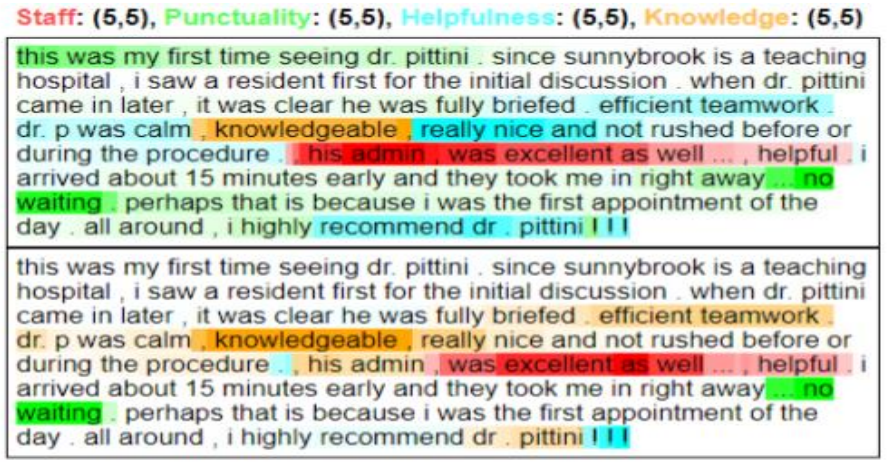
Model	F@S	MSE@S	F@P	MSE@P	F@H	MSE@H	F@K	MSE@K
Major	0.10	4.57	0.10	2.41	0.13	5.59	0.14	4.70
GLVL	0.29	2.69	0.24	2.05	0.29	2.72	0.29	2.65
MT-Base	0.31	2.07	0.27	1.77	0.36	1.40	0.37	1.64
MT-Feat	**0.35**	**1.97**	**0.35**	1.85	0.39	1.33	0.39	1.47
MT-FAKGA	**0.35**	2.02	**0.35**	**1.69**	**0.41**	**1.24**	**0.40**	**1.40**

Attention Visualisation

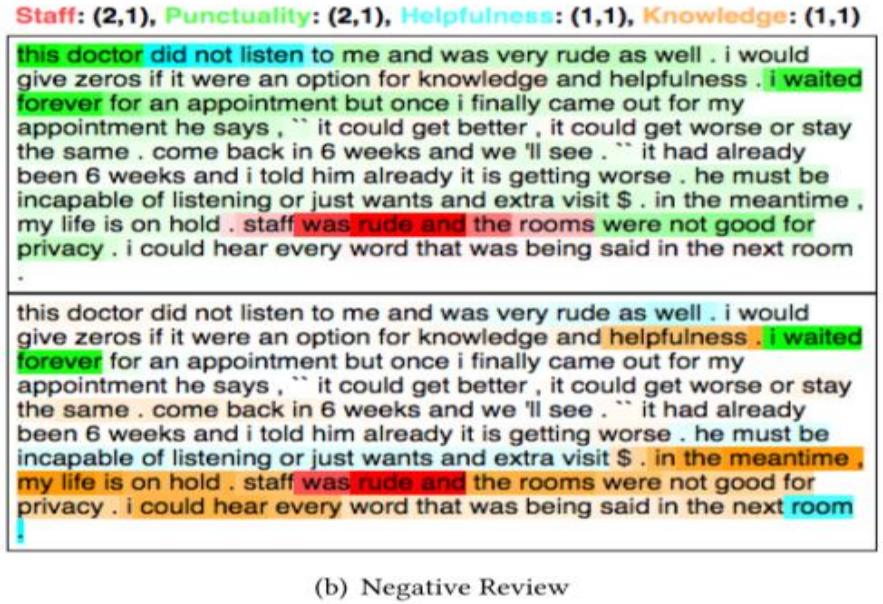
- For first review:
  - Staff : both attentions capture “excellent, helpful”
  - Punctuality : “no waiting”
  - Helpfulness : keyword guided attention does not detect “efficient teamwork, nice and not rush..” possibly because the keywords don’t align with these words.
  - Knowledge : both identify “knowledgeable”

- For second review:
  - Staff : “rude”
  - Punctuality : “I waited forever”
  - Helpfulness : wrong selection “room”
  - Knowledge : only partially correct attention. Ignores second line.

- Helpfulness and knowledge are often difficult to be distinguished.



(a) Positive Review



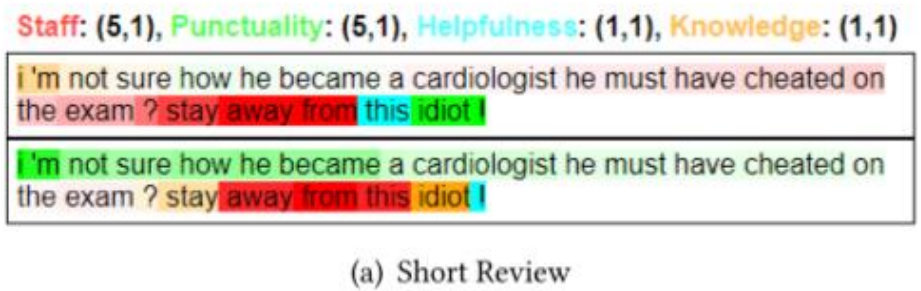
(b) Negative Review

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.

*Figure taken from paper

Reasons for failure of Attention.

- Reviews missing certain aspects.
  - Staff and punctuality are not covered in the review. Attention mechanism makes mistakes.
- When reasoning is needed for making predictions. Example “Dr. X started out as an excellent doctor for us…” and then the review has complaints.
- Many keywords and phrases are ambiguous in different context. Exampe “long” in “wait very long” and “has been my doctor very long”



(a) Short Review

*Figure taken from paper

Attention Visualisation

- Staff : Keyword guided attention correctly identifies “competent” and “caring”
- For most aspects self attention only focusses on “Wonderful”

- For most examples adjectives like “wonderful”, “great”, “amazing”, “horrible”, “poor” etc were the only words that the model paid attention to.
- The model made most mistakes on reviews that did not cover all aspects.
- This method works well if most reviews in dataset are sure to contain all aspects. Otherwise model switches to ratings based on adjectives.

Thank You!