MARTA: Leveraging Human Rationales for Explainable Text Classification – Supplementary Material –

Ines Arous[1], Ljiljana Dolamic[2], Jie Yang[3],
Akansha Bhardwaj[1], Giuseppe Cuccu[1], Philippe Cudré-Mauroux[1]
[1]{ines.arous,akansha.bhardwaj,giuseppe.cuccu,pcm}@unifr.ch,
[2]Ljiljana.Dolamic@armasuisse.ch, [3]j.yang-3@tudelft.nl

1 Model Inference

In this section, we present the proofs of our lemmas. We use the same notational conventions as in the paper.

Lemma 1 (Incremental Document Classification)**.** *The true label distribution $q(z_i)$ can be incrementally computed using the predicted label by the attention-based model θ_i, and the parameters m_j and n_j of the worker reliability distribution r_j.*

$$q(z_i = 1) \propto \left\{ \begin{array}{ll} \theta_i \prod_{j\in\mathcal{J}_i} \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, & \textit{if } \mathbf{A}_{i,j} = 0, \\ \theta_i \prod_{j\in\mathcal{J}_i} \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}, & \textit{if } \mathbf{A}_{i,j} = 1, \end{array} \right. \tag{1}$$

where Ψ is the Digamma function. If $q(z_i = 0)$, then we replace θ_i by $1 - \theta_i$.

Proof. To minimize the KL divergence, we assume the variational distribution follows the same distribution as the latent variable (Tzikas, Likas, and Galatsanos 2008). For $q(z_i)$, we obtain Eq.(2).

$$q(z_i) \propto g_{q(r_j,\alpha_s)}[p(z_i, r, \alpha, \mathbf{A}_{i,*}, \mathbf{B}, \mathcal{W})], \tag{2}$$

where, we use $g_x(\cdot)$ to denote the exponential of expectation term $\exp\{\mathbb{E}_x[\log(\cdot)]\}$ with x being a variational distribution and $x \propto y$ to denote that the two variables x and y are proportionally related (i.e., $x = ky$, where k is a constant). According to the mean field approximation, the probability $p(z_i, r, \alpha, \mathbf{A}_{i,*}, \mathbf{B}, \mathcal{W})$ factorizes over $\mathcal{S}_i$ and $\mathcal{J}_i$ and Eq.(2) can be written as Eq.(3).

$$q(z_i) \propto g_{q(r_j,\alpha_s)}[\prod_{s\in\mathcal{S}_i, j\in\mathcal{J}_i} p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})], \tag{3}$$

where $\mathcal{S}_i$ and $\mathcal{J}_i$ represent respectively the sentences and the workers relevant to document i. Using the properties of the exponential and the logarithm functions in $g_x(\cdot)$, we get Eq.(4):

$$q(z_i) \propto \prod_{s\in\mathcal{S}_i, j\in\mathcal{J}_i} g_{q(r_j,\alpha_s)}[p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})]. \tag{4}$$

By applying the chain rule on $p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})$, we obtain Eq.(5).

$$\begin{aligned} p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W}) = {} & p(\mathbf{A}_{i,j}|z_i, r_j) \times p(\mathbf{B}_{s,j}|r_j, \alpha_s) \\ & \times p(z_i|\mathbf{v}_s, \mathcal{W}) \times p(r_j|m_j, n_j) \\ & \times p(\alpha_s|\mathbf{v}_s, \mathcal{W}_a). \end{aligned} \tag{5}$$

Next we replace the probability in Eq.(4) by the chain rule and keep only the terms that depend on z_i, we get Eq.(6).

$$q(z_i) \propto \prod_{s\in\mathcal{S}_i, j\in\mathcal{J}_i} g_{q(r_j,\alpha_s)}[p(z_i|\mathbf{v}_s, \mathcal{W}) p(\mathbf{A}_{i,j}|z_i, r_j)]. \tag{6}$$

As the probability $p(z_i|\mathbf{v}_s, \mathcal{W})$ is independent from $p(\mathbf{A}_{i,j}|z_i, r_j)$, we get the following:

$$q(z_i) \propto \prod_{s\in\mathcal{S}_i, j\in\mathcal{J}_i} g_{q(r_j,\alpha_s)}[p(z_i|\mathbf{v}_s, \mathcal{W})] g_{q(r_j,\alpha_s)}[p(\mathbf{A}_{i,j}|z_i, r_j)]. \tag{7}$$

Since the probability of z_i does not depend on r_j and α_s, we can simplify Eq.(7) to the following:

$$\begin{aligned} q(z_i) & \propto \prod_{s\in\mathcal{S}_i, j\in\mathcal{J}_i} p(z_i|\mathbf{v}_s, \mathcal{W}) g_{q(r_j,\alpha_s)}[p(\mathbf{A}_{i,j}|z_i, r_j)] \\ & \propto \prod_{s\in\mathcal{S}_i} p(z_i|\mathbf{v}_s, \mathcal{W}) \prod_{j\in\mathcal{J}_i} g_{q(r_j)}[p(\mathbf{A}_{i,j}|z_i, r_j)] \\ & \propto p(z_i|\mathbf{V}, \mathcal{W}) \prod_{j\in\mathcal{J}_i} g_{q(r_j)}[p(\mathbf{A}_{i,j}|z_i, r_j)]. \end{aligned} \tag{8}$$

We show the proof only for $z_i = 1$ since the proof for $z_i = 0$ follow similarly. Using the definition of $q(z_i)$, we have Eq.(9):

$$p(z_i = 1|\mathbf{V}, \mathcal{W}) = \theta_i. \tag{9}$$

We substitute the probability $p(z_i|\mathbf{V}, \mathcal{W})$ and $p(\mathbf{A}_{i,j}|z_i, r_j)$ by their respective definitions in Eq.(9) and Eq.(8) from Section Method:

$$q(z_i = 1) \propto \left\{ \begin{array}{ll} \theta_i \prod_{j\in\mathcal{J}_i} g_{q(r_j)}[1 - r_j], & \text{if } \mathbf{A}_{i,j} = 0, \\ \theta_i \prod_{j\in\mathcal{J}_i} g_{q(r_j)}[r_j], & \text{if } \mathbf{A}_{i,j} = 1. \end{array} \right. \tag{10}$$

By computing the geometric mean of the beta distribution, we can evaluate the exponential terms $g_{q(r_j)}[\cdot]$ as follows:

$$\begin{aligned} g_{q(r_j)}[1 - r_j] & = \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, \\ g_{q(r_j)}[r_j] & = \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}. \end{aligned} \tag{11}$$

Putting (11) into (10), the update equation can be simplified:

$$q(z_i = 1) \propto \\ \begin{cases} \theta_i \prod_{j \in \mathcal{J}_i} \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{A}_{i,j} = 0, \\ \theta_i \prod_{j \in \mathcal{J}_i} \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{A}_{i,j} = 1, \end{cases} \tag{12}$$

which concludes the proof. □

Lemma 2 (Incremental Sentence Importance)**.** *The importance of a sentence for document classification can be incrementally computed using the attributed attention weight by the attention-based model a_s and the parameters m_j and n_j of the worker reliability distribution r_j.*

$$q(\alpha_s = 1) \propto \\ \begin{cases} a_s \prod_{j \in \mathcal{J}_s} \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, & \textit{if } \mathbf{B}_{s,j} = 0, \\ a_s \prod_{j \in \mathcal{J}_s} \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}, & \textit{if } \mathbf{B}_{s,j} = 1. \end{cases} \tag{13}$$

Proof. Similarly to Eq.(2), we assume the variational distribution $q(\alpha_s)$ follows the same distribution as the latent variable α_s. We obtain Eq.(14):

$$q(\alpha_s) \propto g_{q(r_j, z_i)}[p(z, r, \alpha_s, \mathbf{A}, \mathbf{B}_{s,*}, \mathcal{W})]. \tag{14}$$

Using the mean field approximation, the probability $p(z, r, \alpha_s, \mathbf{A}, \mathbf{B}_{s,*}, \mathcal{W})$ factorizes as follows:

$$q(\alpha_s) \propto g_{q(r_j, z_i)}[\prod_{j \in \mathcal{J}_s, i \in \mathcal{I}_s} p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})], \tag{15}$$

where $\mathcal{J}_s$ and $\mathcal{I}_s$ represent the workers and the documents relevant to sentence s. Using the properties of the exponential and logarithm functions, we get:

$$q(\alpha_s) \propto \prod_{j \in \mathcal{J}_s, i \in \mathcal{I}_s} g_{q(r_j, z_i)}[p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})]. \tag{16}$$

By applying the chain rule of Eq.(5) and keeping only the terms that depend on α_s, Eq.(16) is simplified as follows:

$$q(\alpha_s) = \prod_{j \in \mathcal{J}_s} g_{q(r_j)}(p(\alpha_s | \mathbf{v}_s, \mathcal{W}_a) p(\mathbf{B}_{s,j} | \alpha_s, r_j)). \tag{17}$$

Since α_s does not depend on r_j, we get:

$$q(\alpha_s) = p(\alpha_s | \mathbf{v}_s, \mathcal{W}_a) \prod_{j \in \mathcal{J}_s} g_{q(r_j)}[p(\mathbf{B}_{s,j} | \alpha_s, r_j)]. \tag{18}$$

Here as well, we show the proof for $\alpha_s = 1$, as the proof for $\alpha_s = 0$ follows similarly. Using the definition of α_s, we get:

$$p(\alpha_s = 1 | \mathbf{v}_s, \mathcal{W}_a) = a_s. \tag{19}$$

Using the definition of $p(\alpha_s = 1 | \mathbf{v}_s, \mathcal{W}_a)$ in Eq.(19) and the definition of $p(\mathbf{B}_{s,j} | \alpha_s, r_j)$ from Section Method Eq.(7), we get the following:

$$q(\alpha_s = 1) \propto \begin{cases} a_s \prod_{j \in \mathcal{J}_s} g_{q(r_j)}[1 - r_j], & \text{if } \mathbf{B}_{s,j} = 0, \\ a_s \prod_{j \in \mathcal{J}_s} g_{q(r_j)}[r_j], & \text{if } \mathbf{B}_{s,j} = 1. \end{cases} \tag{20}$$

We replace in Eq.(20), $g_{q(r_j)}[1 - r_j]$ and $g_{q(r_j)}[r_j]$ by the expressions given in Eq.(11):

$$q(\alpha_s = 1) \propto \\ \begin{cases} a_s \prod_{j \in \mathcal{J}_s} \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{B}_{s,j} = 0, \\ a_s \prod_{j \in \mathcal{J}_s} \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{B}_{s,j} = 1, \end{cases} \tag{21}$$

which concludes the proof. □

Lemma 3 (Incremental Worker Reliability)**.** *The worker reliability distribution $q(r_j)$ can be incrementally computed using her annotation and rationale quality, and the reliability parameters m_j and n_j from the previous iteration.*

$$q(r_j) \propto \\ \begin{cases} Beta(m'_j + \sum_{s \in \mathcal{S}_j} (1 - a_s), (n'_j + \sum_{s \in \mathcal{S}_j} a_s), \textit{if } \mathbf{B}_{s,j} = 0, \\ Beta(m'_j + \sum_{s \in \mathcal{S}_j} a_s, n'_j + \sum_{s \in \mathcal{S}_j} (1 - a_s)), \textit{if } \mathbf{B}_{s,j} = 1, \end{cases} \tag{22}$$

where $m'_j = m_j + \sum_{i \in \mathcal{I}_j} \theta_i$ *and* $n'_j = n_j + \sum_{i \in \mathcal{I}_j} (1 - \theta_i)$, *if* $\mathbf{A}_{i,j} = 1$ *and* $m'_j = m_j + \sum_{i \in \mathcal{I}_j} (1 - \theta_i)$ *and* $n'_j = n_j + \sum_{i \in \mathcal{I}_j} \theta_i$, *if* $\mathbf{A}_{i,j} = 0$.

Proof. We assume the variational distribution $q(r_j)$ follows the same distribution as the latent variable r_j which translates to Eq.(23).

$$q(r_j) \propto g_{q(z_i, \alpha_s)}[p(z, r_j, \alpha, \mathbf{A}_{*,j}, \mathbf{B}_{*,j}, \mathcal{W})]. \tag{23}$$

Using the mean field approximation, the probability $p(z, r_j, \alpha, \mathbf{A}_{*,j}, \mathbf{B}_{*,j}, \mathcal{W})$ factorizes over $|\mathcal{I}_j|$ and $|\mathcal{S}_j|$ representing respectively the documents and the sentences relevant to worker j.

$$q(r_j) \propto g_{q(z_i, \alpha_s)}[\prod_{i \in \mathcal{I}_j, s \in \mathcal{S}_j} p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})]. \tag{24}$$

Using the properties of the exponential and logarithm, we get:

$$q(r_j) \propto \prod_{i \in \mathcal{I}_j, s \in \mathcal{S}_j} g_{q(z_i, \alpha_s)}[p(z_i, r_j, \alpha_s, \mathbf{A}_{i,j}, \mathbf{B}_{s,j}, \mathcal{W})]. \tag{25}$$

By applying the chain rule in Eq.(5) and keeping only the terms that depend on r_j, we get:

$$q(r_j) \propto \prod_{i \in \mathcal{I}_j, s \in \mathcal{S}_j} g_{q(z_i, \alpha_s)}[p(r_j | m_j, n_j) p(\mathbf{A}_{i,j} | z_i, r_j) p(\mathbf{B}_{s,j} | r_j, \alpha_s)] \tag{26}$$

Since the probability $p(r_j | m_j, n_j)$ does not depend on z_i and α_s, we can simplify Eq.(26) to the following:

$$q(r_j) \propto p(r_j | m_j, n_j) \\ \times \prod_{i \in \mathcal{I}_j, s \in \mathcal{S}_j} g_{q(z_i, \alpha_s)}[p(\mathbf{A}_{i,j} | z_i, r_j) p(\mathbf{B}_{s,j} | r_j, \alpha_s)] \tag{27}$$

By replacing the exponential of expectation term $g_x(\cdot)$ by its expression, we can simplify the term $g_{q(z_i,\alpha_s)}[p(\mathbf{A}_{i,j}|z_i,r_j)p(\mathbf{B}_{s,j}|r_j,\alpha_s)]$ as follows:

$$\begin{aligned}
&g_{q(z_i,\alpha_s)}[p(\mathbf{A}_{i,j}|z_i,r_j)p(\mathbf{B}_{s,j}|r_j,\alpha_s)]\\
&=\exp\{\mathbb{E}_{q(z_i,\alpha_s)}[\log(p(\mathbf{A}_{i,j}|z_i,r_j)p(\mathbf{B}_{s,j}|r_j,\alpha_s))]\}\\
&=\exp\{\mathbb{E}_{q(z_i,\alpha_s)}[\log(p(\mathbf{A}_{i,j}|z_i,r_j))+\log(p(\mathbf{B}_{s,j}|r_j,\alpha_s))]\}\\
&=\exp\{\mathbb{E}_{q(z_i)}[\log(p(\mathbf{A}_{i,j}|z_i,r_j))]+\mathbb{E}_{q(\alpha_s)}[\log(p(\mathbf{B}_{s,j}|r_j,\alpha_s))]\}\\
&=g_{q(z_i)}[p(\mathbf{A}_{i,j}|z_i,r_j))]\times g_{q(\alpha_s)}[p(\mathbf{B}_{s,j}|r_j,\alpha_s)] \qquad (28)
\end{aligned}$$

We distinguish two main cases depending on the values of $\mathbf{A}_{i,j}\in\{0,1\}$. Let's start with the case where $\mathbf{A}_{i,j}=1$. This case covers all documents i a worker j has annotated as positive. In such a case, the probability $p(\mathbf{A}_{i,j}|z_i,r_j)$ can be written as a function of θ_i and r_j as given by eq.(29):

$$g_{q(z_i)}[p(\mathbf{A}_{i,j}=1|z_i,r_j)=r_j^{\theta_i}(1-r_j)^{(1-\theta_i)} \qquad (29)$$

The documents annotated as positive by worker j include two sets of sentences: a set of sentences annotated as rationales, where $\mathbf{B}_{s,j}=1$, and a set of non-rationales where $\mathbf{B}_{s,j}=0$. The probability $p(\mathbf{B}_{s,j}|r_j,\alpha_s)$ is simplified as follows:

$$g_{q(\alpha_s)}[p(\mathbf{B}_{s,j}|r_j,\alpha_s)]\propto\begin{cases}r_j^{a_s}(1-r_j)^{(1-a_s)}, \text{if } \mathbf{B}_{s,j}=1,\\ r_j^{(1-a_s)}(1-r_j)^{a_s}, \text{if } \mathbf{B}_{s,j}=0,\end{cases} \qquad (30)$$

If we take the case of $\mathbf{B}_{s,j}=1$, by putting (29) and (30) in (26), we get:

$$q(r_j)\propto p(r_j|m_j,n_j)\prod_{i\in\mathcal{I}_j,s\in\mathcal{S}_j}r_j^{\theta_i+a_s}(1-r_j)^{(1-\theta_i+1-a_s)} \qquad (31)$$

The probability $p(r_j|m_j,n_j)$ is a beta distribution and hence its probability density function is given by Eq.(32).

$$\begin{aligned}p(r_j|m_j,n_j)&=Beta(m_j,n_j)\\&=r_j^{(m_j-1)}(1-r_j)^{(n_j-1)}\end{aligned} \qquad (32)$$

By substituting the probability $p(r_j|m_j,n_j)$ by its expression in Eq.(32), we get the following result:

$$\begin{aligned}q(r_j)\propto& r_j^{(m_j-1)}(1-r_j)^{(n_j-1)}\\&\times\prod_{i\in\mathcal{I}_j,s\in\mathcal{S}_j}r_j^{\theta_i+a_s}(1-r_j)^{(1-\theta_i+1-a_s)}\\&\propto r_j^{m_j-1+\sum_{i\in\mathcal{I}_j}\theta_i+\sum_{s\in\mathcal{S}_j}a_s}\\&\times(1-r_j)^{(n_j-1+\sum_{i\in\mathcal{I}_j}(1-\theta_i)+\sum_{s\in\mathcal{S}_j}(1-a_s))}\end{aligned} \qquad (33)$$

Similarly for $\mathbf{B}_{s,j}=0$, by putting (29), (30) and (32) in (27), we get:

$$\begin{aligned}q(r_j)\propto& r_j^{(m_j-1)}(1-r_j)^{(n_j-1)}\\&\times\prod_{i\in\mathcal{I}_j,s\in\mathcal{S}_j}r_j^{\theta_i+1-a_s}(1-r_j)^{(1-\theta_i+a_s)}\\&\propto r_j^{m_j-1+\sum_{i\in\mathcal{I}_j}\theta_i+\sum_{s\in\mathcal{S}_j}(1-a_s)}\\&\times(1-r_j)^{(n_j-1+\sum_{i\in\mathcal{I}_j}(1-\theta_i)+\sum_{s\in\mathcal{S}_j}a_s)}\end{aligned} \qquad (34)$$

Method	Implementation
MILNET	github.com/stangelid/oposum
fastText	fasttext.cc/docs/en/supervised-tutorial.html
SciBERT	github.com/allenai/scibert
ALBERT	huggingface.co/transformers/model_doc/albert.html
LSTM	github.com/akashkm99/Interpretable-Attention
InvRAT	github.com/code-terminator/invariant_rationalization
RA-CNN	github.com/yezhang-xiaofan/Rationale-CNN

Table 1: Methods Implementation

Using Eq.(31) and Eq.(34), we get the updating rules for when the documents are labeled as positive by a worker, i.e., $\mathbf{A}_{i,j}=1$:

$$q(r_j)\propto\begin{cases}r_j^{m_j-1+\sum_i\theta_i+\sum_s a_s}(1-r_j)^{(n_j-1+\sum_i(1-\theta_i)+\sum_s(1-a_s))}, \text{if } \mathbf{A}_{i,j}{=}\mathbf{B}_{s,j},\\ r_j^{(m_j-1+\sum_i\theta_i+\sum_s(1-a_s))}(1-r_j)^{(n_j-1+\sum_i(1-\theta_i)+\sum_s a_s)}, \text{if } \mathbf{A}_{i,j}{\neq}\mathbf{B}_{s,j},\end{cases} \qquad (35)$$

where the documents i and sentences s are relevant to worker j, i.e., $i\in\mathcal{I}_j$ and $s\in\mathcal{S}_j$. For the second case, where $\mathbf{A}_{i,j}=0$, the term $g_{q(z_i)}[p(\mathbf{A}_{i,j}|z_i,r_j)$ can be written as a function of θ_i and r_j as given by eq.(36):

$$g_{q(z_i)}[p(\mathbf{A}_{i,j}=0|z_i,r_j)]=r_j^{(1-\theta_i)}(1-r_j)^{\theta_i} \qquad (36)$$

Using the same reasoning of distinguishing the two cases $\mathbf{B}_{s,j}=0$ and $\mathbf{B}_{s,j}=1$ and then replacing the probability $p(r_j|m_j,n_j)$ by its probability density function, we get the following results:

$$q(r_j)\propto\begin{cases}r_j^{(m_j-1+\sum_i(1-\theta_i)+\sum_s(1-a_s))}(1-r_j)^{(n_j-1+\sum_i\theta_i+\sum_s a_s)}, \text{if } \mathbf{A}_{i,j}{=}\mathbf{B}_{s,j},\\ r_j^{m_j-1+\sum_i(1-\theta_i)+\sum_s a_s}(1-r_j)^{(n_j-1+\sum_i\theta_i+\sum_s(1-a_s))}, \text{if } \mathbf{A}_{i,j}{\neq}\mathbf{B}_{s,j},\end{cases} \qquad (37)$$

which concludes the proof. □

2 Implementation Details

Comparison Methods

In our experiments, we compare with text classification and rationale-aware methods. We use the authors' implementation for all methods except MILNET, for which we reimplemented the original code in Python. Table 1 summarizes all methods implementations used in our experiments. For each method, we tune the hyperparameters including the learning rate in {0.00001,0.0001,0.001,0.01,0.1,1}, the batch size in {10,20,50,100} and the number of epochs in {10,20,50,100,500}. For fastText, we also vary the word n-grams in $[1,5]$. For InvRat, we vary the embedding dimension in {50,100,200,300}. We use the hyperparameters that led to the optimal results on the validation set. We report the optimal settings in Table 2.

Parameter Settings for MARTA

The parameters of our framework are empirically set. We search for the best architecture for our attention-based model by applying a grid search in {10, 20, 50, 100} for the batch size and in {0.00001, 0.0001, 0.001, 0.01, 0.1, 1} for the

Method	**Wiki-Tech**				**Amazon**			
	#epochs	lr	Batch Size	Other	#epochs	lr	Batch Size	Other
MILNET	25	1e-3	50	-	20	1e-3	50	-
fastText	20	1	-	N-gram = 5	20	9e-1	-	N-gram = 5
SciBERT	75	1e-4	4	-	250	1e-7	8	-
ALBERT	10	1e-4	8	-	10	1e-4	8	-
LSTM-ortho	8	-	32	-	8	-	32	-
LSTM-diversity	8	-	32	diversity weight=0.5	8	-	32	diversity weight=0.5
InvRAT	20	1e-5	20	embedding dim=300	20	1e-4	20	embedding dim=300
RA-CNN	15	-	50	-	20	-	50	-
MARTA	20	1e-3	50	$m_j = 2.5, n_j = 2$	20	1e-3	50	$m_j = 2.5, n_j = 2$

Table 2: Choice of hyperparameters for baseline methods on the *Wiki-Tech* and *Amazon* datasets. We use '-' to denote if the hyperparameter is not applicable. 'lr' denotes the learning rate and the value reported for m_j and n_j in MARTA, is the value used for initialization.

learning rate. We also test different optimization methods including stochastic gradient descent, ADAM and RMSprop. We initialize the priors m_j and n_j by sampling from a uniform distribution $[0, 10]$ and update them in the E-step according to Lemma 3. The optimal parameter settings we found through the validation set are reported in Table 2.

Hardware and Software

For our framework, we used a Ubuntu 16.04 machine with 32 CPUs and 128GB RAM. For experiments that required GPU, we used a Ubuntu 18.04 with 9 GPUs (1 TITAN V and 8 GeForce RTX 2080 Ti), 64 CPUs and 395GB RAM. In our code repository, we provide a file (requirement.txt) that specifies the versions of all required libraries; this file can be used to install them automatically.

3 Limitations and Future Work

The main limitation of our technique consists in two assumptions. The first one assumes that a rationale given by a worker can be directly extracted from text. While this assumption is valid in our context, there are applications where the rationale is expressed in a syntax different from the original text. For example, in McDonnell et al. (2016), the rationale is expressed as reasoning. We also assume that a document is composed of many sentences, and that the sentences have different levels of importance. In case the document is short (1-3 sentences), the importance of the sentences does not vary a lot, and hence the attention scores assigned by our framework are almost all equal. In future work, we plan to represent the workers' rationales by embedding and leverage the similarity between the rationale and the original text to derive the sentence's importance. Using embedding to represent rationales would allow us to capture the worker reliability on different topics. We also plan to leverage the rationales generated through our framework to learn other domains' rationale through transfer learning.

References

McDonnell, T.; Lease, M.; Kutlu, M.; and Elsayed, T. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 139–148. AAAI Press.

Tzikas, D. G.; Likas, A. C.; and Galatsanos, N. P. 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine* 25(6): 131–146.