MARTA: Leveraging Human Rationales for Explainable Text Classification

Ines Arous[1], Ljiljana Dolamic[2], Jie Yang[3],
Akansha Bhardwaj[1], Giuseppe Cuccu[1], Philippe Cudré-Mauroux[1]

[1]University of Fribourg–Switzerland, [2]armasuisse–Switzerland, [3]Delft University of Technology–Netherlands
[1]{ines.arous,akansha.bhardwaj,giuseppe.cuccu,pcm}@unifr.ch,
[2]Ljiljana.Dolamic@armasuisse.ch, [3]j.yang-3@tudelft.nl

Abstract

Explainability is a key requirement for text classification in many application domains ranging from sentiment analysis to medical diagnosis or legal reviews. Existing methods often rely on "attention" mechanisms for explaining classification results by estimating the relative importance of input units. However, recent studies have shown that such mechanisms tend to mis-identify irrelevant input units in their explanation. In this work, we propose a hybrid human-AI approach that incorporates human rationales into attention-based text classification models to improve the explainablility of classification results. Specifically, we ask workers to provide rationales for their annotation by selecting relevant pieces of text. We introduce MARTA, a Bayesian framework that jointly learns an attention-based model and the reliability of workers while injecting human rationales into model training. We derive a principled optimization algorithm based on variational inference with efficient updating rules for learning MARTA parameters. Extensive validation on real-world datasets shows that our framework significantly improves the state of the art both in terms of classification explainability and accuracy.

1 Introduction

Text classification is a fundamental task in natural language processing (NLP) (Zhang, Marshall, and Wallace 2016; Yang et al. 2016; Arras et al. 2017). State-of-the-art methods are dominated by neural network models, which are generally considered as "black boxes" by end-users. The opaqueness of those models has become a major obstacle for their development, deployment, and improvement, particularly in critical tasks such as medical diagnosis (Lakkaraju, Bach, and Leskovec 2016) and legal document review (Chhatwal et al. 2018; Mahoney et al. 2019). Explainable text classification has, therefore, emerged as an important topic, where the goal is to present end-users with human-readable descriptions of the classification rationale (Ribeiro, Singh, and Guestrin 2016; Sundararajan, Taly, and Yan 2017; Camburu et al. 2018; Liu, Yin, and Wang 2019).

Among existing explainability methods, a popular approach is the *attention* mechanism, which identifies important parts of the input for the prediction task by providing a distribution over attended-to input units (e.g., tokens or sentences) (Xu et al. 2015; Bahdanau, Cho, and Bengio 2014).

Attention-based models have resulted in impressive performance across many NLP tasks including text classification, question-answering, and entity recognition (Bahdanau, Cho, and Bengio 2014; Parikh et al. 2016; Wang et al. 2017); in particular, the self-attention mechanism that underlies the Transformer architecture (Vaswani et al. 2017; Devlin et al. 2018) has been playing a central role in many NLP systems. Despite that, recent studies have shown that the learned attention weights are often uncorrelated with the importance of input components measured by other explainability methods (e.g., gradients (Simonyan, Vedaldi, and Zisserman 2013)), and that one can identify different attention distributions that nonetheless yield equivalent predictions (Jain and Wallace 2019; Wiegreffe and Pinter 2019).

A promising approach to enhance the explainability of attention-based models is integrating human rationales as extra supervision information for attention learning. Prior research (Zaidan, Eisner, and Piatko 2007; Zhang, Marshall, and Wallace 2016) has shown that human rationales represent valuable input for improving model performance and for identifying explainable input features in model prediction (Bahdanau, Cho, and Bengio 2014; Mohankumar et al. 2020). Coincidentally, recent studies in human computation (McDonnell et al. 2016) have demonstrated that asking workers to provide annotation rationales – by highlighting supporting text excerpts from the given text – brings no extra annotation efforts. Human rationales are, therefore, easy-to-obtain information with great potential in improving model explainability and performance. Existing work (Bahdanau, Cho, and Bengio 2014; Mohankumar et al. 2020), however, takes human rationales as gold information that is entirely trustworthy, which is typically not the case in practice; indeed, studies from human computation have found the reliability of human-contributed rationales to be a key problem that requires careful treatment (Zaidan, Eisner, and Piatko 2007; Ramírez et al. 2019).

In this work, we tackle the problem of rationale reliability by introducing a human-AI computational approach that integrates human rationales into an attention-based model while weighing individual reliability. We crowdsource the task of annotating documents and ask workers to justify their labels using text excerpts from the document. We introduce MARTA, a Bayesian framework that jointly learns the workers' reliability and the attention-based model pa-

rameters while MApping human Rationales To Attention. The model parameters and worker reliability are updated in an iterative manner, allowing their learning processes to benefit from each other until agreements on both the label and rationales are reached. We formalize such a learning process with a principled optimization algorithm based on variational expectation-maximization. In particular, we derive efficient updating rules that allow both model parameters and worker reliability to be updated incrementally at each iteration. In summary, we make the following contributions:

- We propose MARTA, a Bayesian framework for explainable text classification that integrates human rationales into attention-based models.
- We derive an efficient learning algorithm based on variational inference with incremental updating rules for MARTA parameter estimation.
- We conduct an extensive evaluation on two real-world datasets and show that MARTA substantially outperforms the state of the art by 5.76% F1-score while offering a human-understandable explanation.

2 Related Work

Explainable Text Classification

Driven by the need for transparency, machine learning explainability has drawn significant attention recently (Ribeiro, Singh, and Guestrin 2016; Doshi-Velez and Kim 2017). Existing explainability methods fall into two broad categories: post-hoc explainability and intrinsic explainability. Post-hoc explainability aims at providing explanations for an existing model. A representative method is LIME (Ribeiro, Singh, and Guestrin 2016), which approximates model decisions with an explainable model (e.g., a linear model) in the local area of the feature space. A recent development of this topic is GEF (Liu, Yin, and Wang 2019), which is designed to explain a generic encoder-predictor architecture by jointly generating explanations and classification results. Another class of methods identifies important features by calculating the gradient of an output with respect to an input feature to derive the contribution of the various features (Simonyan, Vedaldi, and Zisserman 2013; Ross, Hughes, and Doshi-Velez 2017; Ancona et al. 2018). Intrinsic explainability aims at constructing self-explanatory models. This can be achieved by adding explainability constraints in model learning to enforce feature sparsity (Freitas 2014), representation disentanglement (Zhang, Nian Wu, and Zhu 2018), or sensitivity towards input features (Sundararajan, Taly, and Yan 2017). Our work falls into this second category by injecting human rationales into model learning through a unified Bayesian framework.

To explain individual predictions, a more popular approach is attention mechanisms, which identify parts of the input that are attended by the model for specific predictions (Xu et al. 2015; Bahdanau, Cho, and Bengio 2014). These attention mechanisms have been playing an important role in NLP not only for explainability but also for the enhancement they bring to model performance (Parikh et al. 2016; Vaswani et al. 2017; Devlin et al. 2018). Their effectiveness in explainability, however, has recently been questioned by an empirical study, which points to the facts that attention distributions are inconsistent with the importance of input units as measured by gradient-based methods and that adversarial distributions can be found yielding similar model performance (Jain and Wallace 2019). Those findings have triggered heated discussions, e.g., it has been shown that attention mechanisms attribute higher weights to important input units for a given task even when the model architecture for prediction changes (Wiegreffe and Pinter 2019). Our work contributes to the discussion by showing that human rationales, when properly injected into the attention-based models, can enhance the model explainability and performance.

Human Rationale in Machine Learning

The idea of incorporating human rationales for model improvement can be traced back to Zaidan et al. (2007), where a human teacher highlights pieces of text in a document as a rationale to justify label annotation. The rationale is fused into the loss function of an SVM classifier by constraining the prediction labels. Similar ideas have been explored for neural network models (Zhang, Marshall, and Wallace 2016) and through different ways of human rationale integration, e.g., by learning a mapping between human rationales and machine attention (Bao et al. 2018) or ensuring the diversity among the hidden representations learned at different time steps (Mohankumar et al. 2020). The idea of finding a small subset of input units capable of generating the same output has resulted in various selective rationalization techniques (Lei, Barzilay, and Jaakkola 2016; Li, Monroe, and Jurafsky 2016; Chen et al. 2018; Chang et al. 2019; Yu et al. 2019). Despite all existing efforts, few studies have addressed the potential issues in human involvement, such as controlling the quality of rationales contributed by humans with varying levels of expertise and motivation. Unlike them, our framework offers a principled method to model human reliability in integrating human rationales.

A separated line of research in human computation and crowdsourcing has investigated the task design for soliciting human rationales in crowdsourcing settings. When gathering relevance judgments for search results, McDonnell et al. (2016) found out that by asking crowd workers to provide 2-3 sentences of document excerpts for justification, annotation quality can be largely enhanced without the task completion time being increased. However, it is also known that the quality of human-contributed rationales remains a challenging issue, especially for subjective and complex tasks (Zaidan, Eisner, and Piatko 2007; Ramírez et al. 2019). Aligned with these works, our work offers a computational approach that integrates human rationales for explainable text classification while addressing the reliability issue of rationales through a principled learning algorithm.

3 Method

MARTA is a unified Bayesian Framework that integrates an attention-based model with labels and rationales contributed by workers. In this section, we first formally define our problem, and then introduce our framework, followed by a presentation of our algorithm for learning MARTA parameters.

Problem Formulation

Notations. We use boldface lowercase letters to denote vectors and boldface uppercase letters to denote matrices. For an arbitrary matrix $\mathbf{M}$, we use $\mathbf{M}_{i,j}$ to denote the entry at the i-th row and j-th column. We denote the set of documents as $\mathcal{I}$, the set of sentences composing all documents as $\mathcal{S}$, and the set of sentences belonging to document i as $\mathcal{S}_i$. We denote the set of workers who provide noisy labels with rationales as $\mathcal{J}$. The subset of workers who label document i is denoted as $\mathcal{J}_i$. We consider binary classification and use $\mathbf{A}_{i,j} = 1$ to denote that a document i is classified as positive by worker j, and $\mathbf{A}_{i,j} = 0$ otherwise. We use $\mathbf{B}_{s,j} = 1$ to denote that a sentence s is selected as a rationale by worker j. The subset of workers who select the sentence s as a rationale for their annotations is denoted as $\mathcal{J}_s$.

Problem Definition. Let $\mathcal{I}$ be a set of documents, each assigned to a unique binary label representing its relevance to a topic. Each document $i \in \mathcal{I}$ is composed of a set of sentences $\mathcal{S}_i$ that can be used as rationale in determining the relevance of a document to the topic. Let $\mathcal{J}$ be a set of workers who annotate the documents with labels $\mathbf{A}$ and rationales $\mathbf{B}$. Our goal is to infer the true label of a document denoted as z_i while estimating the importance of each sentence $s \in \mathcal{S}_i$, denoted as α_s, in the inference.

The MARTA Framework

MARTA is a probabilistic framework that models the process of worker-provided labels (i.e., $\mathbf{A}$) and rationales (i.e., $\mathbf{B}$), conditioned on the true labels (i.e., z), the importance of rationales (i.e., α), and the reliability of workers (i.e., r). The overall framework is depicted as a graphical model in Figure 1. In the following, we first describe how an attention-based model is embedded into MARTA to allow the integration of human rationales, and then describe the process of worker-provided labels and answers for their integration.

Rationale-Aware Attention Model. Given the true label of a document $z_i \in \{0, 1\}$ as a binary variable, we model it with a Bernoulli distribution. The underlying intuition of our rationale-aware attention model is that the label of a document is determined by its sentences, and that each sentence contributes differently in determining the overall label of the document. Formally, we have:

$$z_i \sim Ber(\theta_i), \theta_i = \sum_{s \in \mathcal{S}_i} a_s P_s, \tag{1}$$

where θ_i is the parameter of the distribution, modeled as the weighted sum of the sentence-level label P_s with attention weight a_s. The sentence-level label P_s is predicted from the contents of the sentence through a neural network of arbitrary architecture:

$$P_s = \text{softmax}(f^{W_p, b_p}(\mathbf{v}_s)), \tag{2}$$

where $\mathbf{v}_s$ is the embedding vector of the sentence s, and $f^{W_p, b_p}(\mathbf{v}_s)$ models the output of the network layers preceding the softmax layer, parameterized by W_p and b_p and shared across all sentences.

To model the attention weight a_s for each sentence, we use a Bidirectional LSTM (BiLSTM) (Schuster and Paliwal 1997) to account for the sequential dependencies among sentences. Specifically, each sentence vector is transformed into a hidden vector h_s through BiLSTM:

$$h_s = \text{BiLSTM}(\mathbf{v}_s). \tag{3}$$

Then, the attention weight of a sentence is modeled through a fully-connected layer and a softmax normalization:

$$a_s = \text{softmax}(h'_s), h'_s = \tanh(W_a h_s + b_a). \tag{4}$$

Finally, we model if a sentence can be viewed as a rationale for the document label as a binary variable $\alpha_s \in \{0, 1\}$ that follows a Bernoulli distribution, parameterized by a_s:

$$\alpha_s \sim Ber(a_s). \tag{5}$$

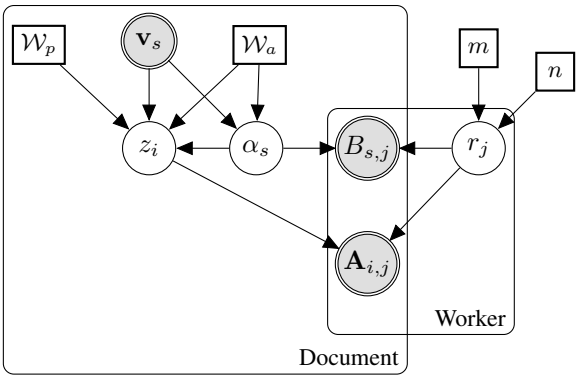


Figure 1: Graphical representation of MARTA. Double (greyed) circles represent observed variables, while single circles represent latent variables. Squares represent model parameters. Edges represent conditional relationships in text classification. On the left side, an attention-based model parameterized by $\{\mathcal{W}_a, \mathcal{W}_p\}$ predicts the label z_i for a document. Each document is composed of sentences $\mathbf{v}_s$, with an importance α_s in the classification. On the right side, a worker is represented with her reliability distribution r_j with parameters m and n. The work annotates a document with label $\mathbf{A}_{i,j}$ and rational $\mathbf{B}_{s,j}$.

Integrating Labeling Rationales. We represent worker reliability by $r_j \in [0, 1]$ where $r_j = 1$ indicates that the worker is fully reliable and $r_j = 0$ otherwise. In practice, we would like to measure our confidence for an estimate r_j as dependent on the number of answers of worker j, i.e., the more annotations a worker provides, the more confident we would like to be about her reliability estimate r_j. To quantify the confidence of our estimates, we adopt a Bayesian treatment of r_j by modeling it with a Beta distribution:

$$r_j \sim Beta(m, n), \tag{6}$$

where m and n are the parameters of the distribution.

We use the reliability of a worker to define the likelihood of her rationale being a true support of the document label:

$$p(\mathbf{B}_{s,j} | \alpha_s, r_j) = r_j^{\mathbb{1}[\alpha_s = \mathbf{B}_{s,j}]} (1 - r_j)^{\mathbb{1}[\alpha_s \neq \mathbf{B}_{s,j}]}, \tag{7}$$

where $\mathbb{1}[\cdot]$ is an indicator function returning 1 if the statement is true and 0 otherwise.

Similarly, we use the reliability of a worker to define the likelihood of her provided label being the true label:

$$p(\mathbf{A}_{i,j} | z_i, r_j) = r_j^{\mathbb{1}[z_i = \mathbf{A}_{i,j}]} (1 - r_j)^{\mathbb{1}[z_i \neq \mathbf{A}_{i,j}]}. \tag{8}$$

Variational Inference

Learning the parameters of MARTA resorts to maximizing the following likelihood function:

$$p(\mathbf{A}, \mathbf{B}) = \int p(\mathbf{A}, \mathbf{B}, \mathbf{z}, \mathbf{r}, \alpha, |\mathcal{W}, \mathbf{V}) d\mathbf{z}, \mathbf{r}, \alpha, \tag{9}$$

where $\mathbf{z}$, $\mathbf{r}$ and α are latent variables, $\mathcal{W}$ represents the set of parameters of the model, i.e. $\mathcal{W} = \{\mathcal{W}_a, \mathcal{W}_p\}$, and $\mathbf{V}$ is the embedding of all the sentences composing the documents. Since Eq.(9) contains more than one latent variable, it is computationally infeasible to optimize (Tzikas, Likas, and Galatsanos 2008). Therefore, we consider the log of our likelihood function, i.e.,

$$\begin{aligned}
&\log\left(p(\mathbf{A}, \mathbf{B})\right) \\
&= \underbrace{\int q(\mathbf{z}, \mathbf{r}, \alpha) \log \frac{p(\mathbf{A}, \mathbf{B}, \mathbf{z}, \mathbf{r}, \alpha | \mathcal{W}, \mathbf{V})}{q(\mathbf{z}, \mathbf{r}, \alpha)} d\mathbf{z}, \mathbf{r}, \alpha}_{\mathcal{L}(\mathcal{W}, q)} \\
&+ \underbrace{\int q(\mathbf{z}, \mathbf{r}, \alpha) \log \frac{q(\mathbf{z}, \mathbf{r}, \alpha)}{p(\mathbf{z}, \mathbf{r}, \alpha | \mathbf{A}, \mathbf{B}, \mathcal{W}, \mathbf{V})} d\mathbf{z}, \mathbf{r}, \alpha,}_{KL(q||p)}
\end{aligned} \tag{10}$$

where $q(\mathbf{z}, \mathbf{r}, \alpha)$ is any probability density function and $KL(\cdot)$ is the KL divergence between two distributions. By doing so, the two parts of the objective function can then be optimized iteratively with a variational expectation-maximization method (Tzikas, Likas, and Galatsanos 2008). Specifically, we iterate between two steps: 1) the E-step, where we approximate the latent variables $p(\mathbf{z}, \mathbf{r}, \alpha | \mathbf{A}, \mathbf{B}, \mathcal{W}, \mathbf{V})$ with the variational distribution $q(\mathbf{z}, \mathbf{r}, \alpha)$, by minimizing the KL-divergence. 2) the M-step, where we maximize the term $\mathcal{L}(\mathcal{W}, q)$ given the newly inferred latent variables.

E step. We use the mean field variational inference approach (Blei, Kucukelbir, and McAuliffe 2017) by assuming that $q(\mathbf{z}, \mathbf{r}, \alpha)$ factorizes over the latent variables.

$$q(\mathbf{z}, \mathbf{r}, \alpha) = \prod_i q(z_i) \prod_s q(\alpha_s) \prod_j q(r_j). \tag{11}$$

To minimize the KL divergence, we choose following forms for the factor functions:

$$q(z_i)=Ber(\theta_i); q(\alpha_s)=Ber(a_s); q(r_j)=Beta(m_j, n_j), \tag{12}$$

where θ_i, a_s, m_j and n_j are variational parameters used to minimize the KL divergence. The latter can be minimized by updating one latent variable at a time and keeping all others fixed.

In the following, we derive the updating rules for $q(z_i)$, $q(\alpha_s)$ and $q(r_j)$. To do so, we simplify Eq.(10) and obtain for each latent variable the following inference equations:

$$\begin{aligned}
q(z_i) &= \prod_{s \in \mathcal{S}_i, j \in \mathcal{J}_i} p(z_i | \mathbf{v}_s, \mathcal{W}) g_{q(r_j)}(p(\mathbf{A}_{i,j} | z_i, r_j)), \\
q(\alpha_s) &= \prod_{j \in \mathcal{J}_s} p(\alpha_s | \mathbf{v}_s, \mathcal{W}_a) g_{q(r_j)}(p(\mathbf{B}_{s,j} | \alpha_s, r_j)), \\
q(r_j) &= \prod_{i \in \mathcal{I}_j, s \in \mathcal{S}_j} p(r_j) g_{q(z_i, \alpha_s)}(p(\mathbf{A}_{i,j} | z_i, r_j) p(\mathbf{B}_{s,j} | \alpha_s, r_j)),
\end{aligned} \tag{13}$$

where $\mathcal{S}_i$ are the sentences in a document i. $\mathcal{J}_i$ and $\mathcal{J}_s$ are the workers annotating document i and those choosing sentence s as a rationale, respectively. $\mathcal{I}_j$ and $\mathcal{S}_j$ are the documents annotated by worker j and the sentences chosen by her as rationales, respectively. We use $g_x(\cdot)$ to denote the exponential of expectation term $\exp\{\mathbb{E}_x[\log(\cdot)]\}$ with x being a variational distribution. With the above equations, we obtain the updating rules for all latent variables. We first give the updating rules of the document label z_i and the sentence's importance α_s by the following lemmas.

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.*

$$q(z_i = 1) \propto \begin{cases} \theta_i \prod_{j \in \mathcal{J}_i} \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{A}_{i,j} = 0, \\ \theta_i \prod_{j \in \mathcal{J}_i} \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{A}_{i,j} = 1, \end{cases} \tag{14}$$

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

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.*

$$q(\alpha_s = 1) \propto \begin{cases} a_s \prod_{j \in \mathcal{J}_s} \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{B}_{s,j} = 0, \\ a_s \prod_{j \in \mathcal{J}_s} \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}, & \text{if } \mathbf{B}_{s,j} = 1. \end{cases} \tag{15}$$

Next, we show the updating rule for the worker reliability $q(r_j)$ with the following lemma.

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.*

$$q(r_j) \propto \begin{cases} Beta(m'_j + \sum_{s \in \mathcal{S}_j}(1 - a_s), n'_j + \sum_{s \in \mathcal{S}_j} a_s), \text{if } \mathbf{B}_{s,j} = 0, \\ Beta(m'_j + \sum_{s \in \mathcal{S}_j} a_s, n'_j + \sum_{s \in \mathcal{S}_j}(1 - a_s)), \text{if } \mathbf{B}_{s,j} = 1, \end{cases} \tag{16}$$

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 $\mathbf{A}_{i,j} = 0$.

Due to page constraints, we provide proofs for all lemmas in the appendix (Arous et al. 2020).

M step. Given the true labels of the documents, the importance of sentences, and the worker reliability inferred by the E-step, the M-step maximizes the first term of Eq.(10) to

Algorithm 1: Learning MARTA Parameters

Input : $\mathbf{A}, \mathbf{B}, \mathcal{S}_i$ $(\forall i \in \mathcal{I})$
Output : Variational distributions: $q(z_i)$, $q(\alpha_s)$ and $q(r_j)$
Initialize: MARTA parameters: θ_i, m_j, n_j, $\mathcal{W}$

1 **while** *Eq. (10) has not converged* **do**
2 **for** $i \in \mathcal{I}$ **do**
3 update $q(z_i)$ using Lemma 1;
4 update $q(\alpha_s)$ using Lemma 2;
5 **for** $j \in \mathcal{J}$ **do**
6 update $q(r_j)$ using Lemma 3;
7 **for** $i \in \mathcal{I}$ **do**
8 Update $\mathcal{W}$;

learn the parameters $\mathcal{W} = \{\mathcal{W}_a, \mathcal{W}_p\}$.

$$\begin{aligned}\mathcal{L}(\mathcal{W}, q) &= \int q(z, \alpha, r) \log\left[p(\mathbf{A}, \mathbf{B}, z, \alpha, r|\mathbf{V}, \mathcal{W})\right] dr + C_1 \\ &= \underbrace{\sum_{z_i} q(z_i) \log[p(z_i|\mathbf{V}; \mathcal{W}_a, \mathcal{W}_p)]}_{\mathcal{T}_1} \\ &+ \underbrace{\sum_{\alpha_s} q(\alpha_s) \log[p(\alpha_s|\mathbf{v}_s; \mathcal{W}_a)]}_{\mathcal{T}_2} + C_1 + C_2, \quad (17)\end{aligned}$$

where $C_1 = \exp\{\mathbb{E}_{q(z,\alpha,r)}[\log((\frac{1}{q(z,\alpha,r)})]$ is a constant and C_2 are the terms that do not depend on the parameters $\mathcal{W}$. The term $\mathcal{T}_1$ is equivalent to the inverse of the cross entropy between the target labels of a document $q(z_i)$ and the predicted label $p(z_i|\mathbf{V}; \mathcal{W}_a, \mathcal{W}_p)$. Similarly, the term $\mathcal{T}_2$ is equivalent to the inverse of the cross entropy between the indication of a sentence as a rationale and the predicted importance $p(\alpha_s|\mathbf{v}_s; \mathcal{W}_a)$. Given the shared parameter $\mathcal{W}_a$, we minimize the prediction loss $\mathcal{T}_1$ together with the loss $\mathcal{T}_2$.

Algorithm

The overall optimization algorithm is given in Algorithm 1. We initialize MARTA's parameters and iterate between an E step (rows 2-6) and an M-step (rows 7-8). The E-step consists of updating the variational distribution of the document labels, the sentence importance and the worker reliability. Our framework is semi-supervised in the sense that when ground truth labels are available, we fix them in the E-step. The M steps consists in updating the parameters of the attention-based model by jointly learning the document labels and the sentence's importance. It is worth noting that for this step, the loss between human rationales and the attention generated by the model is minimized. The convergence is reached when the documents label $q(z_i)$ and the sentences relevance $q(\alpha_s)$ are no longer modified by the workers' reliability and the model's parameters stabilize.

The iterations through the documents (rows 2-4) yields a time complexity of $|\mathcal{I}|$ while the iterations through the workers (rows 5-6) yields a time complexity of $|\mathcal{J}|$. The overall complexity of our algorithm is $O(\#iter \times (|\mathcal{I}| + |\mathcal{J}| + \mathcal{C}_W))$, where $\#iter$ is the number of iterations needed to converge and $\mathcal{C}_W$ is the complexity of learning the parameters $\mathcal{W} = \{\mathcal{W}_a, \mathcal{W}_p\}$ of the attention-based model.

Dataset	#Docs	%Positive	#Judgments	#Workers
Wiki-Tech	1413	17.26%	4488	58
Amazon	400	50%	6744	449

Table 1: Datasets Description

4 Experiments

Experimental Setup

Datasets. We use two datasets for our experiments: *Wiki-Tech* and *Amazon*[1]. *Wiki-Tech* contains 1413 Wikipedia articles with expert annotations on their relevance with respect to the topic "technologies commonly used by companies". We crowdsourced this dataset to collect worker rationales. *Amazon* is developed and published by Ramírez et al. (2019). It contains 400 reviews with ground truth labels about "reviews written about books"; this dataset is released with worker's rationales. Key statistics about both datasets are reported in Table 1.

Crowdsourcing Task. Worker annotations in *Wiki-Tech* were collected through a crowdsourcing task that we published on Amazon Mechanical Turk[2]. We asked workers the following predicate: *Does the Wikipedia article describe a technology commonly used by companies?*. We chose workers with a HIT approval rate above 70%. The task started by explaining the concept of "Technology" and provided a positive and a negative example. Then, workers were asked to annotate the article and provide a snippet from the text as a justification. Workers took on average 1 minute to complete the task and were rewarded 16 cents per answer (we made sure that we pay over 8USD per hour). The crowdsourcing task used to collect the *Amazon* dataset consisted in asking workers: *Is this review written on a book?* The full experiment is described in length in (Ramírez et al. 2019).

Representation Learning. The inputs of our machine learning model are the sentences from the documents. We represent each sentence as a fixed-size vector $\mathbf{v}_s$ by leveraging pre-trained language models. We use SciBERT as pre-trained word embeddings for *Wiki-Tech* since the language in Wikipedia is formal and contains scientific terms and ALBERT for *Amazon* as it contains reviews with less formal language compared to the documents used to train SciBERT. Considering the size of the datasets, we use a neural network with one fully-connected layer for sentence-level label prediction (Eq. (2)).

Comparison Methods. We compare our approach to a wide range of baselines. First, we compare against a set of recent text classification methods: 1) MILNET (Angelidis and Lapata 2018), a Multiple Instance Learning (MIL) neural network model. 2) fastText (Joulin et al. 2017), a linear model

[1]Source code and data are available at https://github.com/eXascaleInfolab/MARTA.

[2]https://www.mturk.com/

Method	Wiki-Tech				Amazon			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
MILNET	0.683	0.340	0.890	0.490	0.840	0.850	0.820	0.840
fastText	0.829	0.521	0.268	0.349	0.780	0.750	0.888	0.804
SciBERT	0.779	0.440	**0.970**	0.600	0.920	0.940	0.900	0.92
ALBERT	0.882*	**0.708**	0.560	0.618*	0.946	0.960*	0.932	0.946
LSTM-ortho	0.799	0.464	0.829	0.590	0.822	0.699	0.756	0.725
LSTM-diversity	0.649	0.365	0.928*	0.506	0.952*	0.960*	**0.944**	0.952*
InvRAT	0.717	0.220	0.210	0.230	0.720	0.750	0.720	0.710
RA-CNN	0.813	0.428	0.442	0.432	0.667	0.652	0.68	0.661
MARTA	**0.886**	0.660*	0.700	**0.680**	**0.960**	**0.980**	0.940*	**0.960**

Table 2: Performance (Accuracy, Precision, Recall and F1-score) comparison with baseline methods. The best performance is highlighted in bold; the second best performance is marked by '*'.

for text classification that uses bags of n-grams as additional features to capture information about the local word order. 3) SciBERT (Beltagy, Lo, and Cohan 2019), a language model trained on scientific text consisting of scholar papers from the computer science and biomedical domains. 4) ALBERT (Lan et al. 2019), a pre-trained language model that takes into account the inter-sentence coherence, which allows to capture fine-grained information in documents.

In addition, we compare against rational-aware models: 1) LSTM-ortho and LSTM-diversity, both proposed in (Mohankumar et al. 2020). These methods extend an LSTM to learn diverse hidden representations at different time steps through an orthogonality and a diversity constraint for hidden states. 2) InvRat (Chang et al. 2020), a game-theoretic approach that is designed to identify and remove features with spurious correlation with the output. 3) RA-CNN (Zhang, Marshall, and Wallace 2016), a sentence-level convolutional model that estimates the probability of a given sentence being a rationale. We note that the LSTM variants (LSTM-ortho and LSTM-diversity) and InvRat generate rationales automatically from the models, while RA-CNN uses the rationale provided by workers. In our experiment, we use the sentences indicated by the majority of workers as rationales to train RA-CNN.

Evaluation Protocol. We split the datasets into training, validation, and test sets. We use 50% of the data for training and the rest for validation and test with equal split. We report the average over 10 runs for each method. Note that we only use worker's annotations and rationales in the training and validation sets. We use accuracy, precision, recall and F1-score over the positive class to measure the performance. Higher values indicate better performance.

Results and Discussion

Table 2 summarizes the performance of MARTA against baseline methods on both *Wiki-Tech* and *Amazon*.

First, we observe that ALBERT and SciBERT perform relatively well compared to the other baseline methods, especially on the *Wiki-Tech* dataset. Recall that both ALBERT and SciBERT leverage textual context for representation learning, which is useful in fine-grained classification tasks, such as *Wiki-Tech* where the model has to capture the relationship between technologies and companies. Second, we observe that among the rationale-aware models, the two LSTM variants, i.e., LSTM-ortho and LSTM-diversity, achieve the highest performance. This confirms the advantage of attention mechanisms and shows the effectiveness of learning non-redundant hidden states for model performance. We also observe that RA-CNN, which uses human rationales, does not necessarily perform well. This is probably due to the way textual data is handled by RA-CNN: as opposed to the LSTM variants where the sequential order in the textual data is modeled (with attention), the textual data is considered as independent tokens by RA-CNN, which can lead to a loss of contextual meaning.

Most importantly, MARTA achieves the best performance in terms of accuracy and F1-score on both datasets. Overall, it improves ALBERT by 0.97% accuracy and 5.76% F1-score and LSTM-diversity by 18.68% accuracy and 17.61% F1-score on average on both datasets. To further confirm that our way of integrating human rationales is effective, we conducted an ablation study comparing MARTA to a simplified version with only the attention-based model (with pre-trained sentence embeddings). Results show that MARTA improves the performance by 23% accuracy and 28% F1-score in the *Wiki-Tech* dataset and by 12.5% accuracy and F1-score in the *Amazon* dataset. Such a result underlines the effectiveness of weighting the reliability of human rationales when integrating them into attention-based models.

MARTA Properties

In addition to better classification performance, MARTA exhibits a number of properties highly desirable in terms of accountability and deployment. In the following, we present some of these properties.

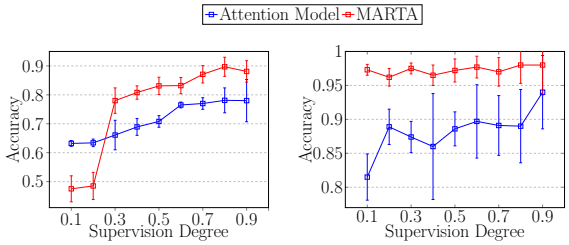
Explainability. MARTA provides explanations to classification results by incorporating human rationales. A comparison of the overlap between the rationales chosen by annotators and those highlighted by MARTA shows a recall of 71.1% on the *Wiki-Tech* dataset and 61.3% on the *Amazon* dataset, which is respectively 45.1% and 22.6% higher than the attention-based model alone. The precision is 21.7% on *Wiki-Tech* and 27.0% on *Amazon*. These results are due to the fact that MARTA typically tends to select multiple rel-

(a) **Microwave transmission is the transmission of information by microwave radio waves.** Although an experimental 40-mile (64 km) microwave telecommunication link across the English Channel was demonstrated in 1931, the development of radar in World War II provided the technology for practical exploitation of microwave communication. In the 1950s, large transcontinental microwave relay networks, consisting of chains of repeater stations linked by line-of-sight beams of microwaves were built in Europeand America to relay long distance telephone traffic and television programs between cities. *Communication satellites which transferred data between ground stations by microwaves took over much long distance traffic in the 1960s. In recent years, there has been an explosive increase in use of the microwave spectrum by new telecommunication technologies such as wireless networks, and direct-broadcast satellites which broadcast television and radio directly into consumers' homes.*
(b) In the Atom Station, Halldor Laxness demonstrates the skill and complexity that led to his being awarded the Nobel Prize in Literature. *The novel tells the story of a simple lass from the north of Iceland who comes face to face with the duplicity of politicians who sell out Icelandic sovereignty for the sake of a nuclear station during the cold war.* **She also comes to some realizations about herself and the importance of social class and knowledge and how these interact in today's modern world.** The novel will be of very special interest to those with some knowledge of Iceland and its history. For those without such knowledge, the novel will compel you to learn more about this fascinating country and its wonderful author laureate, Halldor Laxness.

Table 3: Examples from the *Wiki-Tech* (a) and *Amazon* (b) datasets. Bold letters refer to the weight attributed by an attention-based model. Italic letters indicate a rationale given by a worker. The shades of green refer to the weights given by MARTA: a stronger shade means a higher weight.

evant sentences as a rationale, while workers tend to select only one sentence. Table 3 shows two examples from the test sets of the *Wiki-Tech* and the *Amazon* datasets, respectively. The first example describes a technology used by companies. The attention-based model attributes a high weight to the first sentence (in bold), which defines the concept as a technology but not as used by companies. In comparison, our framework attributes high weights to the last two sentences given as rationales by workers (in italic), as they clearly show the relationship between the technology and companies. In addition, MARTA attributes a high weight to the second sentence that is relevant to the task. These results show that MARTA learns to generalize from human rationales how to identify important sentences. We observe similar results in the example from the *Amazon* dataset: our framework identifies both worker-provided rationales and other relevant sentences for the task.

Adjustable Supervision Degree. Our framework is highly effective even with a relatively small amount of ground truth labels for training. In what follows, we study the impact of the supervision degree to determine the minimum amount of ground truth needed. We split our datasets by s_{deg} where we vary s_{deg} between 10% and 90%, where $s_{deg} = 50\%$ means that we use 50% of the ground truth labels for training. We compare with a variant of our model with only the attention-based model described in Section 3. The results are shown in Figure 2. We observe that the performance of our framework increases along with the increase of s_{deg} on the *Wiki-Tech* data while it is overall stable for the *Amazon* data. This shows that the amount of ground truth needed to train our framework varies across tasks: compared with *Amazon*, *Wiki-Tech* is a more complex task that requires the model to capture fine-grained information; consequently, it requires more labels in model training. We observe that MARTA has better performance than the attention-based model starting from a supervision degree of 30% on the *Wiki-Tech* dataset and 10% on the *Amazon* dataset. This on one hand, confirms the effectiveness of integrating human rationales for model performance. On the other hand, the fact that a small proportion of labels (less than 30%) does not help to improve model performance on the *Wiki-Tech* dataset indicates that when the task is complex, a small proportion of ground truth labels might not be sufficient to correctly identify the workers' reliability, and that the benefits of having the labels might be over-weighted by the disadvantage of the extra parameters to be learned in that case.



(a) Wiki-Tech - Accuracy (b) Amazon - Accuracy

Figure 2: Performance of MARTA with varying s_deg.

In addition, we measure the performance variation across 10 runs with different data split for each s_deg. The results are shown in Figure 2, where the standard deviation is 0.044 and 0.019 on average on *Wiki-Tech* and *Amazon*, respectively. The standard deviation is small compared to the absolute accuracy which demonstrates MARTA's robustness.

5 Conclusion

In this paper, we presented MARTA, a Bayesian framework leveraging human rationales to improve the performance of attention-based models and provide a human-understandable explanation of classification results. Our proposed method incrementally updates the attention distribution by learning from human rationales while taking into account the workers' reliability. Extensive validation on two real-world datasets shows that MARTA is an effective and robust framework that substantially outperforms state-of-the-art methods while providing better, human-understandable explanations.

Acknowledgments

This work was supported by the armasuisse Science and Technology, R&D agency of the Swiss Armed Forces.

References

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*. OpenReview.net.

Angelidis, S.; and Lapata, M. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics* 6: 17–31.

Arous, I.; Dolamic, L.; Yang, J.; Bhardwaj, A.; Cuccu, G.; and Cudré-Mauroux, P. 2020. MARTA: Leveraging Human Rationales for Explainable Text Classification. Supplementary Material. https://exascale.info/assets/pdf/arous2020AAAI-sm.pdf.

Arras, L.; Horn, F.; Montavon, G.; Müller, K.-R.; and Samek, W. 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one* 12(8): e0181142.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1903–1913. Association for Computational Linguistics.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3606–3611. Association for Computational Linguistics.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518): 859–877.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, 9539–9549.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. 2019. A Game Theoretic Approach to Class-wise Selective Rationalization. In *Advances in neural information processing systems*, 10055–10065.

Chang, S.; Zhang, Y.; Yu, M.; and Jaakkola, T. S. 2020. Invariant Rationalization. *arXiv preprint arXiv:2003.09772* .

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610* .

Chhatwal, R.; Gronvall, P.; Huber-Fliflet, N.; Keeling, R.; Zhang, J.; and Zhao, H. 2018. Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding. In *2018 IEEE International Conference on Big Data (Big Data)*, 1905–1911. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Freitas, A. A. 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter* 15(1): 1–10.

Jain, S.; and Wallace, B. C. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* .

Joulin, A.; Grave, É.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Association for Computational Linguistics.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–1684. ACM.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*. OpenReview.net.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* .

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* .

Liu, H.; Yin, Q.; and Wang, W. Y. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5570–5581. Association for Computational Linguistics.

Mahoney, C. J.; Zhang, J.; Huber-Fliflet, N.; Gronvall, P.; and Zhao, H. 2019. A Framework for Explainable Text Classification in Legal Document Review. In *2019 IEEE International Conference on Big Data (Big Data)*, 1858–1867. IEEE.

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.

Mohankumar, A. K.; Nema, P.; Narasimhan, S.; Khapra, M. M.; Srinivasan, B. V.; and Ravindran, B. 2020. Towards

Transparent and Explainable Attention Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4206–4216. Association for Computational Linguistics.

Parikh, A. P.; Täckström, O.; Das, D.; and Uszkoreit, J. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933* .

Ramírez, J.; Baez, M.; Casati, F.; and Benatallah, B. 2019. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 144–152. AAAI Press.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717* .

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11): 2673–2681.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* .

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.

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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 189–198. Association for Computational Linguistics.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11–20.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. JMLR.org.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 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. The Association for Computational Linguistics.

Yu, M.; Chang, S.; Zhang, Y.; and Jaakkola, T. 2019. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. Association for Computational Linguistics.

Zaidan, O.; Eisner, J.; and Piatko, C. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, 260–267. The Association for Computational Linguistics.

Zhang, Q.; Nian Wu, Y.; and Zhu, S.-C. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8827–8836. IEEE Computer Society.

Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, 795. NIH Public Access.