UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

TU Delft

MARTA: Leveraging Human Rationales for Explainable Text Classification

Inès Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj
Giuseppe Cuccu, Philippe Cudré-Mauroux

February 2021, AAAI

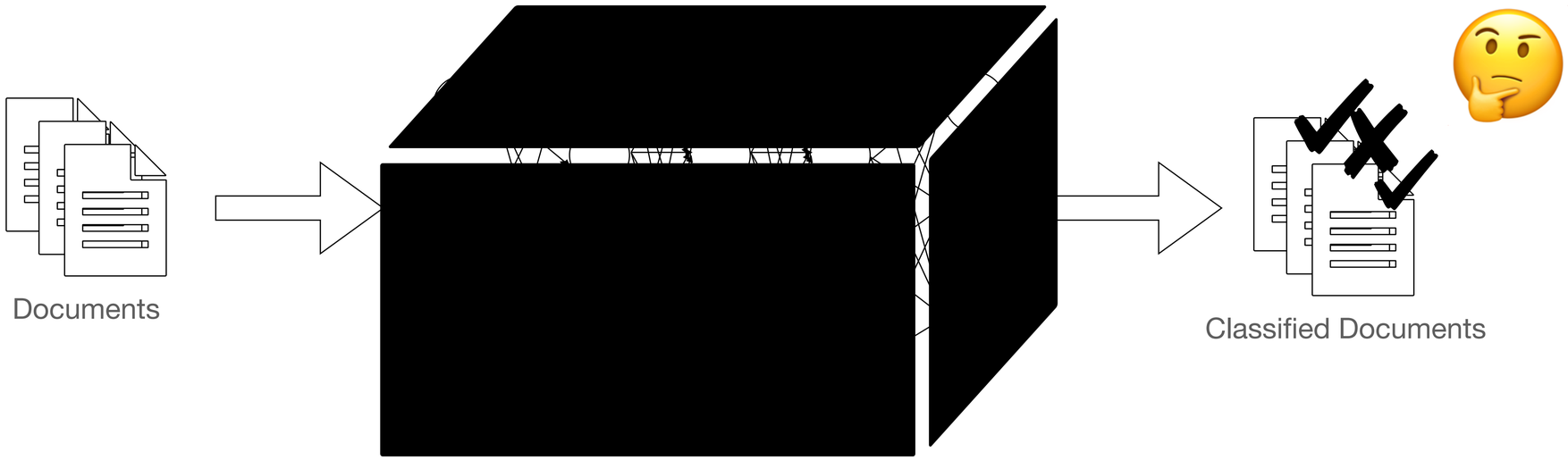
eXascale Infolab

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
armasuisse
Science and Technology

Contents

Introduction

Neural Networks for Text Classification



- State of the art methods for text classification rely on neural networks.
- These methods are often considered as **black boxes** by end users as their output is **hard to interpret**.

Introduction

Explainable Text Classification Methods

- The goal is to present end-users with **human-readable description** of the classification result: “**a rationale**”.
- Among these methods, a popular approach consists of using **attention mechanisms.**
- An attention mechanism assigns an **attention weight** to each sentence of the text.
- Does the attention weight provide a rationale?

Introduction

Example using Attention Mechanisms

Review
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 […] 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. […]

Method	Rationale for the classification of the review as a book
Ideal	The novel tells the story of a simple lass from the north of Iceland.
Attention Mechanism	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 rationale given by the attention mechanism is not accurate.

Introduction

Attention Vs Explanation

- Recent studies have shown that:
  - the attention distribution is inconsistent with the input units [1]
  - two different attention distributions can yield the same result [2].
- Solution: Enhance the explainability of attention based models by integrating **human rationales**.

[1] Jain, S. and Wallace, B.C., 2019. **Attention is not explanation.** *arXiv preprint arXiv:1902.10186*.
[2] Wiegreffe, S. and Pinter, Y. **Attention is not not explanation.** *EMNLP 2019*.

Introduction

Integrating Worker Rationales for Attention Learning

- Rationale provided by crowd workers can guide the learning of the attention distribution.

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 […] 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. […]

Introduction

Integrating Worker Rationales for Attention Learning

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 […] 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. […]

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 […] 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. […]

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 […] 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. […]

- Rationales provided by workers are not all the same and highly **depend on workers reliability**.

Introduction

Challenges

- The rationale given by the attention mechanism is not accurate.
  - How can we make the attention distribution reflect a **higher quality explanation** for text classification?
- Impact of worker's reliability on rationale's quality:
  - How can we quantify **the reliability of workers** when they provide us with rationales and labels?

Introduction

Contributions

☑ We propose **MARTA**, a Bayesian framework for MApping human Rationales To Attention.

☑ 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** outperforms state-of-the-art methods.

Contents

MARTA

The Crowdsourcing Task

- Predicate: Is the review about a book?

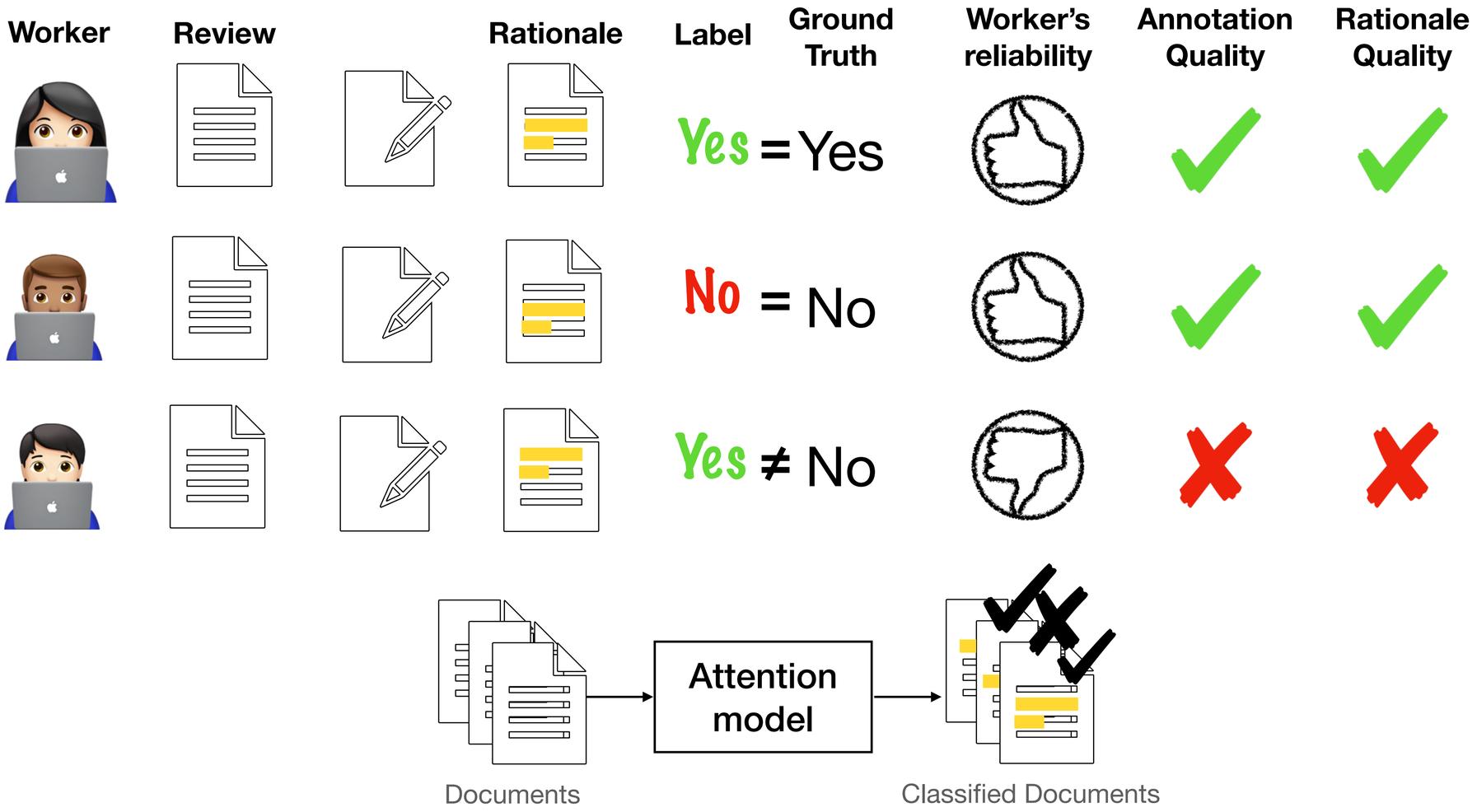
MARTA

Problem Formulation

- Given:
  - A set of textual documents $\mathcal{I}$.
  - Labels and rationales provided from workers of some documents ($\mathcal{I}_w \in \mathcal{I}$).
  - Ground truth labels of some documents ($\mathcal{I}_l \in \mathcal{I}_w$).
- Our goal is to:
  - Infer the **true label** z_i of each document $i \in \mathcal{I}$.
  - Estimate the **importance** α_s of each sentence in a document $i \in \mathcal{I}$.

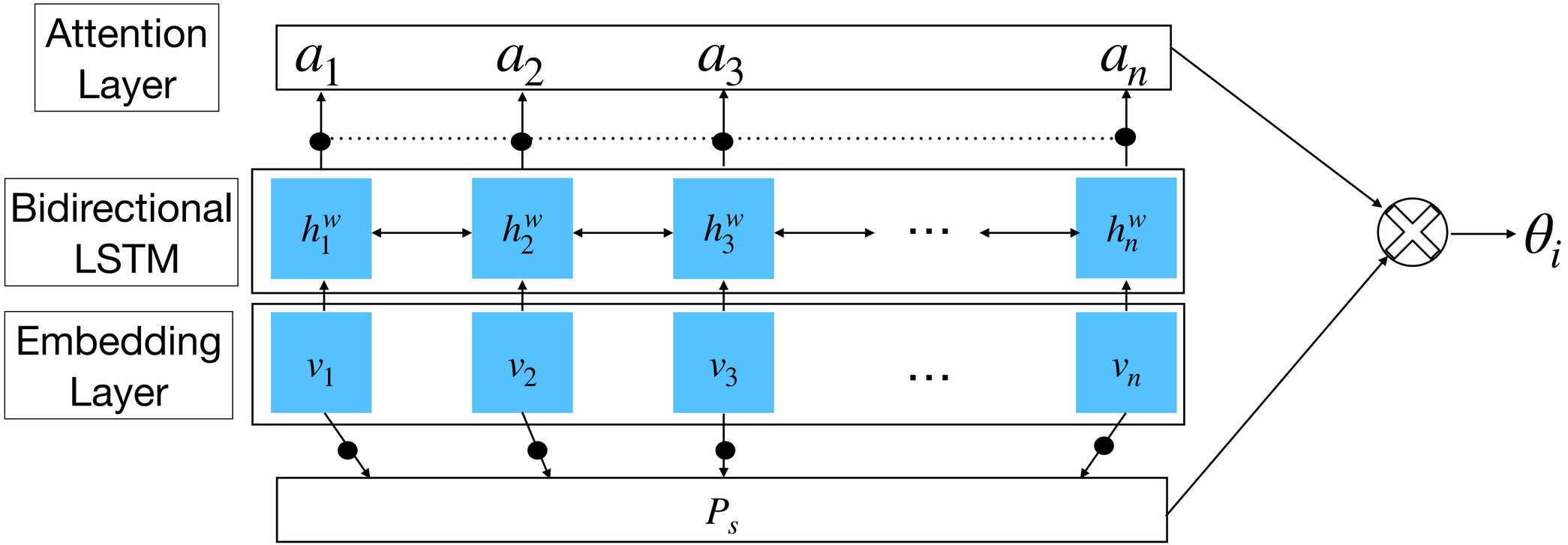
MARTA

In a nutshell



MARTA

Rationale-Aware Attention Model



- We use worker's rationale to guide the learning of the attention distribution.
- Worker’s annotation of rationale are not fully reliable.

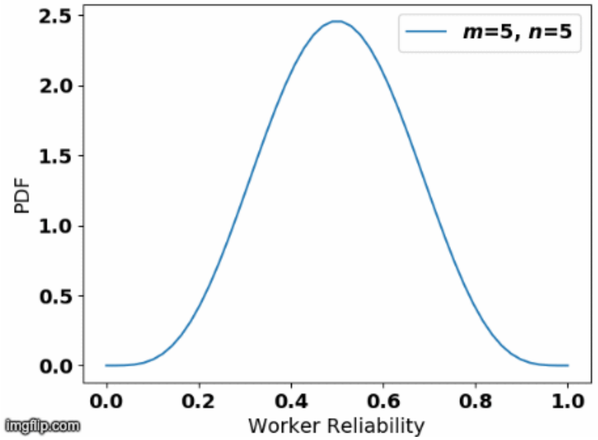
MARTA

Modeling Worker Reliability

- The worker reliability $r_j \in [0,1]$: If $r_j = 1$, the worker is fully reliable and $r_j = 0$ otherwise.
- The support of a Beta distribution is in $[0,1]$

➡ We model the worker reliability with a Beta distribution: $r_j \sim Beta(m, n)$

- The likelihood of worker's rationale being correct depends on her reliability.



MARTA

Updating the Worker Reliability

The worker reliability can be incrementally computed using:

★ her annotation quality θ_i

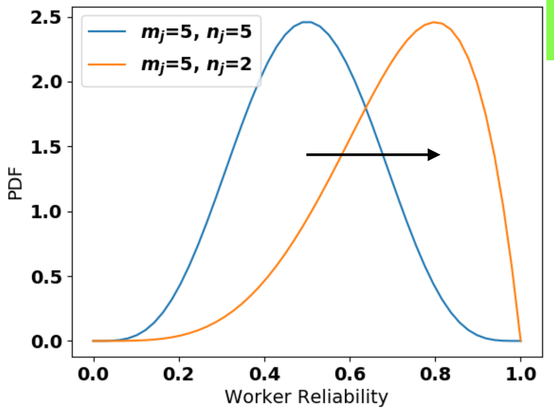
★ her rational quality a_s

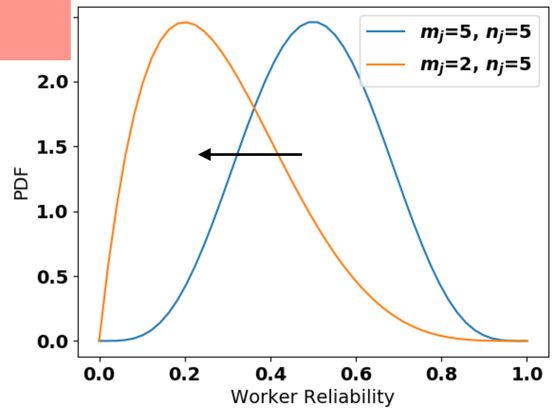
such that ***the more correct answers*** she provides, ***the more reliable*** she is.

$$q(r_j) = Beta(m_j + \sum_{i \in \mathcal{I}_j} \theta_i + \sum_{s \in \mathcal{S}_j} a_s, n_j + \sum_{i \in \mathcal{I}_j} (1 - \theta_i) + \sum_{s \in \mathcal{S}_j} (1 - a_s))$$

correct answers

wrong answers





MARTA

Modeling the Document Classification

- The true label of a document is a binary variable.
- We model the true label of a document with a Bernoulli distribution: $z_i \sim Ber(\theta_i)$

The document classification can be incrementally computed using:

★ The predicted label θ_i

★ The reliability parameters m_j and n_j.

$$q(z_i) = \theta_i \prod \exp\{\Psi(m_j) - \Psi(m_j + n_j)\}$$

Predicted label — Geo. Mean of the reliability

MARTA

Modeling the Sentence Importance

- The importance of a sentence is a binary variable.
- We model if a sentence is a rationale with a Bernoulli distribution: $\alpha_s \sim Ber(a_s)$

The sentence importance can be incrementally computed using:

★ The predicted importance a_s

★ The reliability parameters m_j and n_j.

$$q(\alpha_s) = a_s \prod \exp\{\Psi(n_j) - \Psi(m_j + n_j)\}$$

Predicted importance — Geo. Mean of the reliability

MARTA

Variational Inference Algorithm

Input: Worker's labels, Worker's Rationales, Textual documents

Output: Worker reliability, Document classification, Sentences importance

Repeat

#E step:

Incrementally compute the worker reliability

Incrementally compute the document classification

Incrementally compute the sentence importance

#M step:

Update the rationale-aware model

Until *convergence*

Contents

Experiments

Datasets

- Task: We ask workers to specify the **relevance of documents** to a certain topic and extract the part of text **justifying** their answer.

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

- Metrics: We use **Accuracy**, Precision, Recall and **F1-score**.

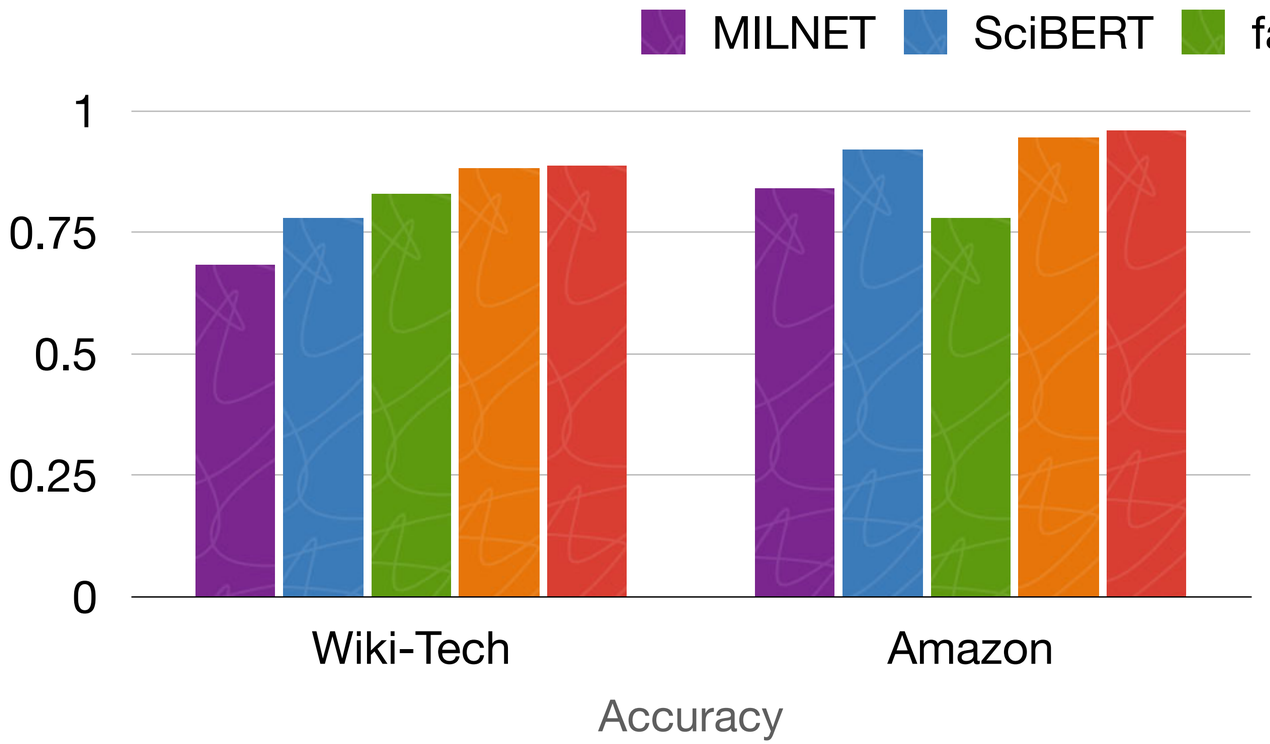
Experiments

Baselines (1/2)

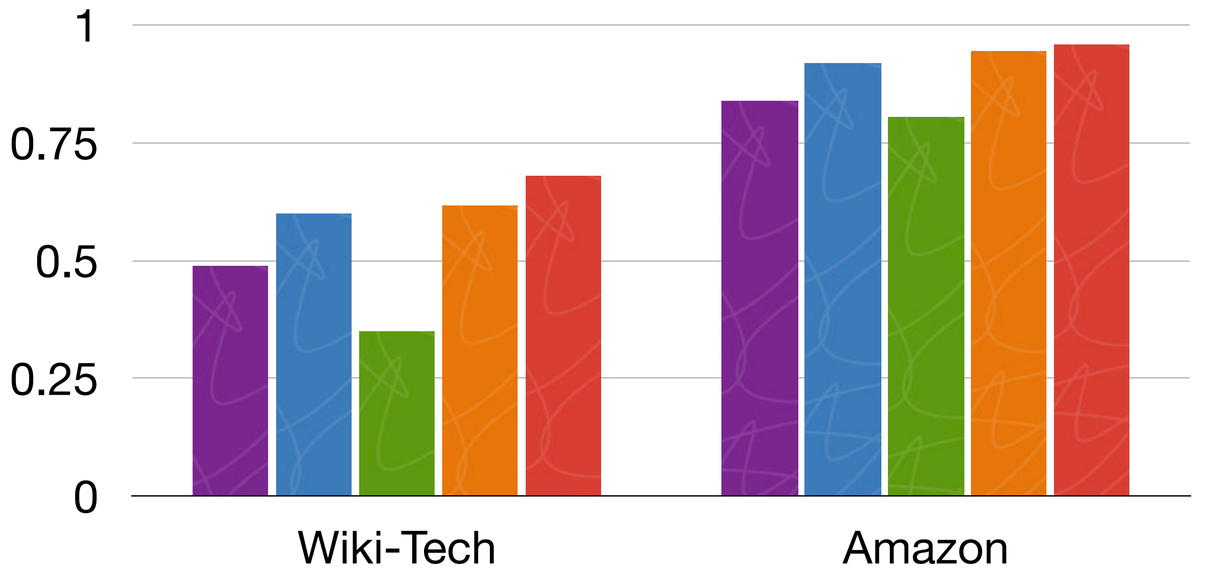
- **Text classification methods:**
  - *MILNET (TACL'18):* a Multiple Instance Learning neural network.
  - *fastText (EACL'17):* Linear model on top of bags of n-grams.
  - *SciBERT (EMNLP'19)* and *ALBERT (ICLR'19)*: pre-trained language models with a linear classifier.

Experiments

Comparison with Text Classification Methods



Accuracy



F1-score

- **MARTA** improves text classification methods by **0.97% accuracy** and **5.76% F1-score**.
- The rationale highlighted by workers help guiding our rationale-aware model.

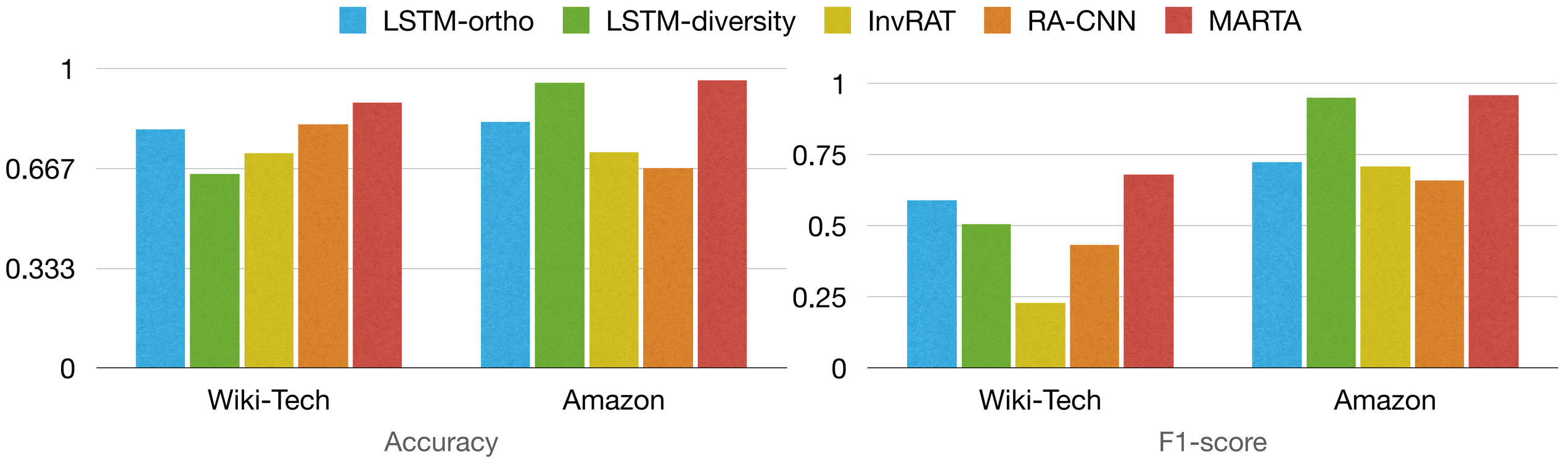
Experiments

Baselines (2/2)

- **Rationale-aware model:**
  - *LSTM-ortho, LSTM-diversity (ACL'20):* An extension of LSTM with diversity constraints on hidden states.
  - *InvRAT (NeurIPS'19):* A game theoretic approach aiming at identifying features correlating with the output.
  - *RA-CNN (EMNLP'16):* A sentence-level convolutional model that estimates if a sentence is a rationale.

Experiments

Comparison with Rationale-aware Methods



- **MARTA** improves rationale-aware methods by **18.68% accuracy** and **17.61% F1-score.**
- **MARTA** is better at leveraging rationale by weighing workers reliability.

Experiments

Explainability

- The overlap between the rationales chosen by workers and those highlighted by MARTA is **66%**.

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 […] 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. […]	Book related review	**Attention Mechanism**
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 […] 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. […]	Book related review	
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 […] 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. […]	Book related review	**MARTA**

Contents

Conclusion

- We introduced **MARTA** for explainable text classification that integrates ***human rationales*** into an ***attention-based model***.
- **MARTA** substantially outperforms state of the art by **5.76% F1-score**.
- **MARTA** offers a **human understandable** explanation for text classification.
- Future work includes: token level explainability and leveraging workers justification expressed in a syntax different from the original text.

Thanks for your attention!
Any questions?

https://github.com/
eXascaleInfolab/MARTA

eXascale Infolab

https://exascale.info