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]{ines.arous,akansha.bhardwaj,giuseppe.cuccu,pcm}@unifr.ch,
[2]Ljiljana.Dolamic@armasuisse.ch, [3]J.Yang-3@tudelft.nl

1. Motivation and Challenges

Motivation: Explainability is a key requirement for text classification in many application domains including medical diagnosis and legal reviews.

Challenges: Existing explainable models rely on attention mechanisms, but:

- Attention-based models often provide **an inaccurate explanation** of text classification.

Using human rationales is challenging because:

- Workers provide **different rationales** for the same text classification.
- **Quality** of rationales provided by workers depend on their **reliability**.

2. Problem Definition

Given

- A set of textual documents
- Labels and rationales provided from workers of some documents

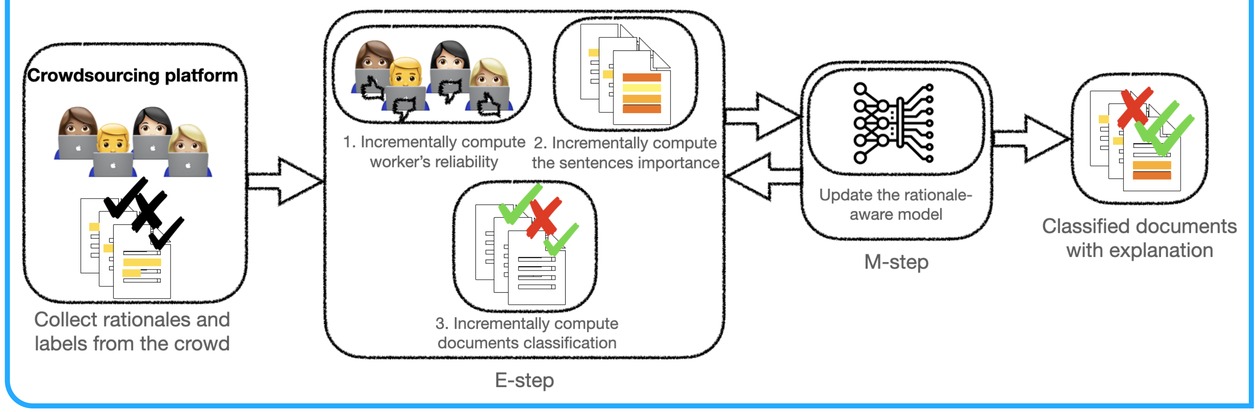
Our goal is to:

- Infer **the true label** of each document
- Estimate **the importance of each sentence** in a document.

3. Contributions

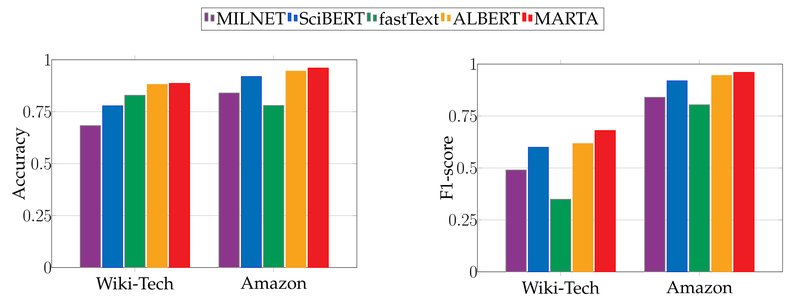
- A Bayesian framework called MARTA for MApping human Rationales To Attention.
- A Variational Inference algorithm with incremental updating rules.

4. The MARTA Framework

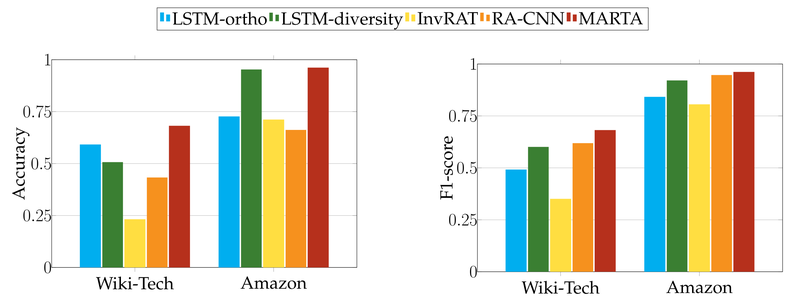


5. Results

- MARTA improves **text classification methods** by **0.97% accuracy** and **5.76% F1-score**.



- MARTA improves **rationale-aware methods** by **18.68% accuracy** and **17.61% F1-score**.



- MARTA is better at leveraging rationale by weighing workers reliability.

6. Properties

- *Explainability*: The overlap between the rationales chosen by workers and those highlighted by MARTA is **66%** on average on two datasets.
- *Adjustable Supervision Degree*: MARTA requires a small amount (max 30%) of ground truth labels for training.
- *Robustness*: The standard deviation is small (~ 0.03) compared to the absolute accuracy which demonstrates MARTA's robustness.

7. Conclusion

- We proposed MARTA for explainable text classification that integrates **human rationales** into an **attention-based model**.
- As future work, we plan to take into account worker's rationales expressed in a syntax different from the original text.

Published in: *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*, A Virtual Conference, February 2021.