XAIT: An Interactive Website for Explainable AI for Text

Erick Oduor
ericko@ke.ibm.com
IBM Research – Africa

Kun Qian
qian.kun@ibm.com
IBM Research – Almaden

Yunyao Li
yunyaoli@us.ibm.com
IBM Research – Almaden

Lucian Popa
lpopa@us.ibm.com
IBM Research – Almaden

ABSTRACT

Explainable AI (XAI) for text is an emerging field focused on developing novel techniques to render black-box models more interpretable for text-related tasks. To understand the recent advances in XAI for text, we have done an extensive literature review and user studies. Allowing users to easily explore the assets we created is a major challenge. In this demo we present an interactive website named XAIT. The core of XAIT is a tree-like taxonomy, with which the users can interactively explore and understand the field of XAI for text through different dimensions: (1) the type of text tasks in consideration; (2) the explanation techniques used for a particular task; (3) who are the target and appropriate users for a particular explanation technique. XAIT can be used as a recommender system for users to find out what are the appropriate and suitable explanation techniques for their text-related tasks, or for researchers who want to find out publications and tools relating to XAI for text.

CCS CONCEPTS

• **Human-centered computing** → **Visualization toolkits**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Explainable AI; Natural Language Processing

ACM Reference Format:
Erick Oduor, Kun Qian, Yunyao Li, and Lucian Popa. 2020. XAIT: An Interactive Website for Explainable AI for Text. In *25th International Conference on Intelligent User Interfaces Companion (IUI '20 Companion), March 17–20, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3379336.3381468

IUI '20 Companion, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7513-9/20/03.
https://doi.org/10.1145/3379336.3381468

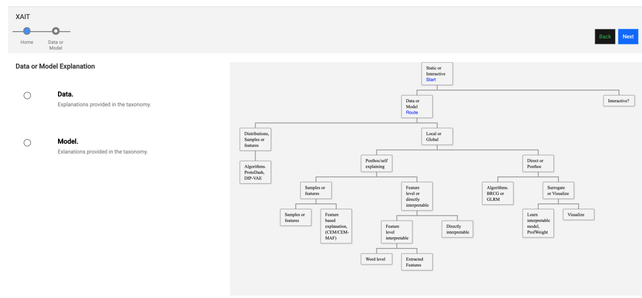

Figure 1: Checkmark options for navigating Data or Model explanations (L) of the XAI taxonomy (R).

1 INTRODUCTION

We are increasingly experiencing the impact of artificial intelligence in our routine life as machine learning algorithms are applied to provide us with improved services in areas such as financial credit applications, hiring in human resource and even prediction of cancer in healthcare. The complex implementation of ML models and advances that have led to using AI in diagnosing illness or managing finances have raised questions on whether we can trust the blackbox calculations of AI. Whether to trust or not, is a basic question that we practice in our daily lives around safe medication or even fair treatment when we apply for financial credit. Even though some machine learning models are easy to comprehend, blackbox models are more complex and challenging to understand and trust [3]. We contribute towards the growing research around interpretability and explainability of AI models with a focus on text related tasks.

Background

Our work started with an extensive literature review over 200 papers recently published (2014-2019) in top AI and NLP conferences (e.g., AAAI, IJCAI, KDD, ACL, EMNLP) that focused on text related tasks. We also conducted an interview study with individuals working on real-world text-related projects within IBM to identify practical challenges related to the understandability of different explanation techniques [2].

For each paper we reviewed, we identify (1) what is the target text task (e.g., machine translation, question answering), (2) what is the explanation technique it uses (e.g., human-readable rules, highlighting important words), and (3) who are the appropriate target users (AI experts or lay users).

Valuable insights can be obtained by exploring our findings. For example, AI engineer can use XAIT to quickly explore different possible options and identify the most suitable technique(s) for their use cases. However, it is not easy to present our findings in an intuitive and accessible way so that users can quickly find the most relevant information. Clearly, listing our results in a spreadsheet will not serve the purpose.

XAIT was designed to address these issues. The core of XAIT is a tree-like structure that organizes research publications, explanation techniques, and target users in a taxonomy that is based on the one originally presented in [1, 3]. The taxonomy presented in [1, 3] was for general XAI tasks, the main goal of which is to suggest suitable toolkits for different XAI tasks. We reused the components (both the taxonomy structure and the available toolkits) that are text related. More importantly, we extend their taxonomy by adding more nodes, including relevant research publications, and suggesting target users according to our findings.

In this demo, we illustrate work in progress towards our overall goal to build a system that can recommend a list of possible explanation techniques as well as provide a list of existing approaches that adopted these explanation techniques.

2 OVERVIEW OF XAIT

As shown in Figure 1, our taxonomy is visually presented to the user. This visualization provides a global and structural view of different explanation techniques included in our taxonomy. The exploration starts from the root node, and the user can navigate down to the leaf nodes (according to the choices s/he made). Each inner node is a burst node that splits the exploration into different paths. The leaf nodes of the taxonomy contain the actual explanation techniques, toolkits (if available), target users, and existing publications that adopted these techniques. Each node also has a clickable link that can pop up a dialog that provides some high-level description of the purpose of this node. Below we introduce several key concepts used in XAIT.

Static or Interative Explanation. Whether the explanation is provided in a static way or an interactive way.

Data Explanation or Model Explanation. Whether the explanations are generated for data or an AI model.

Local or Global Explanation. Whether the goal is to provide only local explanation (i.e., explain the prediction of a particular input instance) or global explanations (i.e., explain a model's decision making process in general).

Surrogate model. In some machine learning scenarios, when we want to explain the predictions of a non-interpretable model, we learn a second interpretable surrogate model that approximates the predictions of the first model.

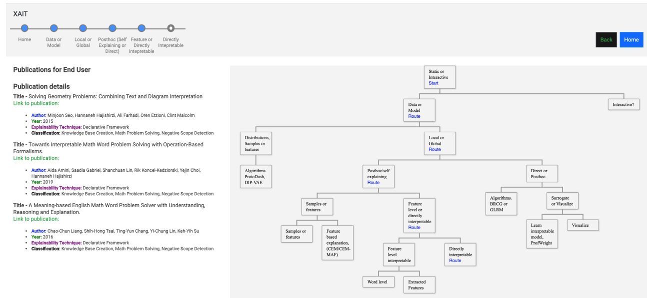

Figure 2: Publications and toolkits for directly and globally interpretable models.

Posthoc or self-explaining. Whether the explanation is generated via a self-explaining interpretable model (the learned model itself also generates some sort of explanations for its prediction) or post-hoc process (i.e., the learned model is not interpretable, and we need some post-hoc processes to explain its prediction, for example, learn a interpretable surrogate model).

Feature-level or directly interpretable. The explanations are based on feature-level elements (e.g., sparse word embeddings, or highlighted words) or the explanations are directly interpretable by human users (e.g., human-readable rules).

3 A CONCRETE DEMO EXPERIENCE

Assuming there is an AI engineer who needs to design an entity resolution (the task of identifying and linking different representations of the same real-world objects) model for end users who want to have a globally transparent AI model. The AI engineer can start with XAIT by selecting *static* mode at the root node, then selects *global* to indicate that s/he wants to provide global explanation. Then, s/he can select *Direct* to indicate that s/he wants to provide human-comprehensible explanations. After that, a leaf node with a list of relevant publications with detailed information and toolkits will be presented to the user (see Figure 2).

REFERENCES

[1] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[2] Shipi Dhanorkar, Christine Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2020. Tutorial: Explainability for Natural Language Processing. In *(AACL 2020)*. to appear.

[3] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.