AI Explainability WG

January 2024

January 2024 updates

- Happy new year! 🎉
- TrustyAI core / service
  - 0.10.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.10.0
  - quay.io/trustyai/trustyai-service:v0.10.0
- TrustyAI operator
  - 1.15.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.15.0
  - quay.io/trustyai/trustyai-service-operator:v1.15.0
- TrustyAI Python
  - 0.4.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability-python/releases/tag/0.4.0
  - https://pypi.org/project/trustyai/0.4.0/

What’s new?

TrustyAI - What’s new?

TrustyAI core/service 0.10.0

- Fixes: Fix name mapping mis-application and infinite loop

TrustyAI operator 1.15.0

- Fixes: Correct operator's instance label

Python TrustyAI 0.4.0

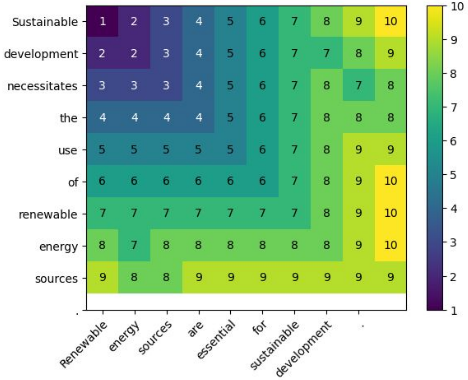
- Inclusion of language performance metrics (WER, Levenshtein)
- Switch to PyPi’s “Trusted Providers” release pipeline completed
- Update dependencies (pyarrow, pillow)

Python TrustyAI examples

Language performance metrics

- Visualisation for distance matrices

  -
    ```
    from trustyai.metrics.distance import levenshtein
    A = "Renewable energy sources are essential for sustainable development."
    B = "Sustainable development necessitates the use of renewable energy sources."
    d = levenshtein(A, B)
    d.plot()
    ```



	Renewable	energy	sources	are	essential	for	sustainable	development	.	
Sustainable	1	2	3	4	5	6	7	8	9	10
development	2	2	3	4	5	6	7	7	8	9
necessitates	3	3	3	4	5	6	7	8	7	8
the	4	4	4	4	5	6	7	8	8	8
use	5	5	5	5	5	6	7	8	9	9
of	6	6	6	6	6	6	7	8	9	10
renewable	7	7	7	7	7	7	7	8	9	10
energy	8	7	8	8	8	8	8	8	9	10
sources	9	8	8	9	9	9	9	9	9	9
.										

Python TrustyAI

Detoxification (HAP)

- **trustyai-detoxify** repo
  - https://github.com/trustyai-explainability/trustyai-detoxify/
  - Next step is to move into Python TrustyAI as an optional package
  - `pip install trustyai[detoxify]`
- Provides default expert / anti-expert models[1]
  - Ability to supply custom models
- Token scoring
- “I then realised that the number of **idiots** behaving like this are just a minority, as in background noise in the society.”
  - Text masking: “I then realised that the number of **<mask>** behaving like this are just a minority, as in background noise in the society.”
  - Rephrasing: “I then realised that the number of **people** behaving like this are just a minority, as in background noise in the society.”

[1]: https://huggingface.co/trustyai (gplus / gminus models)

Community

Documentation

- New TrustyAI documentation page
  - Central location for TrustyAI documentation
  - Antora (Asciidoc based)
  - https://trustyai-explainability.github.io/trustyai-site/main/main.html
  - Source
    - https://github.com/trustyai-explainability/trustyai-explainability.github.io

Roadmap

TrustyAI 2024 roadmap

- **December 2023 – March 2024 (proposal / discussion)**
    - Out-of-distribution (OOD) metrics completion
        - Finalise OOD metrics, namely data uploading aspect
    - Explainability service endpoints completion
        - Formalise explainability payloads schema
        - Refining handling of synthetic payloads to avoid interference with metrics
        - Handle potential large computational times in a service setting
    - Detoxification at the library and service level
        - Integrate detoxification with Python TrustyAI (Jupyter main target)
        - Token scoring at the service level
    - Database backend
        - Address scalability
        - Replace PVC with DB?
    - Expand supported types (*eg* image data)
        - Metrics and explainability for non-tabular data
    - Model drift/data drift/anomaly detection
    - Other model architecture