PieVal: an Open-Source, Efficient, Secure, Gamified, Rapid Document Classification Annotation Tool

Albert William Riedl, MS[1]; Aaron Seth Rosenberg, MD[1]; JP Graff, DO[1]; Matthew S Renquist[1]; Joseph M Cawood[1]; Nicholas R Anderson, PhD [1]

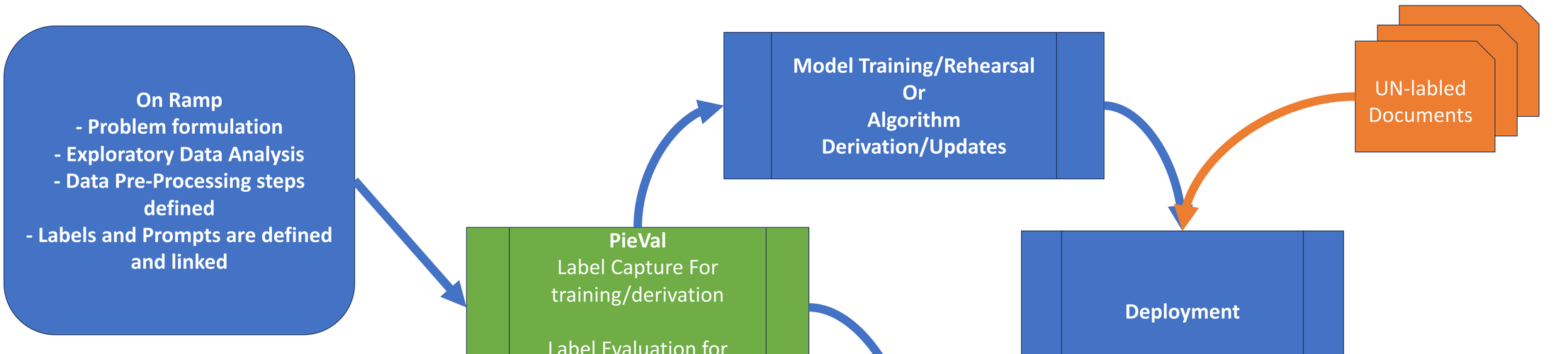
[1] University of California at Davis, Sacramento CA

SCAN ME

Check us out on GitHub for:
- More extensive Documentation
- Implementation Instructions

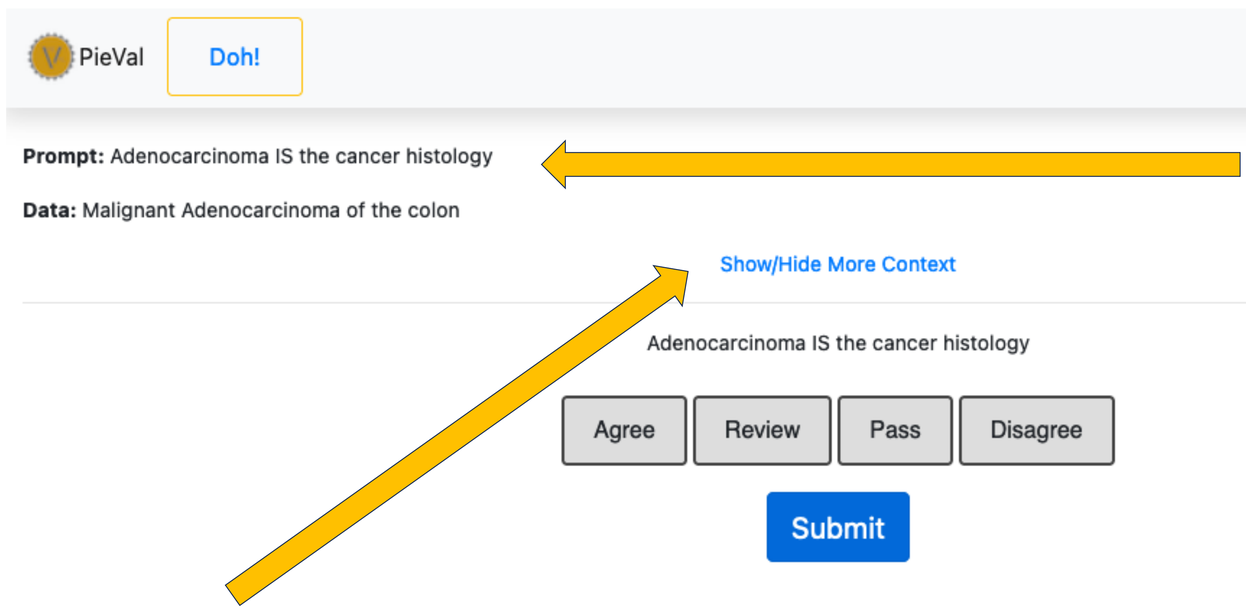
Introduction

- As much as **83% of biomedical information is contained within clinical notes** [1]
- Advancing Natural Language Processing **(NLP) tools can address this challenge.**
- **NLP is bottlenecked by the need for annotated datasets** for model training and evaluation
- **Annotated data is expensive** due to the need for highly trained personnel and the time required for annotation.
- **Expertise is required but inefficient annotation tools are not**
- **We developed PieVal, a web-based, secure, high-throughput document classification annotation tool**
- Document level annotations in **less than 30 seconds per document**, based on 20K annotations captured to-date



Built for continuous improvement

- The key to this integration is the relationship between labels and prompts
- Labels are a machine learning concept
- Prompts are human readable 'label' used in PieVal
- PieVal tests the assertion that a prompt (aka label) and some data are concordant – thus creating labels that can be used to train machine learning models
- After you have a deployed model, you can feed the machine generated lables back into PieVal for validation and to monitor drift over time

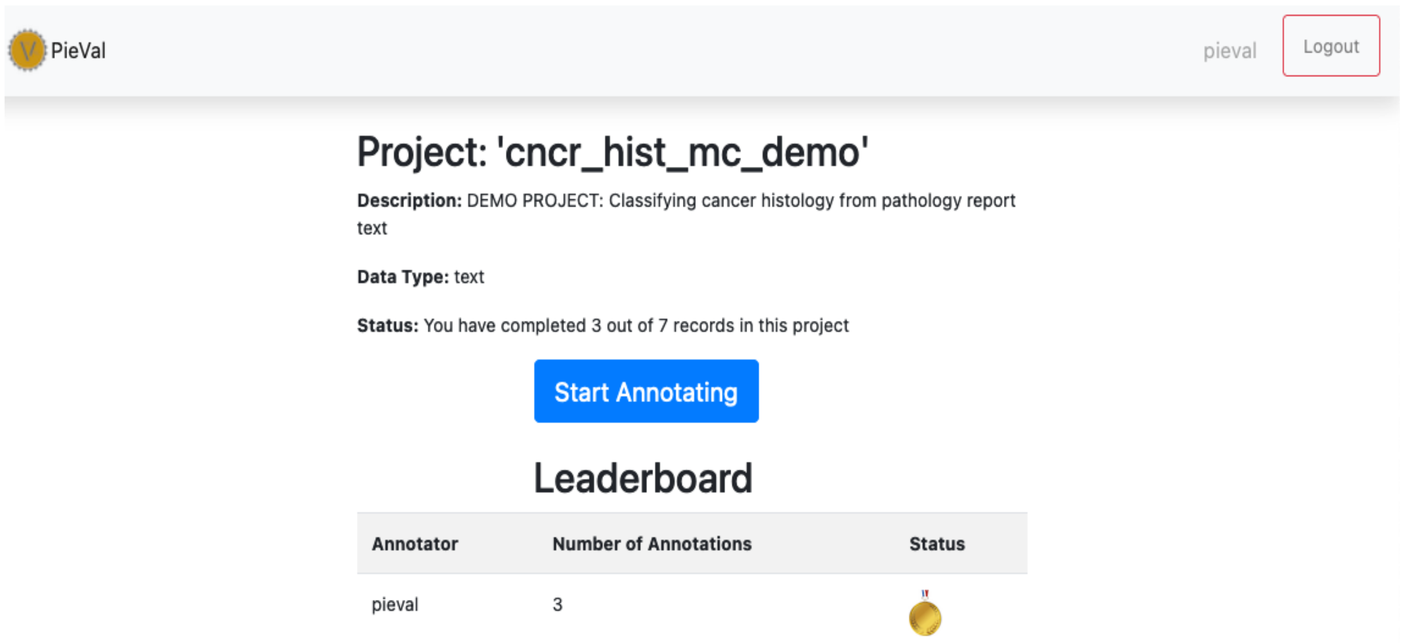


Annotations as assertion tests

- Asking about agreement allows for a consistent User Interface not matter what data OR the number of classes
  - Users presented with Prompt and Data and asked if the prompt applies to the data
    - Agree – yes, it does
    - Review – Unclear, and I have enough expertise to help clarify
    - Pass – Unclear, and I don't have enough expertise to clarify
    - Disagree – No, it does not
      - Multi-class annotations include a 'correction' workflow when disagreement is selected

Data Enrichment

- Data is encouraged to be populated with 'enriched' text
- Enriched text is the result of pre-processing that pulls text more likely to answer a question out of a larger document
- This is done purely for operational efficiency
- The less text to review, the faster annotations can be captured
- There is risk – pre-processing strategies can be flawed and sometimes eliminate the wrong text!
- PieVal includes a hidden Extended Data field that should contain the full orginal text
- It is accessible by clicking 'show/hide more context'
- Annotators can always review the original document if the enriched text leaves room for uncertainty
- Full document views are tracked, which helps gather data about when an enrichment strategy may be failing



Future Directions

- Continue to grow community support for a thriving Open-Source future
- Add Image and table data annotation workflows
- Include helpful pre-processors and post-processors
- Build in data quality metrics as well as operator agreement scores

Gamefied – Leaderboard with Gold, Silver Bronze Medals

- Use friendly competition to incentivize annotation adherence

References

1. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. *JAMA*2013;**309**:1351--