Blinded Experiments with TIRA
Towards Reproducible Shared Tasks

June 30, Leipzig

Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast

University of Jena University of Glasgow University of Leipzig University of Weimar

@webis_de www.webis.de
Blinded Experiments with TIRA

Motivation

How to advertise your Dataset / Paper?

- You are Rich / Famous?
- People will come anyway
Blinded Experiments with TIRA

Motivation

How to advertise your Dataset / Paper?

- You are Rich / Famous?
- People will come anyway

- Organize a shared task (e.g., 2023):
  - Clickbait Spoiling (30 of 83 teams submitted; 24 countries)
  - ValueEval (41 submitted)
  - Pan (32 submitted)
  - Touché (8 submitted)
Blinded Experiments with TIRA

Motivation

Your Shared Task?
Blinded Experiments with TIRA

Motivation

Potential problems (run submissions):

- Problem 1: Internal validity
- Problem 2: External validity

[Fuhr’21]
Blinded Experiments with TIRA

Motivation

Potential problems (run submissions):

- Problem 1: Internal validity
- Problem 2: External validity
- Problem 3: Blinded experimentation with LLMs
Blinded Experiments with TIRA

Problem 1: Internal Validity [Fuhr’21]

Goal

The hypothesis is supported by the data.
Blinded Experiments with TIRA

Problem 1: Internal Validity  [Fuhr’21]

Goal

The hypothesis is supported by the data.

Possible problems

- Wrong baseline  
  [Armstrong’09, Lin’18]

- Formulate hypothesis after experiments  
  [Fuhr’21]
Blinded Experiments with TIRA

Problem 1: Internal Validity  [Fuhr’21]

Goal

The hypothesis is supported by the data.

Possible problems

- Wrong baseline  
  [Armstrong’09, Lin’18]
- Formulate hypothesis after experiments  
  [Fuhr’21]

Possible solutions

- Centralized leaderboards
  - E.g., Run uploads to EvaluateIR  
    [Armstrong’09]
- Task-specific leaderboards
  - E.g., MS MARCO, MIRACL  
    [Lin’22, Zhang’22]
Blinded Experiments with TIRA

Problem 1: Internal Validity  [Fuhr’21]

Goal

The hypothesis is supported by the data.

Possible problems

- Wrong baseline  
  [Armstrong’09,Lin’18]

- Formulate hypothesis after experiments  
  [Fuhr’21]

Possible solutions

- Centralized leaderboards
  - E.g., Run uploads to EvaluateIR  
    [Armstrong’09]

- Task-specific leaderboards
  - E.g., MS MARCO, MIRACL  
    [Lin’22,Zhang’22]
Blinded Experiments with TIRA
Problem 1: Internal Validity [Fuhr’21]

Goal

The hypothesis is supported by the data.

Possible problems

- Wrong baseline
  [Armstrong’09,Lin’18]
- Formulate hypothesis after experiments
  [Fuhr’21]

Possible solutions

- Centralized leaderboards
  - E.g., Run uploads to EvaluateIR
    [Armstrong’09]
- Task-specific leaderboards
  - E.g., MS MARCO, MIRACL
    [Lin’22,Zhang’22]

“EvaluateIR never gained traction, and a number of similar efforts following it have also floundered” [Lin’18]
Blinded Experiments with TIRA

Problem 2: External Validity [Fuhr’21]

Goal

Repeating an experiment on similar data yields similar observations.
Blinded Experiments with TIRA

Problem 2: External Validity  [Fuhr’21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results
Blinded Experiments with TIRA

Problem 2: External Validity  [Fuhr’21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results

Possible Solutions

- TREC Open Runs  
  [Voorhees’16]
- Reproducibility initiatives
  - OSIRRC: Archive artifacts  
    [Arguello’15,Clancy’19]
  - CENTRE: Reimplementation  
    [Ferro’19,Sakai’19]
- Platforms + documentation
  - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- Meta evaluations: BEIR  
  [Thakur’21]
Blinded Experiments with TIRA

Problem 2: External Validity  [Fuhr’21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- Non-reproducible results

Possible Solutions

- TREC Open Runs  
  [Voorhees’16]
- Reproducibility initiatives
  - OSIRRC: Archive artifacts  
    [Arguello’15, Clancy’19]
  - CENTRE: Reimplementation  
    [Ferro’19, Sakai’19]
- Platforms + documentation
  - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- Meta evaluations: BEIR  
  [Thakur’21]

- 19 of 69 runs (Problems: 11)
- 2015: 8 systems archived  
  2019: 1 system fully reproducible  
  [Lin’19]
- Limited adoption of jig + CIFF  
  [Clancy’19]
- Additional effort
- Evaluations on subsets
- Often sparse judgments
I worry about language models being trained on test sets. Recently, we emailed support@openai.com to opt out of having our (test) data be used to improve models. This isn't enough though: others running evals could still inadvertently contribute those test sets to training.
Blinded Experiments with TIRA
Problem 3: Blinded Experimentation with LLMs

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces. Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

I worry about language models being used to improve models. This could still inadvertently cause contamination.

Percy Liang
@percyliang
TIRA to the Rescue?
Reproducible Shared Tasks with TIRA

Evolution of TIRA
[Gollub’12,Potthast’19,Fröbe’23]

- 2012–2022: Software submissions with virtual machines
- 2023–today: Immutable software submissions with Docker + Git CI/CD
  - Shared task = git repository
  - Software execution = commit
Reproducible Shared Tasks with TIRA

Evolution of TIRA
[Gollub’12,Potthast’19,Fröbe’23]

- 2012–2022: Software submissions with virtual machines
- 2023–today: Immutable software submissions with Docker + Git CI/CD
  - Shared task = git repository
  - Software execution = commit

Procedure:

1. Implement approach in Docker image
2. Upload image to dedicated image registry in TIRA
3. Your approach is executed in a Kubernetes cluster via a commit

http://tira.io

TIRA — Evaluation as a Service
Improving the replicability of shared tasks in computer science
Benefits of TIRA

Blinded Experimentation

- Software executed in sandbox: No internet connection
- 2 types of datasets:

<table>
<thead>
<tr>
<th>Type</th>
<th>Blinded</th>
<th>Unblinding</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>Nothing</td>
<td>Direct</td>
<td>Everything</td>
</tr>
<tr>
<td>Test</td>
<td>Everything</td>
<td>Manual</td>
<td>✓ vs ✗</td>
</tr>
</tbody>
</table>
Benefits of TIRA
Blinded Experimentation

- Software executed in sandbox: No internet connection
- 2 types of datasets:

<table>
<thead>
<tr>
<th>Type</th>
<th>Blinded</th>
<th>Unblinding</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>Nothing</td>
<td>Direct</td>
<td>Everything</td>
</tr>
<tr>
<td>Test</td>
<td>Everything</td>
<td>Manual</td>
<td>✓ vs ✗</td>
</tr>
</tbody>
</table>

Repeat, Replicate, and Reproduce in One Line of Code

- Git repository of the shared task can be published after the task

```python
import tira
df = tira.load_data('<dataset-name>')
predictions, evaluation = tira.run(
    '<task-name>/<user-name>/<software-name>',
    data=df, evaluate='<evaluator-name>'
)
```

- SemEval’23: 2 tasks, 83 + 91 reg. teams (active: 31 + 42; Docker: 21 + 7)
Benefits of TIRA
Run Experiments / Analysis on Confidential Data
Example: Paper for which you can't publish the Dataset

📖 The Archive Query Log

Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives.

Start now by running your custom analysis/experiment, scraping your own query log, or just look at our example files.
How you could set it up
What would you need?

- A (small) public dev/train dataset
- A (private) test dataset
- A dockerized baseline
- A Tutorial: How to run the baseline on dev?
- Evaluation measures (we already have many)
How you could set it up

What would you need?

- A (small) public dev/train dataset
- A (private) test dataset
- A dockerized baseline
- A Tutorial: How to run the baseline on dev?
- Evaluation measures (we already have many)

What we often do:

- Step 1: Publish Paper with Baseline + (confidential) Dataset
- Step 2: Run shared task
How you could set it up

What would you need?

- A (small) public dev/train dataset
- A (private) test dataset
- A dockerized baseline
- A Tutorial: How to run the baseline on dev?
- Evaluation measures (we already have many)

What we often do:

- Step 1: Publish Paper with Baseline + (confidential) Dataset
- Step 2: Run shared task

- We often add this to proposals
- We combine this with teaching:
  - IR course in Leipzig
  - NLP course in Jena and Weimar
Conclusion

TIRA allows experiments / shared tasks on confidential data with software submissions

- Improved Reproducibility
- Blinded Experimentation

Better benefit/effort ratio than previous approaches for shared tasks?

- One software submission, evaluation on many datasets
- Evaluate on datasets to which you don't have access
Conclusion

TIRA allows experiments / shared tasks on confidential data with software submissions

- Improved Reproducibility
- Blinded Experimentation

Better benefit/effort ratio then previous approaches for shared tasks?

- One software submission, evaluation on many datasets
- Evaluate on datasets to which you don't have access

Future Work

- Shared tasks over multiple university courses
  - Currently in discussion with 5 courses
Conclusion

TIRA allows experiments / shared tasks on confidential data with software submissions

- Improved Reproducibility
- Blinded Experimentation

Better benefit/effort ratio then previous approaches for shared tasks?

- One software submission, evaluation on many datasets
- Evaluate on datasets to which you dont have access

Future Work

- Shared tasks over multiple university courses
  - Currently in discussion with 5 courses

github.com/tira-io/tira

Thank You!
Example: TIREx
Example: TIREx

TIREx does “one thing”: Integrate Existing Tools

TIRA

- Reproducible shared tasks: Software submissions + blinded experiments

ir_datasets

- Unified + random data access: Documents + queries + rel. Judgments

PyTerrier

- Declarative reproducibility pipelines
TIREx: Overview

- Organizer provides (private) docker image with ir_datasets integration
- Participants provide docker images with retrieval approaches

Covers a shared task end-to-end
TIREx: Feasibility Study

50 Transferrable Retrieval Models in TIRA

- Derived from tira-starters from 4 starters
- Retrieve against default text in ir_datasets
- Selecting suitable baseline → improves internal validity
- Diversification of pools for shared tasks with few participants

<table>
<thead>
<tr>
<th>Framework</th>
<th>Type</th>
<th>Description</th>
<th>Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEIR [78]</td>
<td>Bi-Encoder</td>
<td>Dense Retrieval</td>
<td>17</td>
</tr>
<tr>
<td>ChatNoir [7]</td>
<td>BM25F Retrieval</td>
<td>Elasticsearch Cluster</td>
<td>1</td>
</tr>
<tr>
<td>ColBERT@PT [55]</td>
<td>Late Interaction</td>
<td>Pyterrier Plugin</td>
<td>1</td>
</tr>
<tr>
<td>DuoT5@PT [71]</td>
<td>Cross-Encoder</td>
<td>Pairwise Transformer</td>
<td>3</td>
</tr>
<tr>
<td>PyGaggle [59]</td>
<td>Cross-Encoder</td>
<td>Pointwise Transformer</td>
<td>8</td>
</tr>
<tr>
<td>PyTerrier [64]</td>
<td>Lexical</td>
<td>Traditional Baselines</td>
<td>20</td>
</tr>
</tbody>
</table>

∑ = 6 = 4 frameworks + 2 forks

50
TIREx: Feasibility Study

32 Exchangeable Benchmarks in TIRA

- Models can be transferred to new corpora ⇒ improves external validity

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Docs.</th>
<th>Size</th>
<th>Included Benchmarks</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name</td>
<td>Size</td>
<td>Details</td>
<td>Details</td>
</tr>
<tr>
<td>Argueme</td>
<td>0.4 m</td>
<td>8.3 GB</td>
<td>Touché 2020–2021 [9, 10]</td>
<td>2</td>
</tr>
<tr>
<td>Antique</td>
<td>0.4 m</td>
<td>90.0 MB</td>
<td>QA Benchmark [47]</td>
<td>1</td>
</tr>
<tr>
<td>ClueWeb09</td>
<td>1.0 b</td>
<td>4.0 TB</td>
<td>Web Tracks 2009–2012 [22–25]</td>
<td>4</td>
</tr>
<tr>
<td>ClueWeb12</td>
<td>731.7 m</td>
<td>4.5 TB</td>
<td>Web Tracks [29, 30], Touche [9, 10]</td>
<td>4</td>
</tr>
<tr>
<td>ClueWeb22B</td>
<td>200.0 m</td>
<td>6.8 TB</td>
<td>Touché 2023 [8] (ongoing)</td>
<td>1</td>
</tr>
<tr>
<td>CORD-19</td>
<td>0.2 m</td>
<td>7.1 GB</td>
<td>TREC-COVID [85, 90]</td>
<td>1</td>
</tr>
<tr>
<td>Cranfield</td>
<td>1,400</td>
<td>0.5 MB</td>
<td>Fully Judged Corpus [27, 28]</td>
<td>1</td>
</tr>
<tr>
<td>Disks4+5</td>
<td>0.5 m</td>
<td>602.5 GB</td>
<td>TREC-7/8 [87, 88], Robust04 [81, 82]</td>
<td>3</td>
</tr>
<tr>
<td>Gov2</td>
<td>25.2 m</td>
<td>87.1 GB</td>
<td>TREC TB 2004–2006 [18, 21, 26]</td>
<td>3</td>
</tr>
<tr>
<td>Medline</td>
<td>3.7 m</td>
<td>5.1 GB</td>
<td>Trec Genomics [48, 49], PM [73, 74]</td>
<td>4</td>
</tr>
<tr>
<td>MS MARCO</td>
<td>8.8 m</td>
<td>2.9 GB</td>
<td>Deep Learning 2019–2020 [35, 36]</td>
<td>2</td>
</tr>
<tr>
<td>NFCorpus</td>
<td>3,633</td>
<td>30.0 MB</td>
<td>Medical LTR Benchmark [12]</td>
<td>1</td>
</tr>
<tr>
<td>Vaswani</td>
<td>11,429</td>
<td>2.1 MB</td>
<td>Scientific Abstracts</td>
<td>1</td>
</tr>
<tr>
<td>WaPo</td>
<td>0.6 m</td>
<td>1.6 GB</td>
<td>Core 2018</td>
<td>1</td>
</tr>
</tbody>
</table>

Σ = 15 corpora 1.9 b 15.3 TB 32
TIREx: Feasibility Study
Initial Leaderboards: 1600 runs

- Running all 50 models on all benchmarks took 1 Week
- See https://github.com/tira-io/ir-experiment-platform
- Additional use-cases: LTR, QPP, etc.

Teaser of results:

- Observe system preferences on TREC DL 2019
- Use repro_eval to measure the proportion of reproducible preferences

[Breuer'20,Breuer'21]

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Rank</th>
<th>Succ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC DL 2020</td>
<td>1</td>
<td>85.2</td>
</tr>
<tr>
<td>Touché 20 (Task 2)</td>
<td>2</td>
<td>81.0</td>
</tr>
<tr>
<td>Touché 21 (Task 2)</td>
<td>3</td>
<td>72.6</td>
</tr>
<tr>
<td>Web Track 2004</td>
<td>4</td>
<td>72.1</td>
</tr>
<tr>
<td>CORD-19</td>
<td>5</td>
<td>70.0</td>
</tr>
<tr>
<td>Terabyte 2006</td>
<td>10</td>
<td>62.1</td>
</tr>
<tr>
<td>TREC PM 2017</td>
<td>15</td>
<td>53.4</td>
</tr>
<tr>
<td>Terabyte 2005</td>
<td>20</td>
<td>42.2</td>
</tr>
<tr>
<td>TREC PM 2018</td>
<td>25</td>
<td>33.2</td>
</tr>
<tr>
<td>Cranfield</td>
<td>30</td>
<td>28.8</td>
</tr>
</tbody>
</table>
Human Value Detection Demo

Demo for the Adam Smith human value detector by Schroter et al. (2023) [paper under review], which performed best in the ValueEval’23 competition, an ensemble of three models that performed best in the ablation tests. [code: original, docker image, server docker image]

Enter an argument in the text area and click on submit. After a few seconds, the detected value categories will be highlighted in the value table.

Speed limits should be abandoned.
We should allow gay marriage
Backup: Limitations

- Computational resources.
  Potential Solution:
  - Hybrid submissions: Run upload, Software submission only for plausibility checks
  - OSF infrastructure

- How to avoid big ensembles?
- Evaluation measures required that combine efficiency with effectiveness?
- New iteration of the IRF?
Backup: Use in Teaching

- Cover the “full cycle” with students in IR exercises?
  - We do this next term
Backup: Definition of Multi-Stage Software

Figure 3: The definition of a full-rank retrieval software in TIRA that consists of two modularized components.
pipeline = tira.pt.retriever(
    '<task-name>/<user-name>/software',
    dataset
)
advanced_pipeline = pipeline >> advanced_reranker

Listing 1: Full-Rank Retrieval from a complete corpus.
Backup: Load Submissions

```python
first_stage = tira.pt.from_submission(
    '<task-name>/<user-name>/<software>',
    dataset='<dataset>'
)
advanced_pipeline = first_stage >> advanced_reranker

Listing 3: Re-Rank a run created by a software submission.