A Platform for Aligning Classroom Assessments to Job Postings

PR#800

Who Are We?

Ram Dantu

Professor, Director of Center for:

Information and Cyber Security at The University of North Texas

Tyler Parks

Principal Researcher,
CS Master's Graduate at The University of North Texas

We sincerely acknowledge and thank the ***National Centers of Academic Excellence in Cybersecurity****, housed in the* ***Division of Cybersecurity Education, Innovation and Outreach****, at the* ***National Security Agency (NSA)*** *for partially supporting our research through grants H98230-20-1-0329, H98230-20-1-0414, H98230-21-1-0262, H98230-21-1-0262, and H98230-22-1-0329.*

Abstract

Proposed tool will provide users with a platform

- Access a side-by-side comparison of skills between
  - Classroom assessments
  - Job postings
  - Other text volumes (resumes, etc.)

Using techniques and methodologies from **NLP**, **Machine Learning**, **Data Analysis**, and **Data Mining**, the employed algorithm:

1. Analyzes job postings and classroom assessments
2. extracts and classifies skill units within
3. compares sets of skills from different input volumes

This tool describes the alignment between:

- 50 UNT assessments
- 5,000 industry and federal job postings

This comparison demonstrate a compatibility of 75.5%; and, that this measure was calculated using a tool operating at an 82% precision rate.

Introduction

Audience and Motivation

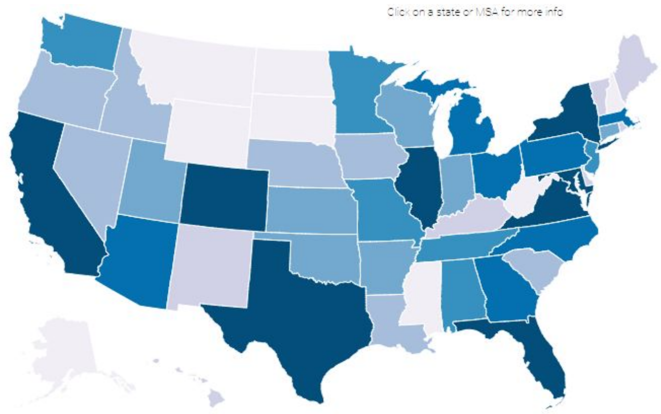
Recent events occurring in today's job market have demonstrated:

- **Mismatch** between job seekers and employers
- Increasing **talent gap** in Cybersecurity and Computer Science

Position vacancy is one such example of a skills gap in the job market. As industry roles and job markets change, so will the curriculum in academia.

CyberSeek's supply and demand heat map shows in the United States:

- **700,000 unfilled cybersecurity positions**
- …out of 1.8 million total cybersecurity jobs nationwide



What's the Problem?

…and how to solve it.

There exists no widely-used **golden standard** set of skills, knowledge, or experience that is used by both academia and industry.

- Lack of transferability
  - Learned material from university
  - Applicable skills in-industry

Without a standard of skill and knowledge items or a meaningful way to provide **transparency** between industry and academia, the entire hiring process succumbs to subjectivity.

This eventually leads to a rise in other types of job recruitment, i.e. **networking**.

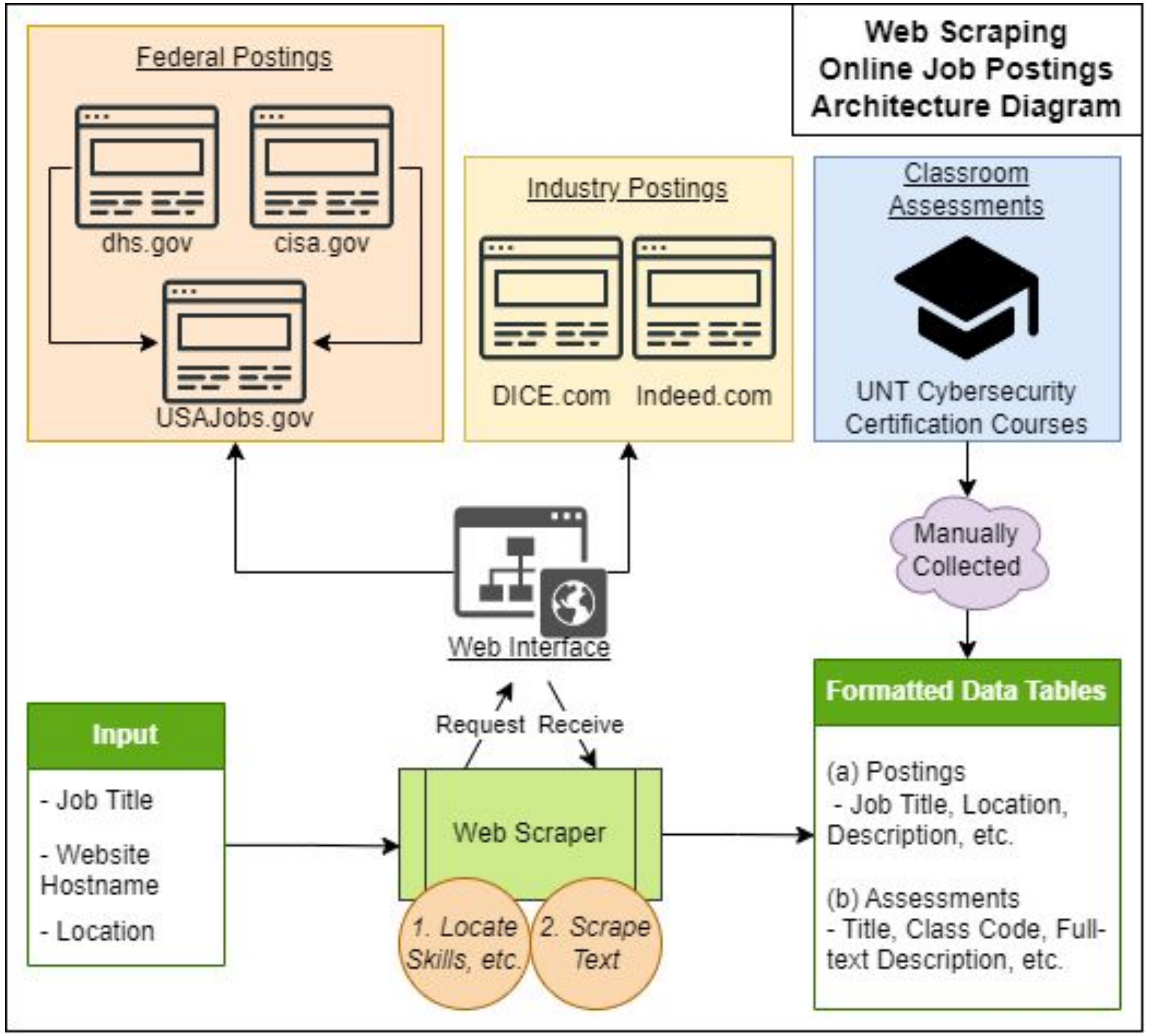
Data Collection

via Web Scraping

A web scraper will collect data features of selected job postings and store that information:

- Using **user defined input**
- Within a formatted data table
  - Comma-separated value file
  - Panda's Dataframe

Additionally, manually-collected assessments are stored the same way.



Data Collection

cont.

Analysis Type	Total Samples	Contain ULs	Ratio of Contained ULs	# of Skills, inside ULs	# of Skills, no ULs	Skill Loss (%)	% of Available Skills Collected
Manual	20	16	0.8000	176	16	8.33	**73.34**
[redacted]							
Scripted	333	296	0.8889	3256	148	4.35	**85.02**
Scripted	1197	1057	0.8830	11627	560	4.60	**84.24**
Scripted	4505	3966	0.8804	43626	2156	4.71	**83.89**

Approximately **88%** of tested postings contain at least 1 bulleted list, with an average of 6 bulleted lists per those same postings.

If only those postings' bulleted lists are scraped:

- **5%** of available skills will go uncollected (loss)
- Overall scraping accuracy of roughly **84%** across all tests

Data Collection

cont.

Webpage Section	Number of Skill Phrases Found	Percent of Total Skills Collected (%)	Avg. Skills per Posting
Duties	1053	**57.10**	5.27
Qualifications	721	**39.10**	3.61
4 Other Sections	70	3.80	0.35
Totals	1844	**96.20**	9.22

The location of skill and knowledge units across the Duties, Qualifications, and other sections of USAJobs.gov job postings. From the analysis, we found that:

- Duties Section - **57.10%**
- Qualifications Section - **39.10%**

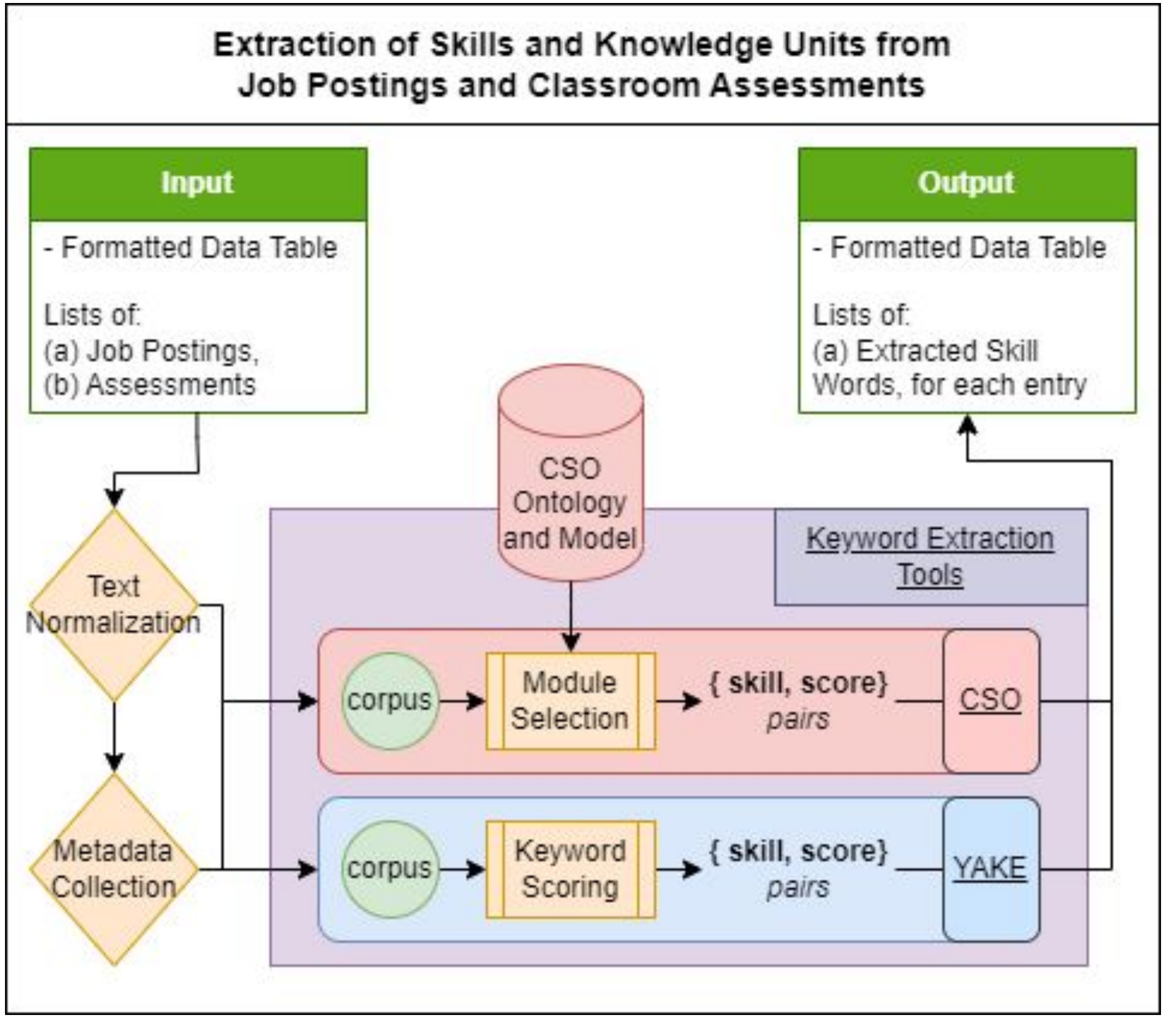
…of available skills over that posting -- summing to **96.20%**

Skill Identification

using Keyword Extraction

The end-to-end architecture of skill and knowledge extraction:

1. Receives formatted table of data
2. Each full-text description is parsed using keyword extraction tools:
   a. **cso-classifier**
   b. **YAKE**
3. The resultant output contains a list of extracted skill words/phrases for each sample processed



Skill Identification

cont.

Corpus Type	Corpus Entries	True Positives	False Positives	False Negatives	Avg Precision	Avg Recall	Avg F1-Score
Assessments-1	20	561	106	26	0.8411	0.9557	**0.8802**
Assessments-2	50	1407	253	103	0.8476	0.9318	**0.8786**
Federal Job Postings	200	5589	2418	449	0.6980	0.9256	**0.7815**
Industry Job Postings-1	2000	62757	21985	4049	0.7406	0.9394	**0.8191**
Industry Job Postings-2	4000	132054	45334	8414	0.7444	0.9401	**0.8308**

- Assessments
  - **87.86%**
  - **88.02%**
- Federal, Industry job postings
  - **78.15%**
  - **81.91%**
  - **83.08%**

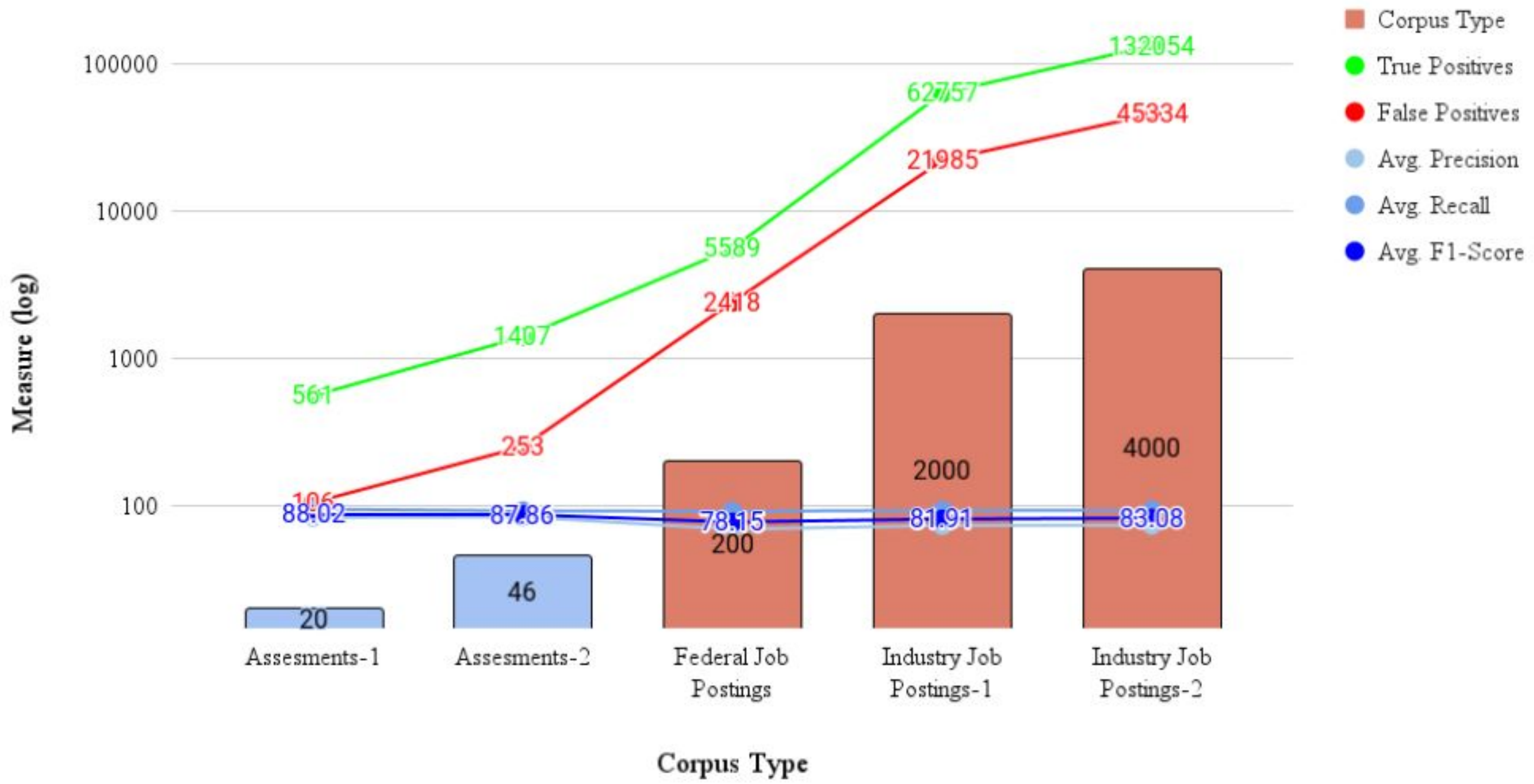
From each test, the cardinality of TPs, FPs, and FNs, are used to calculate each test's average accuracy.

- Why are accuracy results separated by nearly 7%?

Skill Identification

cont.

Composite Results: Gap Analysis between True and False Positives vs. Average Accuracy and other Metrics

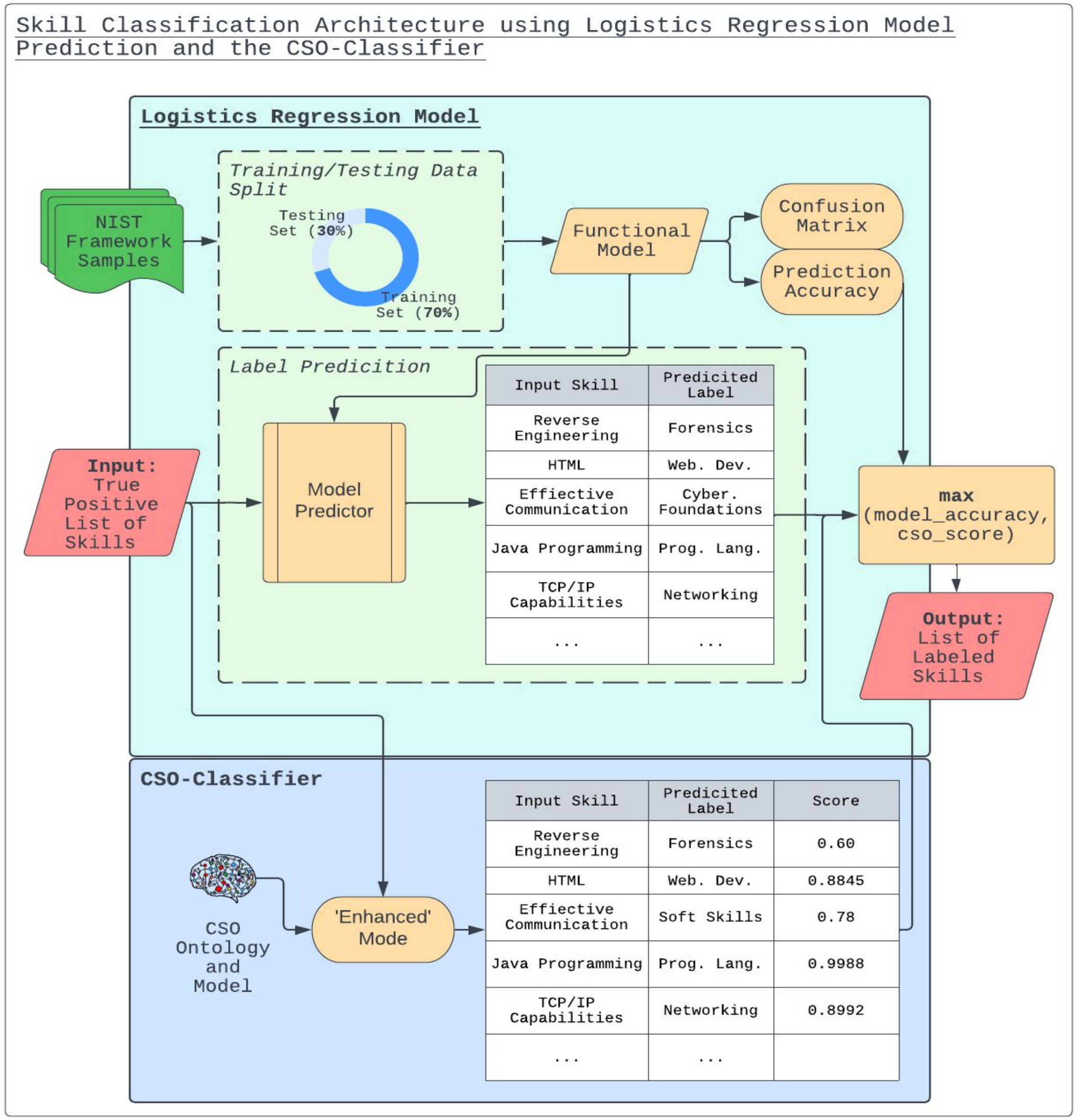


A visual representation of corpus processing data.

- Gaps between TPs and FPs are visually closer for **job postings**
- results in **lower** accuracy scores compared to **assessments**

Skill Clustering

with Logistic Regression



The end-to-end architecture of skill classification and outcome deduction.

Each skill is processed by being:

1. Passed through a **custom classifier**
   a. **Categorized** using a trained logistic regression model

2. Passed through cso-classifier's **Enhanced Module**

Skill Clustering

cont.

Over the course this research, our language model's accuracy has improved from:

- **3.0%...**
  - Custom labeling from scratch
- **... to 76.0%**
  - Gradual completion of sample mapping
  - Improvements in model optimizations

Metric	Precision	Recall	F1_Score	Support
accuracy	n/a	n/a	**0.76**	852
macro avg	0.78	0.75	0.75	852
weighted avg	0.78	0.76	0.76	852

Skill Clustering

cont.

Recorded individual class accuracy information:

- Sorted on highest **F1 Score** to lowest
- **Support** and **skew** represents number of individual samples

Class Name	Precision	Recall	F1 Score	Support	Skew Factor
Networking	0.97	0.88	**0.92**	73	4.29
Communication Architecture and Security	0.90	0.92	**0.91**	61	3.59
Database Systems	0.92	0.86	**0.89**	28	1.65
Analysis	0.84	0.81	**0.82**	77	4.53
Intrusion Detection Systems	0.90	0.75	**0.82**	12	0.71
Data Mining	0.86	0.76	**0.81**	66	3.88
Cyber Threats	0.84	0.77	**0.80**	113	6.65
Encryption and Cryptography	0.73	0.89	**0.80**	9	0.53
Operating Systems	1.00	0.64	**0.78**	14	0.82
Education/Teaching	0.82	0.73	**0.77**	44	2.59
Cyber Crime and Law	0.80	0.65	**0.72**	66	3.88
Disaster Management	0.75	0.69	**0.72**	13	0.76
Information Technology and Forensics	0.74	0.64	**0.69**	86	5.06
Data Security	0.61	0.73	**0.67**	52	3.06
Programming Fundamentals	0.51	0.86	**0.64**	59	3.47
Software Development	0.56	0.63	**0.59**	62	3.65
Cybersecurity Foundations	0.43	0.53	**0.47**	17	1.00

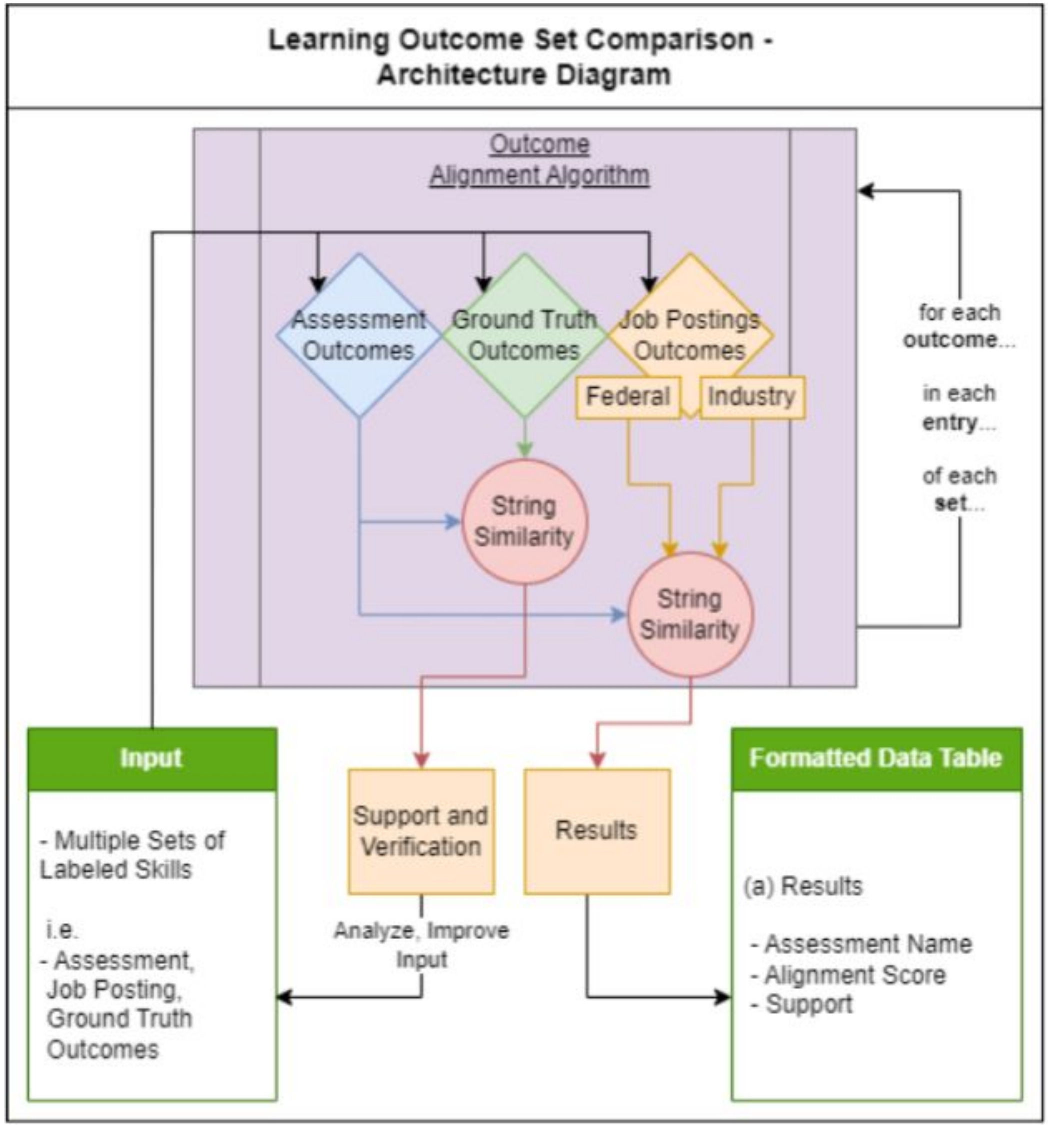
Learning Outcome Alignment

between input volumes

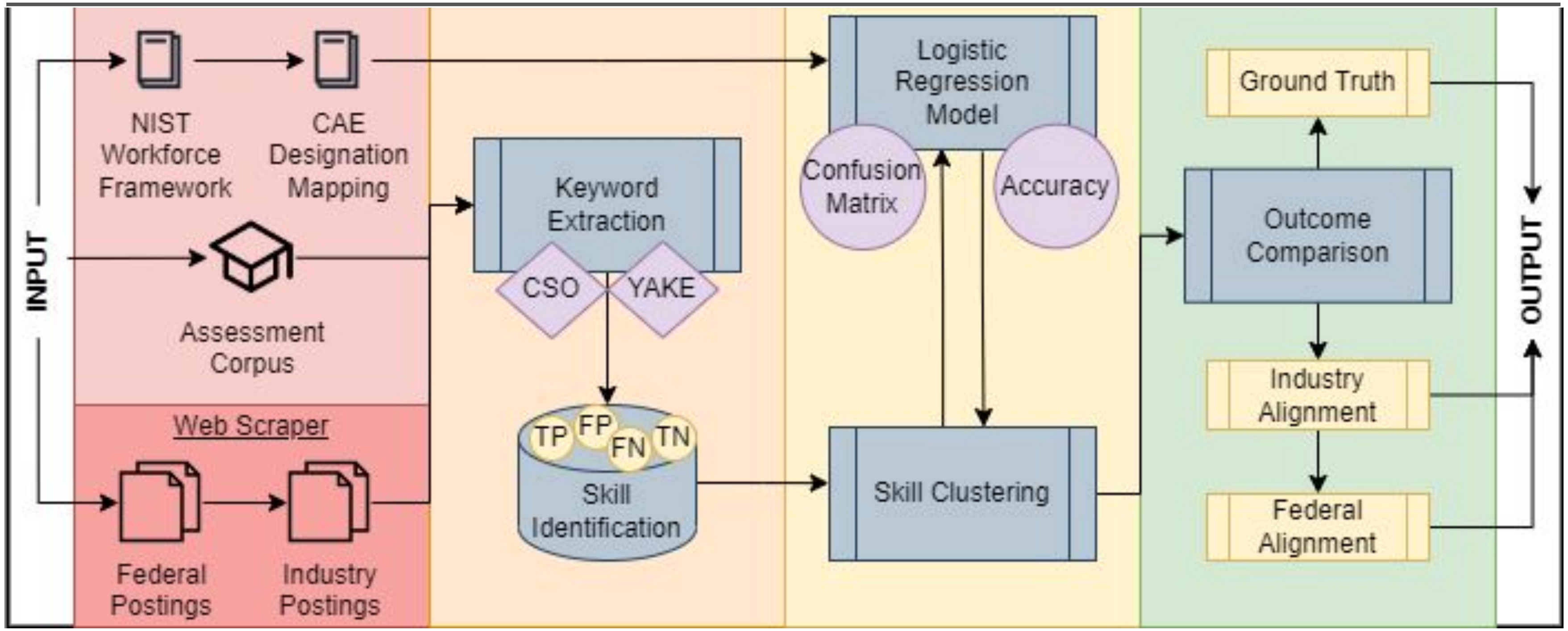
The end-to-end architecture of assessment alignment score between ground-truth and job postings.

Using outcomes generated from each corpus type:

- Calculate the **alignment of outcomes** between:
  - Assessment set
  - Ground-truth set
  - Job posting set
- This is performed by string similarity averaging



Overall Architecture



Overall Architecture of the Platform across each step

- Data Collection
- Keyword Extraction
- Skill Clustering
- Outcome Comparisons

What did we Discover?

Measure Description	Sub-measure	High Value (%)	Avg. Value (%)	Low Value (%)
Federal Job Posting Collection	USAJobs.gov, others	-	96.20	-
Industry Skill Collection	Indeed.com, DICE.com	85.02	82.25	73.34
Keyword Extraction Tool Validation	CSO-Classifier	-	86.74	-
	YAKE	-	83.87	-
	CSO + YAKE	-	87.00	-
Skill Extraction Accuracy	Classroom Assessments	-	87.94	-
	Federal Job Postings	-	78.15	-
	Industry Job Posting	-	82.50	-
Logistic Regression Model Prediction	-	-	76.00	-
Training Data Validation	Most, Least Accurate Classes	92.00	76.00	47.00
Learning Outcome Alignment Score	Assessment vs. Ground Truth	77.85	68.78	58.79
	Assessment vs. Federal Job Postings	88.92	76.76	62.14
	Assessment vs. Industry Job Postings	86.48	75.02	59.14

CAE IN CYBERSECURITY COMMUNITY

Outlier Values:

1. 96% **skill collection accuracy** on federal job postings
2. Individual **class accuracy**
   a. High of 92%
   b. Low of 47%
3. Low minimum value for all **outcome comparisons**
   a. 58.79% -> 62.14%

Conclusion, Discussion

These averages all converge at approximately **82%**, producing a single encompassing metric.

Across step:

- Outcome Alignment: **73.5%**
- Skill Classification: **76.0%**
- Skill Extraction: **82.9%**
- Web Scraping: **89.2%**

Across input type:

- Federal Job Postings: **82.8%**
- Industry Job Postings: **80.6%**
- Assessment, Ground Truth: **79.9%**

Corpus Type	Outcome Alignment	Classifier Accuracy	Corpus Extraction Accuracy	Extraction Tool Accuracy	Corpus Scraping Accuracy	Averages	Uncertainty
Federal	0.7676	0.7600	0.7815	0.8700	0.9620	**0.8282**	0.0669
Industry	0.7502	0.7600	0.8250	0.8700	0.8225	**0.8055**	0.0975
Assessm.	0.6878	0.7600	0.8794	0.8700	n/a	**0.7993**	0.0953
Averages	**0.7352**	**0.7600**	**0.8286**	**0.8700**	**0.8923**	**0.8172**	

Thank you for your time!

We’ll use any remaining time for questions.