RISE-NLP: Interactive Residual Distribution Diagnostics for Fairness Auditing in Text Classification

Herbert Wertheim College of Engineering
Department of Computer & Information Science & Engineering
UNIVERSITY of FLORIDA

Ray Chen • Kevin Kurian • Christan Grant
University of Florida

EMNLP2026 BUDAPEST OCTOBER 24-29

Why this demo?

Scalar fairness metrics can summarize group-level differences, but they often do not show where subgroup errors are concentrated. RISE-NLP helps auditors see where the distributions separate and inspect the underlying text examples.

Project Website | Ray's LinkedIn | Demo Website

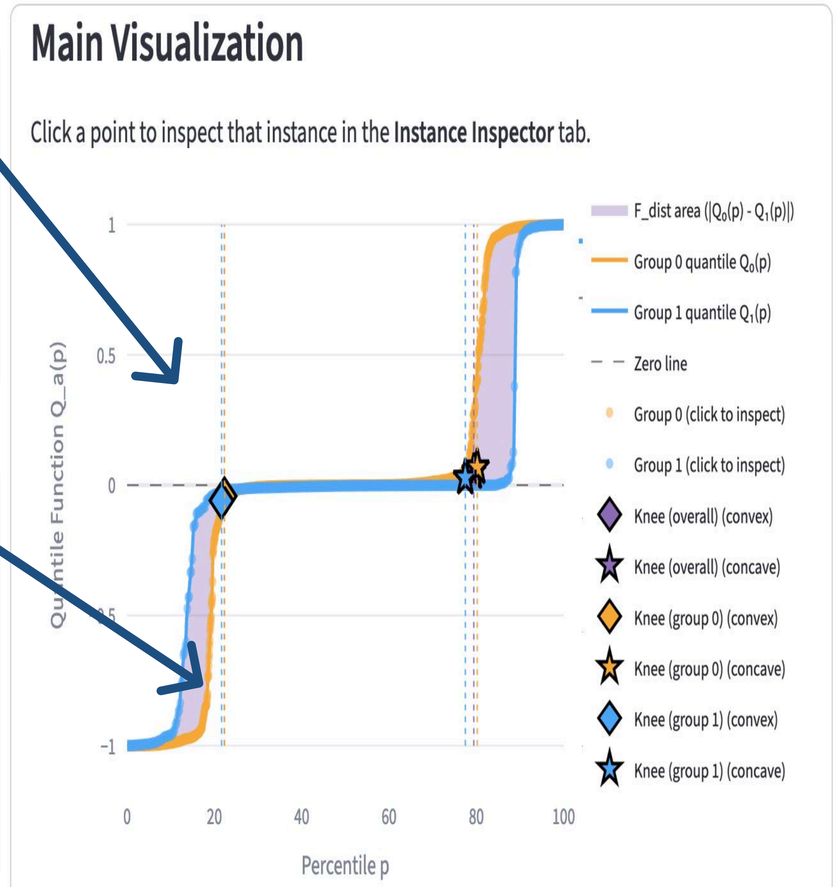
Email: chenz1@ufl.edu

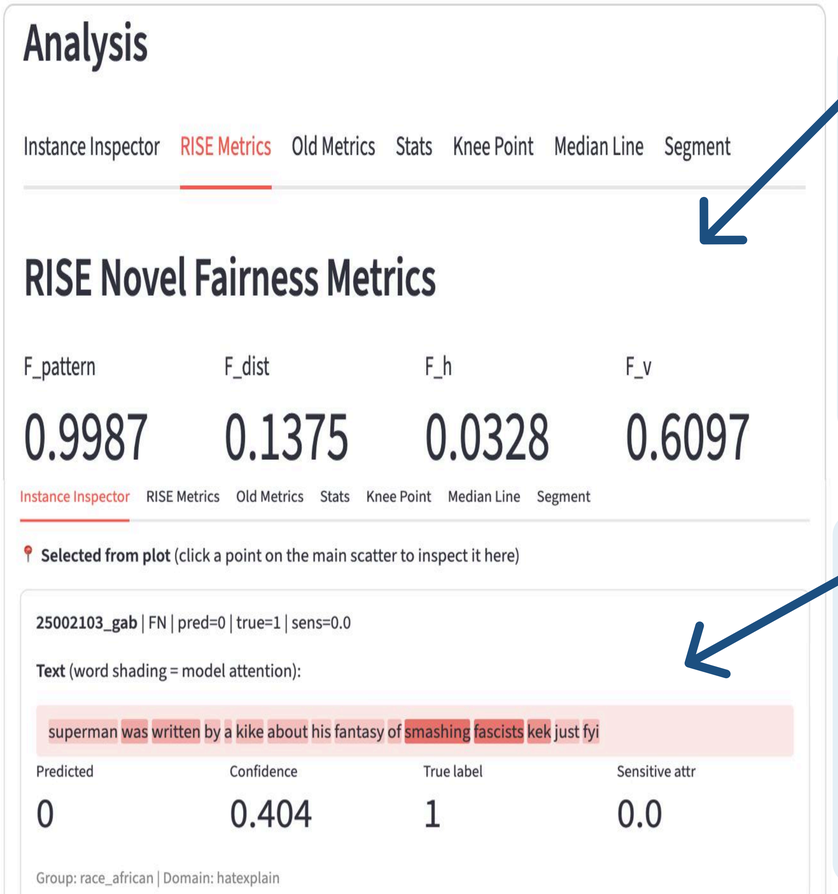
What users see

INTERACTIVE · VISUAL · EXAMPLE-DRIVEN FAIRNESS ANALYSIS

Residual curve visualization
Compare quantile functions across groups and see where they separate.

Interactive subgroup comparison
View group-specific curves, quantiles, knee points, and distributional differences.





Metric summary panel
View distributional fairness metrics and key statistics for the selected region or instance.

Click a region to inspect examples
Select points or percentile regions to view concrete examples and model predictions.

FROM PATTERNS TO EXAMPLES

11790924544920494O8_twitter | TN | pred=0 | true=0 | sens=0.0

Text (word shading = model attention; darker = more impact on prediction):

for the queer readers you deserve every adventure nina varela baby i already love u so much

Predicted	Confidence	True label	Sensitive attr
0	0.013	0	0.0

Group: race_african | Domain: hatexplain

16501013_gab | FN | pred=0 | true=1 | sens=1.0

Text (word shading = model attention; darker = more impact on prediction):

why should i want my descendant future tied to somali immigrants or their children how is that in their best interest

Predicted	Confidence	True label	Sensitive attr
0	0.004	1	1.0

Group: race_african | Domain: hatexplain

5840737_gab | TP | pred=1 | true=1 | sens=0.0

Text (word shading = model attention; darker = more impact on prediction):

lol why why should i move just becos there is nazi morons here in germany

Predicted	Confidence	True label	Sensitive attr
1	0.998	1	0.0

Group: race_african | Domain: hatexplain

11932935221425479G8_twitter | FP | pred=1 | true=0 | sens=0.0

Text (word shading = model attention; darker = more impact on prediction):

<user> i hate you

Predicted	Confidence	True label	Sensitive attr
1	0.990	0	0.0

Group: race_african | Domain: hatexplain

True Positive pred=1, true=1 | True Negative pred=0, true=0 | False Positive pred=1, true=0 | False Negative pred=0, true=1

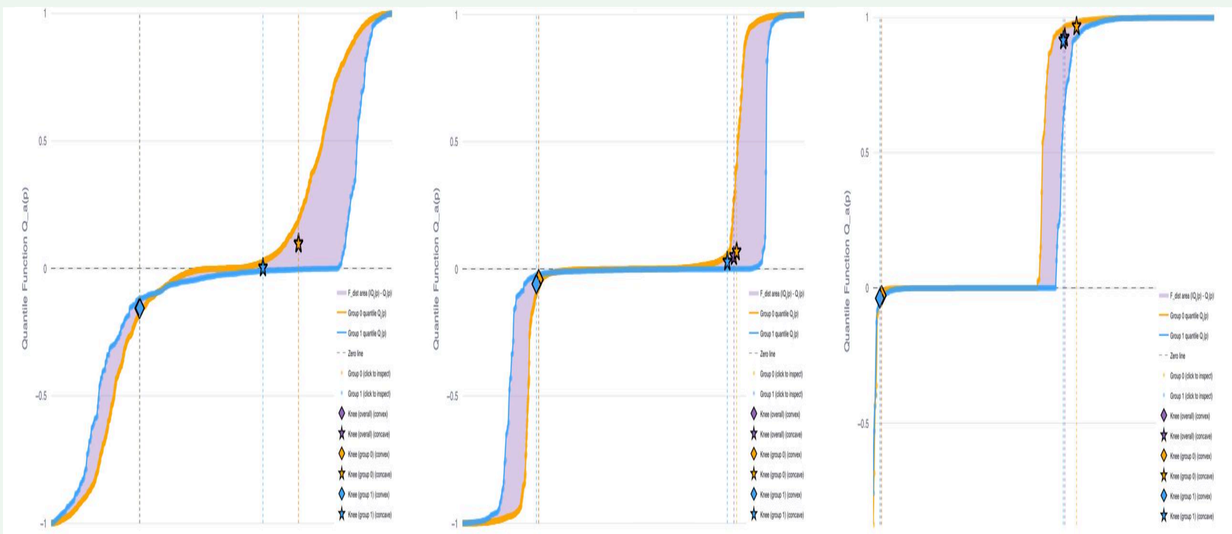
Auditors can inspect true positives, true negatives, false positives, and false negatives, together with model confidence and highlighted text evidence.

HOW THE DEMO WORKS

1. Load predictions and subgroup metadata
2. Compare residual curves across groups
3. Select separated percentile regions
4. Inspect text behind the visulization

VISUALIZATION

DIFFERENT DATASETS, DIFFERENT FAIRNESS BEHAVIORS



Model	Data	$\mathcal{F}_{\text{dist}}$	MD	DP	EO
BERT	SST-2 random control	0.058	0.001	0.998	1.001
	HateXplain African target	0.150	0.242	1.500	1.298
RoBERTa	HateXplain African target	0.137	0.294	1.580	1.340

Identity-linked toxicity settings show larger residual gaps than the random-group control.

USER STUDY

- 16 participants
- 10 said they would use the system
- 12 said the interface helped them reason about fairness–accuracy trade-offs
- Common feedback: clearer metric definitions and a less dense layout.

TAKEAWAY

RISE-NLP helps users move from aggregate metrics to localized residual patterns and then to the concrete examples that drive those disparities.