

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

Ray Zeyao Chen¹, Kevin Kurian¹, and Christan Grant¹

¹University of Florida, Gainesville, FL, USA

{chenz1, kuriankevin, christan}@ufl.edu

Abstract

Fairness metrics such as demographic parity and equalized odds summarize group-level behavior, while the location of subgroup errors in the score distribution often remains unclear. RISE-NLP is a diagnostic interface for auditing probabilistic text classifiers through signed residual distributions. Given prediction scores, labels, group metadata, and text, RISE-NLP plots group residual-quantile curves, reports distributional metrics, and links selected percentile regions to the examples behind the gap. Using BERT and RoBERTa predictions, we observe $F_{\text{dist}} = 0.097\text{--}0.150$ in the two identity-linked toxicity settings and $0.055\text{--}0.058$ in a randomly grouped SST-2 control used as a descriptive reference. A 16-participant formative study evaluates users’ understanding and perceived usefulness of the residual visualizations, showing their potential to support reasoning about fairness–accuracy trade-offs. The project is available at <https://ufdatastudio.com/projects/risenlp>.

1 Introduction

Fairness auditing for text classifiers often starts with scalar group metrics. Demographic parity (DP), equalized odds (EO), accuracy, and calibration answer different questions, and prior impossibility results show that several desirable criteria cannot generally be satisfied simultaneously (Kleinberg et al., 2017; Chouldechova, 2017). Auditors often face two recurring problems: (1) **metric disagreement without a clear source** and (2) **localized errors hidden behind moderate scalar values**, especially in low-confidence regions, transition bands, or high-confidence tails.

RISE-NLP complements DP, EO, calibration, and domain-specific judgment as a threshold-free diagnostic interface for locating subgroup disparity along the residual distribution. For each target class, RISE-NLP uses the signed residual between the predicted target-class probability and the true

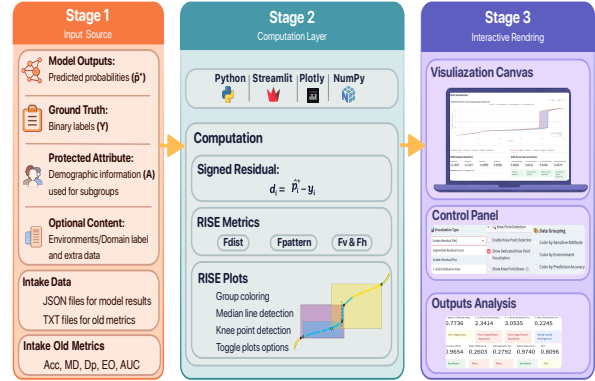


Figure 1: RISE-NLP auditing workflow and system architecture. Auditors upload or select model predictions containing text, labels, scores, group metadata, and model/dataset identifiers. The computation layer derives signed residuals, group quantile curves, scalar metrics, and distributional summaries. The interface coordinates a control panel, residual visualization, instance table, metric explanations, and report-oriented outputs so users can move from aggregate metrics to localized residual regions and then to concrete examples.

target label. Sorting these residuals within each group yields group residual-quantile curves. Curve separation shows whether disparity is concentrated in the center, near the decision boundary, or in a high-confidence tail.

In our experiments, the SST-2 random-group control has nearly overlapping residual curves and small F_{dist} values (0.055–0.058). Identity-linked toxicity settings show larger residual distribution gaps on HateXplain (0.137–0.150) and UC Berkeley hate speech data (0.097–0.131). These gaps identify percentile regions where the interface retrieves representative text examples for inspection.

Our contributions are:

- We introduce RISE-NLP as a complementary diagnostic interface for existing fairness and machine learning metrics, helping users examine model behavior with residual distributions alongside scalar scores.
- We design a visualization view that gives users an

System	NLP support	Fairness metrics	Residual view	Region localization
Fairlearn Dashboard (Bird et al., 2020)	×	✓	×	×
What-If Tool (Wexler et al., 2020)	×	Part.	×	×
AllenNLP Interpret (Gardner et al., 2018)	✓	×	×	×
AIF360 (Bellamy et al., 2018)	×	✓	×	×
FairSight (Ren et al., 2022)	Part.	✓	×	Part.
RISE-NLP	✓	✓	✓	✓

Table 1: Systems comparison focused on the novelty claim. RISE-NLP combines NLP prediction uploads, classical fairness summaries, residual distribution views, region-level disparity localization, and instance drill-down in one auditing workflow.

audit summary of subgroup disparity and model behavior, making patterns behind scalar metrics visible in the residual curves.

- We implement an interactive workflow from distribution-level subgroup disparity to the text instances that drive selected residual regions.
- We present case studies where scalar metrics give limited or confusing signals, including a random-control setting with little residual structure, identity-linked toxicity settings with localized gaps, and cross-model comparisons where RISE-NLP helps locate the errors behind similar metric summaries.

2 Related Work

Group fairness in classification is commonly reported with criteria such as equality of opportunity (Hardt et al., 2016). Prior work shows that calibration, error-rate balance, and related fairness criteria can conflict across settings and operating points (Kleinberg et al., 2017; Chouldechova, 2017; Corbett-Davies and Goel, 2018). Toolkits such as AI Fairness 360 (Bellamy et al., 2018) and Fairlearn (Bird et al., 2020) support metric reporting and mitigation workflows. What-If Tool and FairSight further support interactive exploration (Wexler et al., 2020; Ren et al., 2022), while AllenNLP Interpret focuses on NLP model interpretation (Gardner et al., 2018).

RISE-NLP builds on this line of work by linking fairness metrics with residual-quantile curves and instance inspection for text classification. The original RISE framework introduced sorted-residual visualization for fairness (Chen and Grant, 2026). RISE-NLP adapts this idea to probabilistic NLP classifiers and uses $\mathcal{F}_{\text{dist}}$, the one-dimensional Wasserstein distance between subgroup residual distributions, as the plot-aligned distributional summary. Because the demo can surface raw hate-speech examples, we follow recommendations for presenting harmful text (Kirk et al., 2022). Table 1

compares existing fairness toolkits and visualization systems with RISE-NLP along dimensions relevant to interactive auditing. RISE-NLP links distribution-level fairness patterns to concrete text instances, with scalar summaries used as context.

3 System Overview

RISE-NLP centers on the residual distribution $d = \hat{P}^+ - Y$ and its group-wise quantile functions. The interface pairs residual curves (with $\mathcal{F}_{\text{dist}}$ and $\mathcal{F}_{\text{pattern}}$) with common scalar metrics and linked instance inspection.

3.1 Signed-Error Residual Representation

Let $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^n$, where $y_i \in \{0, 1\}$ indicates membership in the target class under a one-vs-others setup and a_i denotes the group attribute used for auditing. For a classifier producing $\hat{P}_i^+ = \Pr(y = 1 \mid x_i)$, we define the signed-error residual

$$d_i = \hat{P}_i^+ - y_i \in [-1, 1], \quad (1)$$

where $d_i > 0$ indicates overestimation and $d_i < 0$ underestimation. For toxicity classification, $d_i > 0$ indicates the model over-predicts toxicity (a non-toxic comment assigned high toxicity probability), while $d_i < 0$ indicates the model under-predicts toxicity (a toxic comment assigned low toxicity probability). Large-magnitude residuals correspond to high-confidence errors (false positives or false negatives).

Let $D = \{d_1, \dots, d_n\}$ and subgroup sets $D_0 = \{d_i \mid a_i = 0\}$, $D_1 = \{d_i \mid a_i = 1\}$. Denote medians by $m_g = \text{Med}(D)$, $m_0 = \text{Med}(D_0)$, $m_1 = \text{Med}(D_1)$.

Residuals unify magnitude and direction of error in a single continuous variable, enabling threshold-free, distribution-level fairness diagnostics.

3.2 From Residuals to Interactive Geometry

Sorted residual curves. Residuals are sorted to construct quantile functions $Q_a(p)$ over percentile

$p \in [0, 1]$ for each group a . Plotting $Q_0(p)$ and $Q_1(p)$ yields sorted residual curves. Curve separation indicates different residual distributions; overlap indicates similar residual distributions.

Sorted residual curves are often S-shaped, with steep transitions where predictions switch from mostly correct to mostly incorrect. RISE-NLP marks knee points using a curvature-based heuristic (Kneedle (Satopaa et al., 2011)) to help users navigate these transition regions. Knee markers are visual aids for navigation and percentile selection for both individual groups and overall distribution; distributional disparity is summarized by $\mathcal{F}_{\text{dist}}$.

3.3 Interface and Interaction Design

RISE-NLP provides an interactive workflow that links distribution-level diagnostics to instance-level inspection. The interface synchronizes three components: configuration controls, residual-distribution visualization, and example inspection.

Control panel. Users upload a CSV file or select a dataset–model pair, define group attributes, optionally apply probability calibration (temperature scaling (Guo et al., 2017)), and filter subsets or metadata. Each interaction triggers recomputation of residual summaries and updates the views, enabling rapid comparison across models and groups through the interface.

Visualization. The central view displays group-wise RISE curves. The horizontal axis represents percentile rank within each group, and the vertical axis represents signed residual magnitude. Separation between curves indicates differences in error distributions across groups. The shaded region between curves corresponds to $\mathcal{F}_{\text{dist}}$, providing a visual summary of distributional disparity aligned with the plotted geometry. Median markers and knee annotations highlight structural transitions in prediction behavior. Each displayed score includes a one-sentence tooltip: $\mathcal{F}_{\text{dist}}$ is the area between group residual curves, $\mathcal{F}_{\text{pattern}}$ is median residual alignment, and scalar metrics are thresholded group summaries. The metric panel acts as a compact audit summary. For interface readability, $\mathcal{F}_{\text{pattern}}$ and $\mathcal{F}_{\text{dist}}$ are accompanied by heuristic status labels. For $\mathcal{F}_{\text{pattern}}$, Excellent, Good, Fair, and Poor correspond to values ≥ 0.95 , ≥ 0.85 , ≥ 0.70 , and < 0.70 , respectively. For $\mathcal{F}_{\text{dist}}$, the corresponding bands are ≤ 0.10 , ≤ 0.30 , ≤ 0.60 , and > 0.60 . These labels serve as interface heuristics for quick visual comparison, with the thresholds chosen only for the dashboard display. The metric panel records

the selected model, dataset, group definition, scalar metrics, residual summaries, and selected residual region before the user opens individual examples.

Users select percentile regions directly on the residual plot to retrieve representative instances contributing to observed disparities. For each instance, the interface displays predicted probability, ground-truth label, group membership, raw text context, and the derivation rule used by any automatic explanation. The report view records the selected model, dataset, group definition, metrics, residual region, retrieved examples, and dashboard notes, supporting export of an audit report or download of selected examples. Together, these views let users move from a residual pattern to the examples that shaped it.

3.4 Generality

RISE-NLP takes prediction outputs as input and can be used with different models and datasets. For binary text classification, the audit uses the label, predicted class, target-class probability, and group attribute. For multi-class tasks, RISE-NLP uses a one-versus-others view: each class can be selected as the target class, and the residual is computed from the probability assigned to that class.

A user can apply RISE-NLP to a new dataset by uploading a prediction table with text instances, labels, group information, and probability scores. The group variable can come from dataset labels, identity targets, keyword rules, synthetic controls, or intersectional attributes, and the interface keeps this setting with the audit run for repeated comparisons. After upload, RISE-NLP computes group-wise residual quantiles, $\mathcal{F}_{\text{dist}}$, $\mathcal{F}_{\text{pattern}}$, and standard scalar metrics, then connects these values to residual curves and text instances. Figure 1 summarizes the workflow.

4 Residual Distribution Metrics

RISE-NLP reports two plot-aligned residual summaries. $\mathcal{F}_{\text{dist}}$ summarizes the area between subgroup residual-quantile curves and is used to locate distributional separation. $\mathcal{F}_{\text{pattern}}$ summarizes median residual alignment and gives a compact view of whether the typical residual is similar across groups. Both summaries are interpreted with the residual curves, scalar metrics, calibration evidence, and retrieved examples.

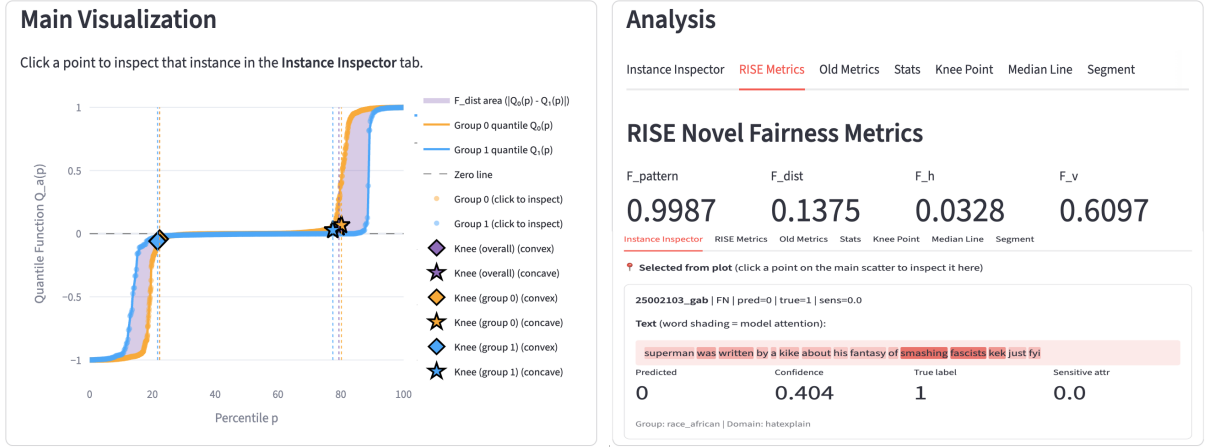


Figure 2: RISE-NLP interface screenshot with residual curves and instance views. Left: group-wise RISE curves $Q_0(p)$ (orange) and $Q_1(p)$ (blue). The purple shaded region between curves represents the distributional disparity. Diamond/Star markers indicate knee points, highlighting transition regions; dashed vertical lines mark their median. Right (top): RISE metrics. Right (bottom): Instance Inspector, which displays the selected example, including prediction, confidence, true label, sensitive attribute, and token-level attention linked to the selected point in the plot.

4.1 Distributional Distance ($\mathcal{F}_{\text{dist}}$)

$\mathcal{F}_{\text{dist}}$ corresponds to the one-dimensional Wasserstein-1 (W_1) distance between subgroup residual distributions. For distributions on the real line with finite first moments, W_1 admits the equivalent quantile representation

$$W_1(P, Q) = \int_0^1 |Q_0(p) - Q_1(p)| dp, \quad (2)$$

which equals the area between subgroup quantile curves (Villani, 2009). In the interface, this value appears as the shaded area between the sorted residual curves.

The key use of $\mathcal{F}_{\text{dist}}$ is localization. A large value prompts the auditor to inspect the percentile bands contributing most of the area; a small value suggests that the group residual distributions are similar under the selected group definition and subset. Because $\mathcal{F}_{\text{dist}}$ is threshold-free, it can expose residual gaps even when thresholded scalar metrics appear similar. Residual gaps still require task context and policy goals for fairness interpretation.

4.2 Median Alignment ($\mathcal{F}_{\text{pattern}}$)

Let m_a denote the median residual for group a .

$$\mathcal{F}_{\text{pattern}} = 1 - \frac{|m_1 - m_0|}{2}. \quad (3)$$

$\mathcal{F}_{\text{pattern}}$ measures whether the typical residual is aligned across groups. It intentionally captures only central tendency. Auditors should therefore read it as a complement to the residual curves, $\mathcal{F}_{\text{dist}}$, and tail inspection.

Analysis

Instance Inspector RISE Metrics Old Metrics Stats Knee Point Median Line Segment

RISE Novel Fairness Metrics

F_pattern F_dist F_h F_v
0.9987 0.1375 0.0328 0.6097

Instance Inspector RISE Metrics Old Metrics Stats Knee Point Median Line Segment

Selected from plot (click a point on the main scatter to inspect it here)

25002103_gab | FN | pred=0 | true=1 | sens=0.0

Text (word shading = model attention):

superman was written by a kike about his fantasy of smashing fascists kek just fyi

Predicted	Confidence	True label	Sensitive attr
0	0.404	1	0.0

Group: race_african | Domain: hateexplain

4.3 Knee-Point Annotations

The interface optionally annotates knee points on the sorted residual curves using a curvature-based heuristic (Kneedle (Satopaa et al., 2011)). Let K_ℓ and K_r denote the left and right knee points. For group $g \in \{0, 1\}$ and the pooled sample, let h_k^g denote the percentile location and d_k^g the residual value at knee $k \in \{\ell, r\}$. We summarize the group-wise knee displacement as

$$F_h = \frac{1}{2} (|H_\ell| + |H_r|), \quad H_k = \frac{h_k^1 - h_k^0}{h_k^{\text{all}}}, \quad (4)$$

and

$$F_v = \frac{1}{2} (|V_\ell| + |V_r|), \quad V_k = \frac{d_k^1 - d_k^0}{d_k^{\text{all}}}. \quad (5)$$

F_h summarizes the normalized difference in where the two group curves change curvature along the percentile axis, while F_v summarizes the corresponding difference along the residual axis. Smaller values indicate closer knee alignment, with zero corresponding to coincident group knees. These quantities serve as auxiliary geometric descriptors, while $\mathcal{F}_{\text{dist}}$ remains the primary distributional summary.

5 Case Studies and Evaluation

We evaluate RISE-NLP around a central auditing question: when does a residual-quantile view provide information that scalar group metrics leave unclear? The current evidence uses fixed predictions from BERT and RoBERTa toxicity classifiers

plus a random-group control sentiment setting. We report the demonstrated coverage separately from extensions that the interface supports but that require additional empirical validation.

5.1 Datasets and Models

Across experiments, datasets are mapped into the schema described in Section 3.4. Each prediction file contains labels, model scores, text, dataset name, model name, and a group attribute used for subgroup comparison.

SST-2 (Socher et al., 2013) is included as a random-group control. Because SST-2 lacks demographic annotations, we assign a random one-group-versus-others attribute. This setting provides a descriptive reference for residual behavior when the audit groups have no intended demographic meaning. The resulting values are specific to this dataset, model, and group construction and carry no general fairness-baseline or validation-threshold meaning for $\mathcal{F}_{\text{dist}}$. **HateXplain** (Mathew et al., 2021) provides toxicity and hate-speech examples with target annotations. We binarize labels into hate/offensive content versus normal content and use African-targeted content as the protected-attribute construction for the main toxicity audit. **UC Berkeley hate speech** (Kennedy et al., 2020) provides hate-speech scores and target metadata. We label examples as hate speech when the score exceeds 0.5 and use race-targeted content as the group attribute.

RISE-NLP audits fixed predictions from external classifiers, and we use two pretrained toxicity classifiers to produce comparable probability outputs: unitary/toxic-bert (AI, 2020; Devlin et al., 2019) and s-nlp/roberta_toxicity_classifier (Versae and contributors, 2020; Liu et al., 2019). For toxicity datasets, the target-class probability is the probability of toxicity versus other labels. For SST-2, probabilities are transformed so the residual sign has a consistent interpretation within the control task.

5.2 Case Study 1: Scalar Metrics and Residual Localization

This case study examines how scalar metrics and residual curves support different parts of a fairness audit. Table 2 summarizes the scalar and residual summaries. In the random-control SST-2 setting, DP and EO are close to parity, and $\mathcal{F}_{\text{dist}}$ is small. In the UC Berkeley setting, BERT has DP = 1.098

and EO = 0.979, while the residual distribution gap is larger ($\mathcal{F}_{\text{dist}} = 0.131$). The scalar metrics describe thresholded group rates; the RISE curves show where the residual distributions separate.

Figure 3 shows the corresponding residual geometry. The SST-2 curves nearly overlap for both models, matching the small $\mathcal{F}_{\text{dist}}$ values in Table 2. The HateXplain and UC Berkeley curves show wider shaded gaps. These gaps are localized to particular percentile regions, giving the auditor a concrete region to inspect after reading the scalar metrics.

DP and EO describe selected thresholded rates across groups. RISE-NLP points to percentile bands that deserve closer inspection. After selecting a separated band, the auditor can review the corresponding text instances and identify whether the gap comes from false positives, false negatives, or uncertainty near the decision boundary.

5.3 Case Study 2: Cross-Model Residual Comparison

This case study compares BERT and RoBERTa under the same datasets and group definitions. On UC Berkeley hate speech, RoBERTa has a lower residual distribution gap than BERT (0.097 versus 0.131), while both models have scalar ratios close to parity. The residual view supports a model-comparison audit: two models can appear similar under thresholded group metrics while distributing residual errors differently across the score range.

On HateXplain, both models have larger scalar disparities and larger residual gaps. RISE-NLP lets the auditor select the curve region that contributes to the shaded area and inspect the examples behind it. This is the role of the instance inspector shown in Figure 2 and Figure 4.

5.4 Formative User Study

We conducted a formative study with 16 participants using a survey. Participants included undergraduate students, master’s students, Ph.D. students, ML engineers, data scientists, and professors. Their ML experience ranged from less than one year to more than ten years. The study asked how users interpreted the residual visualization, which evidence they trusted when judging fairness, and which parts of the interface were unclear.

Participants generally found the visualization useful for auditing. Six participants selected visualization plots alone as the most trusted evidence for judging fairness, and five selected visualization together with the distributional summaries. Twelve

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
	UC Berkeley race target	0.131	0.057	1.098	0.979
RoBERTa	SST-2 random control	0.055	0.003	0.994	0.993
	HateXplain African target	0.137	0.294	1.580	1.340
	UC Berkeley race target	0.097	0.091	1.154	1.017

Table 2: Scalar fairness metrics and RISE residual-distribution summaries across datasets and models. MD, DP, and EO summarize group-level scalar behavior. $\mathcal{F}_{\text{dist}}$ summarizes residual-distribution separation as the area between subgroup residual-quantile curves. **Takeaway:** The SST-2 random-group control serves as a descriptive reference and shows smaller residual gaps than the two identity-linked toxicity settings in these experiments.

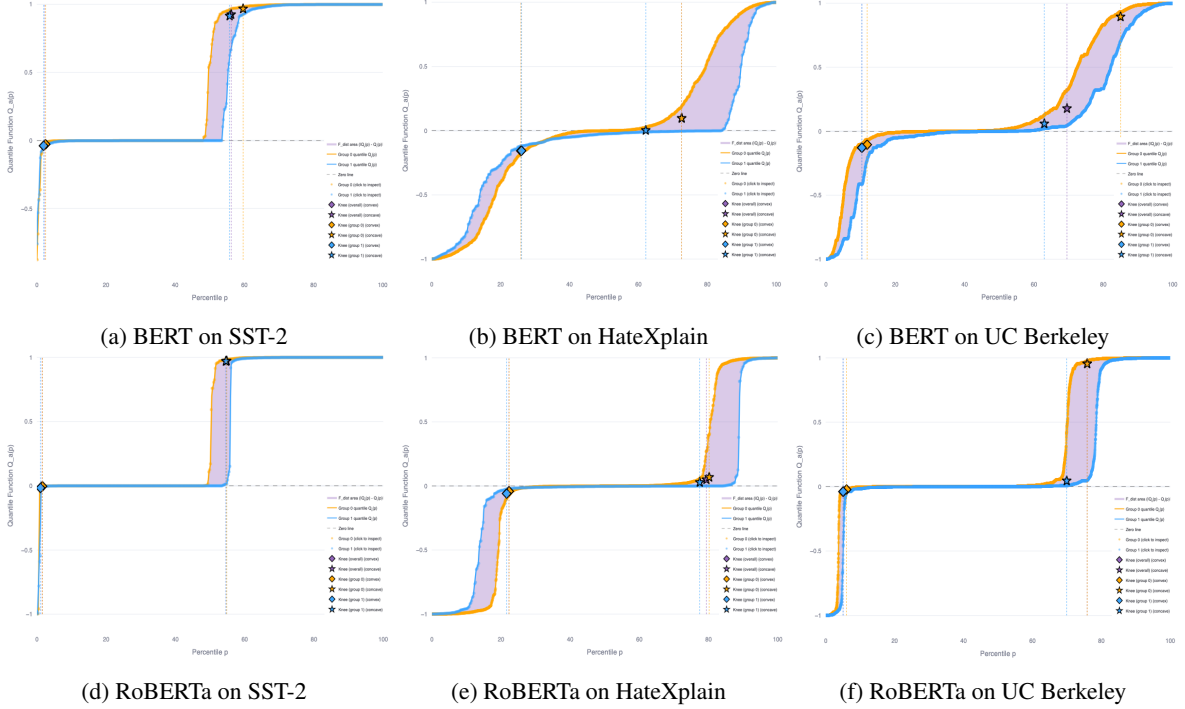


Figure 3: Residual curves for BERT and RoBERTa across the three audit settings. SST-2 random-group control shows near-overlapping curves and the smallest $\mathcal{F}_{\text{dist}}$ values in Table 2 (BERT: 0.058, RoBERTa: 0.055). HateXplain shows the largest gaps (BERT: 0.150, RoBERTa: 0.137), while UC Berkeley remains larger than the control setting (BERT: 0.131, RoBERTa: 0.097). The shaded regions show that these gaps concentrate in specific percentile bands across the curve. **Takeaway:** $\mathcal{F}_{\text{dist}}$ gives the gap size; the curve shows where to inspect.

participants said the system helped them view fairness and accuracy as a trade-off; two said it made accuracy more important, and one said it made fairness more important. For adoption, 10 participants said they would use the system, three answered “Maybe,” and three preferred traditional metrics or requested further improvements.

Open-ended responses identified useful interface elements and remaining sources of confusion. Participants often named knee markers, residual curves, horizontal disparity, and group comparison as helpful. They also pointed to visual density, metric definitions, curve interpretation, and knee detection as areas needing clearer presentation.

Takeaway. The formative study captures participants’ perceived usefulness and interpretation of the residual visualization. Participants reported using the curves to compare subgroup behavior and reason about fairness–accuracy trade-offs, while their comments point to clearer metric definitions and a less dense layout as priorities for the next interface version.

Limitations. This study is formative and focuses on self-reported perceptions of the interface, with auditing accuracy and efficiency left for future task-based evaluation. The current experiments use fixed BERT and RoBERTa predictions on one-vs-others toxicity tasks, together with an SST-2

random-group control used as a descriptive reference. RISE-NLP relies on probability scores and group metadata, so calibration quality, group construction, and dataset provenance can affect the interpretation of residual gaps. A large $\mathcal{F}_{\text{dist}}$ indicates greater residual-distribution separation under the selected audit setting, while fairness judgments still depend on task context and policy goals. The SST-2 control reflects the specific dataset, model, and random grouping used in this study and carries no general fairness-baseline or validation-threshold meaning. Future work will include controlled task-based studies, multi-class and intersectional audits, and comparisons with debiased models.

6 Conclusion

RISE-NLP supports fairness auditing by connecting scalar metrics, residual-distribution views, and instance-level inspection. The system plots group residual-quantile curves, summarizes their separation, and links selected percentile regions to concrete text examples.

Across the case studies, the SST-2 random-group control provides a descriptive reference with smaller residual gaps, while the two identity-linked toxicity settings show larger localized gaps, and BERT and RoBERTa show different residual geometry even when scalar ratios are similar. The formative study indicates that participants perceived the residual visualization as useful for reasoning about fairness–accuracy trade-offs and points to interface clarity as the next design priority. RISE-NLP is best understood as a diagnostic aid: it narrows where auditors should look, while fairness judgments still depend on task context, group construction, and policy goals.

Acknowledgments

Thanks to the National Science Foundation and Amazon for support under award number 2147375. We also thank Gatorade, Arnold, and Lisa Goldberg for their funding support.

References

- Unitary AI. 2020. Toxic bert: Pretrained toxic comment classifier. <https://huggingface.co/unitary/toxic-bert>.
- R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943*.
- Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai.
- Ray Chen and Christan Grant. 2026. Rise: Interactive visual diagnosis of fairness in machine learning models. *arXiv preprint arXiv:2602.04339*.
- Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.
- Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Si Chen, Pradeep Dasigi, Jonathan Durrett, Yoav Goldberg, Hannaneh Hauptmann, and 1 others. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.
- Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multi-task deep learning: A hate speech application. *arXiv preprint arXiv:2009.10277*.
- Hannah Rose Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. *arXiv preprint arXiv:2204.14256*. V3, 2023.
- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hateexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jiayuan Ren and 1 others. 2022. Fairsight: Visual analytics for discovering intersectional bias in text classification. In *IEEE VIS Conference*.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Versae and contributors. 2020. Roberta toxicity classifier. <https://huggingface.co/s-nlp/roberta-toxicity-classifier>. Accessed: 2026-02-24.

Cédric Villani. 2009. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jim Wilson. 2020. The what-if tool: Interactive probing of machine learning models. In *IEEE Transactions on Visualization and Computer Graphics*. Originally released 2019.

A Appendix

A.1 Handling Harmful Text and Instance Examples

RISE-NLP may display offensive language because it operates on toxicity and hate-speech datasets. The demo starts from distribution-level plots and reveals only small user-selected sets of examples from the RISE curve. We include content warnings and recommend access control and dataset provenance documentation for deployments that expose raw text (Kirk et al., 2022).

Figure 4 shows the Instance Inspector view used in the demo workflow. In this example, the auditor selects a residual region from HateXplain with RoBERTa and the race_african attribute. The view shows the selected examples with outcome

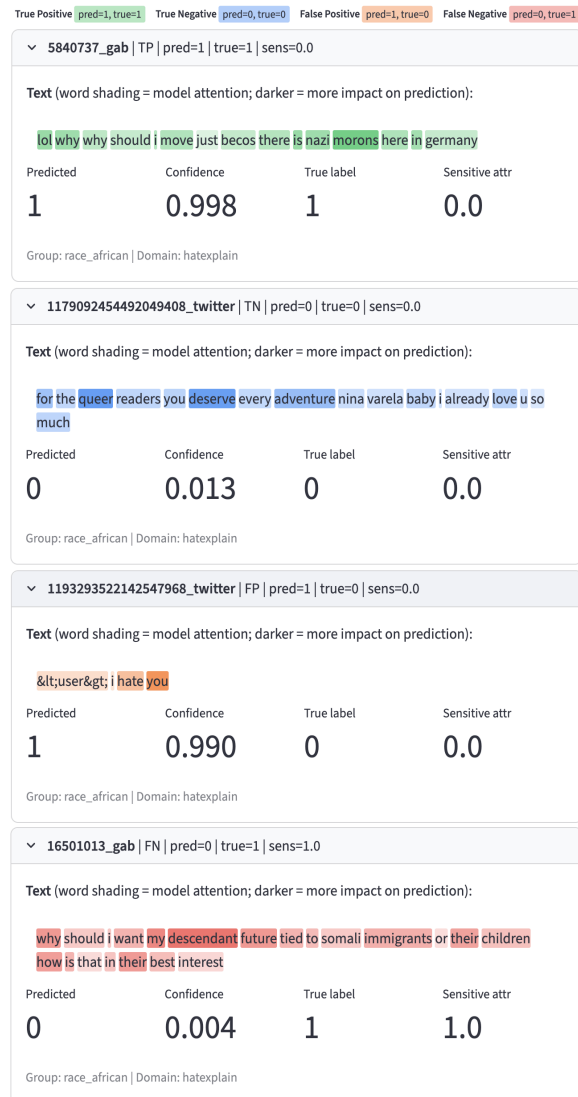


Figure 4: Instance Inspector view for selected HateXplain examples from a residual region.

type, model confidence, true label, sensitive attribute, and token-level attention shading. The bottom false-negative example has low predicted toxicity confidence despite a hateful label, illustrating the type of instance an auditor can retrieve from a separated residual band.