Real-Time Visual Feedback to Guide Benchmark Creation: A Human-and-Metric-in-the-Loop Workflow

Anonymous ACL submission

Abstract

Recent research has shown that language models exploit 'artifacts' in benchmarks to solve tasks, rather than truly learning them, leading to inflated model performance. In pursuit of creating *better benchmarks*, we propose VAIDA, a novel benchmark creation paradigm for NLP, that focuses on *guiding crowdworkers*, an under-explored facet of addressing benchmark idiosyncrasies. VAIDA facilitates sample correction by providing realtime visual feedback and recommendations to improve sample quality. Our approach is domain, model, task, and metric agnostic, and constitutes a paradigm shift for robust, validated, and dynamic benchmark creation via human-and-metric-in-the-loop workflows. We evaluate via expert review and a user study with NASA TLX. We find that VAIDA decreases effort, frustration, mental, and temporal demands of crowdworkers and analysts, simultaneously increasing the performance of both user groups with a 45.8% decrease in the level of artifacts in created samples. As a by product of our user study, we observe that created samples are adversarial across models, leading to decreases of 31.3% (BERT), 22.5% (RoBERTa), 14.98% (GPT-3 fewshot) in performance[1].

1 Introduction

Researchers invest significant effort to create benchmarks in machine learning, including ImageNet (Deng et al., 2009), SQUAD (Rajpurkar et al., 2016), and SNLI (Bowman et al., 2015), as well as to develop models that solve them. *Can we rely on these benchmarks?* A growing body of recent research (Schwartz et al., 2017; Poliak et al., 2018; Kaushik and Lipton, 2018) is revealing that models exploit spurious bias/artifacts– unintended correlations between input and output (Torralba and Efros, 2011) (e.g. the word 'not' is associated with the label 'contradiction' in Natural Language Inference (NLI) (Gururangan et al., 2018))– instead of the actual underlying features, to solve many popular benchmarks. Models therefore fail to generalize, and experience drastic performance drops when testing with out of distribution (OOD) data or adversarial examples (Bras et al., 2020; McCoy et al., 2019; Zhang et al., 2019; Larson et al., 2019b; Sakaguchi et al., 2019; Hendrycks and Gimpel, 2016). This begs the question: *Shouldn't ML researchers consequently focus on creating 'better' datasets rather than developing increasingly complex models on bias-laden benchmarks?*

Deletion of samples based on bias baseline reports– hypothesis-only baseline in NLI (Dua et al., 2019))– and mitigation approaches such as AFLite (Sakaguchi et al., 2019) (adversarial filtering which deletes targeted data subsets), (Clark et al., 2019; Kaushik et al., 2019), have the following limitations: (i) data deletion/augmentation and

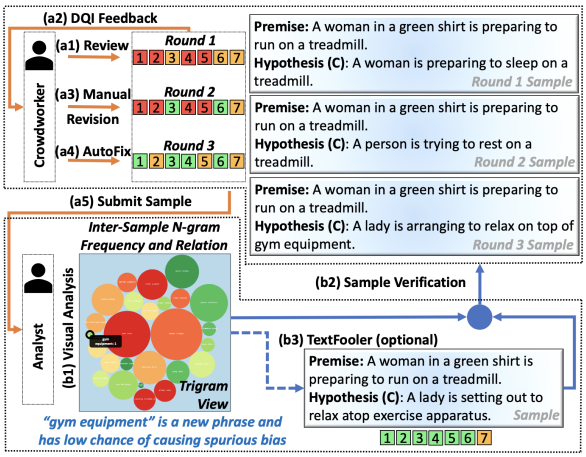


Figure 1: VAIDA workflow– here, *(a)* and *(b)* branches represent crowdworker/analyst functions respectively. Visual feedback is provided based on several aspects of inter and intra-sample artifact presence (numbered 1 through 7), where red>>yellow>>green. Analysts are provided with detailed visualizations of artifact levels during sample validation.

[1]A video description of VAIDA, generated samples, and detailed analyses are available in the Supplemental Material.

residual learning do not justify the original investment in data creation, and (ii) crowdworkers are not provided continuous feedback to learn what constitutes high quality data– and so have additional overhead due to the manual effort involved in sample creation/validation. One potential solution to these problems is *in situ* feedback about artifacts while benchmark data is being created. *To our knowledge, there are no approaches which provide realtime artifact identification, feedback, and reconciliation opportunities to data creators, nor guide them on data quality*.

Contributions: (i) We propose *VAIDA* (Visual Analytics for Interactively Discouraging Artifacts), a novel system for benchmark creation that provides continuous visual feedback to data creators in real-time. VAIDA supports artifact identification and resolution, implicitly educating *crowdworkers* and *analysts* on data quality (Figure 1). (ii) We design a *crowdworker* workflow to create new data samples for benchmark inclusion. Feedback from VAIDA guides crowdworkers on why a sample likely constitutes an artifact. To assist with sample modification, we propose an AutoFix module, that allows for machine-assisted sample modification to achieve *higher quality (i.e., lower presence of artifacts)*. (iii) We develop a series of visualizations for *analysts* to analyze and verify submitted samples to build an optimal dataset. VAIDA allows visual exploration of the effect of a sample's addition to a dataset in both cold-start and pre-existing data scenarios. We also propose the use of TextFooler (Jin et al., 2019) for adversarial transformation to increase benchmark robustness using model-in-the-loop. (iv) We leverage DQI (Mishra et al., 2020) which identifies artifacts by decomposing samples according to their language properties and scores them. (v) We evaluate VAIDA empirically through expert review and a user study to understand the cognitive workload it imposes. The results indicate that VAIDA decreases mental demand, temporal demand, effort, and frustration of crowdworkers (31.1%) and analysts (14.3%); it increases their performance by 34.6% and 30.8% respectively, and educates crowdworkers on how to create *high quality* samples. Overall, we see a 45.8% decrease in the presence of artifacts in created samples. (vi) Even though our main goal is to reduce artifacts in samples, we observe that samples created in our user study are adversarial across language models with performance decreases of 31.3% (BERT), 22.5% (RoBERTa), and 14.98% (GPT-3 fewshot).

2 Related Work

This work sits at the intersection of two primary areas: (1) *visual analysis of data quality* (where we consider the presence of artifacts to lower quality), and (2) *development of a novel data collection pipeline*[2].

2.1 Sample Quality and Artifacts

Data Shapley (Ghorbani and Zou, 2019) has been proposed as a metric to quantify the value of each training datum to the predictor performance. However, the metric might not signify bias content, as the value of training datum is quantified based on predictor performance, and biases might favor the predictor. Moreover, this approach is model and task-dependent. VAIDA uses DQI (Data Quality Index), proposed by (Mishra et al., 2020), to: (i) compute the overall data quality for a benchmark with n data samples, and (ii) compute the impact of a new $(n+1)^{th}$ data sample. Table 1 broadly defines DQI components, along with their interpretation in VAIDA , and juxtaposes them against evaluation methods used in prior works on crowdsourcing pipelines, as discussed in 2.2. (Wang et al., 2020) concurrently propose a tool for measuring and mitigating artifacts in image datasets.

2.2 Crowdsourcing Pipelines

Several pipelines have been proposed to handle various aspects of artifact presence in samples. These primarily focus on: (i) improving sample diversity via word and phrase recommendation schemes that prevent the repetition of over-represented features by crowdworkers (Yaghoub-Zadeh-Fard et al., 2020; Larson et al., 2019a, 2020; Stasaski et al., 2020), and (ii) prompting users to alter samples by highlighting portions of text that are important for the model to make a prediction to promote either adversarial sample creation (Wallace et al., 2019; Kiela et al., 2021) or identification of 'unknown unknowns' (i.e., instances for which a model makes a high confidence prediction that is incorrect) (Attenberg et al., 2015; Vandenhof, 2019).

As shown in Table 1, DQI encompasses the aspects of artifacts studied by the aforementioned works; it further quantifies the presence of many

[2]Detailed related work is in the Supplemental Material.

Component Name	DQI Implication	VAIDA Usage	Artifacts Evaluated
Vocabulary	Ambiguity and diversity of a dataset's language	Does the sample contribute new words?	Sample Length (Wallace et al., 2019), New Words Introduced (Yaghoub-Zadeh-Fard et al., 2020; Larson et al., 2020), Jaccard Index between n-grams (Larson et al., 2019a)
Inter-Sample N-gram Frequency and Relation	Word/phrase repetition and similarity between samples	Does the sample contribute new combinations of words and phrases?	N-gram overlap (Wallace et al., 2019; Yaghoub-Zadeh-Fard et al., 2020), Mean-IDF (Stasaski et al., 2020)
Inter-Sample STS	Syntactic, semantic, and pragmatic sentence parsing	How similar is the hypothesis to all other premises or hypotheses?	Multi-hop reasoning (Wallace et al., 2019), Similarity and overlap (Yaghoub-Zadeh-Fard et al., 2020), Diversity (Larson et al., 2019a)
Intra-Sample Word Similarity	Word overlap and similarity within sample statements	How similar are all words within a sample?	Coreference Resolution, Multi-hop reasoning (Wallace et al., 2019), Word Overlap (Larson et al., 2020)
Intra-Sample STS	Phrase/sentence level overlap within a sample	How similar is the hypothesis to the premise?	N-gram repetition and overlap (Yaghoub-Zadeh-Fard et al., 2020)
N-gram Frequency per Label	Distribution of samples according to annotation	Is the hypothesis too obvious for the system?	Logic and Calculations (Wallace et al., 2019), Diversity (Larson et al., 2019a), Outliers, Entropy (Stasaski et al., 2020)
Inter-Split STS	Optimal similarity between train and test samples	Is the sample too similar to an existing sample?	Entity Distractors, Novel Clues (Wallace et al., 2019), Coverage (Larson et al., 2019a)

Table 1: Language properties considered in DQI that indicate artifact presence, their interpretation in VAIDA, and corresponding methods used in crowdsourcing pipeline evaluation; STS: semantic textual similarity.

more inter and intra-sample artifacts[3], and provides a one stop solution to address artifact impact on multiple fronts. VAIDA leverages DQI to identify artifacts, and further focuses on educating crowdworkers on exactly 'why' an artifact is undesirable, as well as the impact its presence will have on the overall corpus. This is in contrast to the implicit feedback provided by word recommendation and/or highlighting in prior works– VAIDA facilitates the elimination of artifacts without the unintentional creation of new artifacts, something that has hitherto remained unaddressed.

3 Workflow and Modules

In this section, we describe VAIDA's high-level workflow and important backend processes.

3.1 Crowdworker and Analyst Workflows

VAIDA's high-level workflow is shown in Figure 1. Both crowdworkers and analysts work in parallel to create benchmark data points.

For crowdworkers, (a1) artifacts in newly created samples are identified by DQI and (a2) realtime feedback is given to the user. To fix an artifact, users can (a3) manually revise the sample, (a4) run AutoFix to automatically update it, or simply discard the sample and create a new one. After review (and potentially iterative DQI evaluations/revisions), (a5) the sample can be submitted for benchmark inclusion. For analysts, (b1) VAIDA provides several visual interfaces[4] to support detailed analysis and review of submitted samples, and to assess artifact presence (i.e., quality) in the overall benchmark. Submitted samples enter a *pending* state until reviewed by the analyst, who *accepts*, *rejects*, or *modifies* the sample. (b2) Sample decisions are communicated back to crowdworkers to provide continuous feedback about performance and allow them to correct such samples. (b3) Analysts can also submit low quality samples to TextFooler for adversarial transformation and augument with high quality samples to improve the robustness of the dataset, thereby ensuring minimal data loss.

3.2 Modules

DQI and Traffic Signal Scheme: VAIDA communicates sample quality using an intuitive traffic signal color coding (red, yellow, green) to indicate levels of artifacts in samples. Based on overall sample quality, VAIDA computes the probability the sample will be accepted/rejected.[5]. Table 2 shows the explanations VAIDA provides for the DQI feedback shown to crowdworkers during sample revisions in Figure 1, given 100 pre-existing dataset samples.

AutoFix: We propose AutoFix as a module to help crowdworkers avoid creating bad samples by recommending changes to a sample to improve its quality. The AutoFix algorithm is explained in Figure 2. Given a premise, hypothesis, and DQI values for the hypothesis, AutoFix sequentially masks each word in the hypothesis and ranks words based on their impact on model output, i.e. their impor-

[3]See Supplemental Material for details on artifacts that DQI identifies.

[4]See Supplemental Material: Interface Design for interface intuitions and detailed description, with full-resolution images.

[5]Hyperparameters for color mapping depend on the application type. See Supplemental Material: Hyperparameters.

Component Name	Round 1	Round 2	Round 3	TextFooler
Vocabulary	0 new words	0 new words	2 new words	3 new words
Inter-Sample N-gram Frequency and Relation	0 new phrases	0 new phrases	1 new phrase	2 new phrases
Inter-Sample STS	2 cases highly similar	1 case highly similar	1 case highly similar	0 cases highly similar
Intra-Sample Word Similarity	High Overlap	High Overlap	Low Overlap	Low Overlap
Intra-Sample STS	High Similarity	High Similarity	Moderate Similarity	Low Similarity
N-gram Frequency per Label	sleep	rest	—	—
Inter-Split STS	1 case highly similar	1 case highly similar	1 case highly similar	1 case moderately similar

Table 2: Feedback from VAIDA for each DQI component over the iterations of sample revision shown in Figure 1.

tance. Hypothesis words are replaced in the order of importance to achieve at least moderate quality. DQI hence controls the amount and aspect of changes made by AutoFix. By incrementally changing the sample, users can understand how and why their sample is being modified and how DQI values are affected.

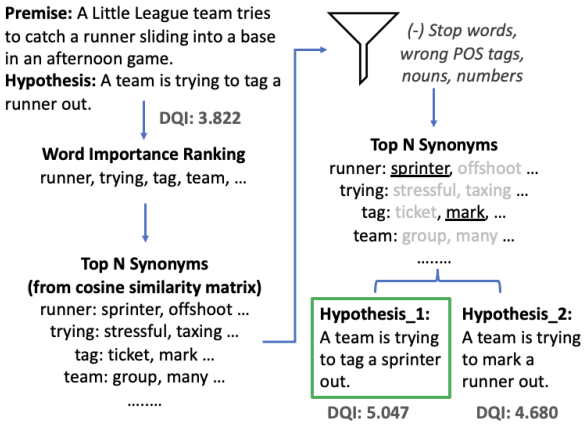


Figure 2: AutoFix Algorithm applied to an SNLI entailment sample, replacing one word per iteration. The DQI of the hypothesis changes from 3.822 to 5.047.

TextFooler: From an analyst's perspective, the quality of a submitted sample might be "too low" because (i) the crowdworker might not employ AutoFix, or (ii) there is a narrow acceptability range due to the criticality of the application domain, such as in BioNLP (Lee et al., 2020). We therefore implement TextFooler (Jin et al., 2019) to adversarially transform low quality samples (instead of discarding them), to improve benchmark robustness, and ensure that crowdsourcing effort is not wasted. We initially use AFLite (Bras et al., 2020), to bin samples into *good* (retained samples) and *bad* (filtered samples) splits. Using TextFooler, we adversarially transform bad split data to flip the label; we revert back to the original label and identify sample artifacts using DQI as shown in Figure 1.

4 Interface Design Choices

VAIDA provides customized interfaces for both crowdworkers and analysts[4].

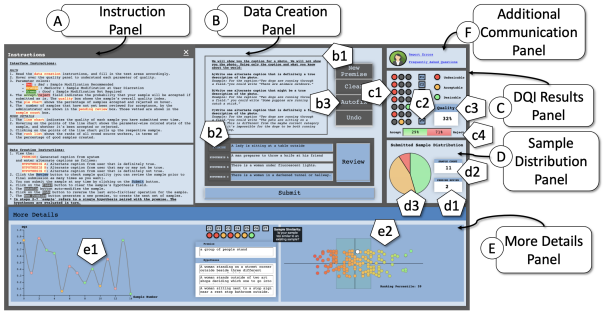


Figure 3: VAIDA's crowdworker interface consists of six linked panels: (A) Instructions, (B) Data creation, (C) DQI results, (D) Sample distribution, (E) More details, and (F) Additional communication.[4]

4.1 Crowdworker Interface

In addition to workflow functionalities, the crowdworker interface (Figure 3) provides interface navigation, data creation, and feedback interpretation instructions (A). Sample creation (B) mimics the original SNLI crowdsourcing interface– examples (b1) are given, and the premise field (b2) autopopulates with captions from the Flickr30 corpus; three hypotheses (for entailment, neutral, contradiction labels) are to be entered at a time, though they are reviewed individually. DQI feedback (C) is shown for each component (c1), and hovering on these displays a tooltip that suggests sample fixes to improve quality (c2). (c3) Overall sample quality and (c4) estimated probability it will be accepted provide additional feedback. AutoFix (b3) can be used for automatic fixes. Samples enter a pending state (d1), until analyst review, upon which the count (d2) and pie chart (d3) update. Historical quality of samples submitted by the user (e1) , and (e2) current rank of the user are shown to help crowdworkers gauge their performance. Communication links for FAQs, and error reporting are also provided (F).

4.2 Analyst Interfaces

While crowdworkers work within a single tightly-coordinated interface to create, submit, and review samples, analysts can navigate between a set of nine interfaces (Figure 4) to review samples in detail to make 'accept', 'reject', and 'modification' decisions, and to assess overall benchmark quality. (UI) The *single crowdworker view* provides a view similar to the crowdworker interface, and allows the analyst to review the work of a single crowdworker. The data creation panel is modified to allow the analyst to iterate over and review submitted samples. For low quality samples, the TextFooler module can be invoked (via a 'Generate Adversarial Example' button). (C1–C7) Other interfaces available to the analyst support detailed review of specific DQI components and allow the analyst to simulate how adding one or more submitted samples affects the benchmark's quality. Several visualization techniques are employed (treemap, node-link diagram, bubble chart, heatmap, bar chart, etc.) tailored to the specific DQI component of interest, but all interfaces consistently utilize the traffic signal color scheme to represent quality.

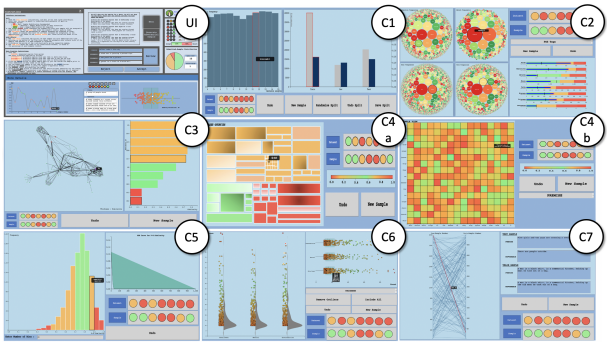


Figure 4: VAIDA provides a collection of interfaces for the analyst supporting detailed analysis and investigation of submitted samples and the overall benchmark.[4]

5 Artifact Case Study with DQI

We evaluate SNLI (Bowman et al., 2015) samples for artifact presence using our traffic signal scheme (based on DQI). We also report results for MNLI(Williams et al., 2017), SQUAD 2.0 (Rajpurkar et al., 2018), and Story CLOZE Task (Schwartz et al., 2017) in the supplemental material. [6]

Setup: We filter samples using AFLite and divide them into two categories: *good* and *bad*, where each category respectively refers to the set of samples retained and removed after adversarial filtering. We get DQI feedback in two different settings: (i) no preexisting samples, and (ii) 100 preexisting samples corresponding to the good category. For (ii), random sampling of 100 pre-existing samples is done 10 times, for a fair comparison. In (ii), we: (a) compute DQI for the existing sample set as x_1, (b) recompute DQI for the sample set after a new sample is added as x_2, and (c) calculate $\Delta x = x_1 - x_2$. The crowdworker interface shows the DQI components corresponding to Δx. In the analyst interface, both Δx as 'sample' and x_2 as the 'dataset' quality are shown component-wise in each view. For fair comparison, we have taken illustrative samples from the AFLite paper (Bras et al., 2020) for SNLI[6]. We tune hyperparameters separating the boundary between red, yellow, and green flags on 0.01% of data manually in a supervised manner (Mishra et al., 2020).

Split	Label	DQI Color
Good	Entailment	
	Neutral	
	Contradiction	
Bad	Entailment	
	Neutral	
	Contradiction	

Table 3: Evaluating samples from SNLI over the most sensitive DQI component, Intra-Sample Word Similarity. Successes: green/orange for good split, red/orange for bad split. Failures: red for good split, green for bad split.

Results: On average, DQI component colors are predicted with 83.3% accuracy according to AFLite categorization of good and bad splits[6] for (i) and (ii), as illustrated in Table 32. False positives and false negatives can be attributed to the limitation of AFLite in incorrectly classifying samples (Mishra et al., 2020). On expanding our evaluation to the other 3 datasets, we have two further observations: (i) prediction accuracy decreases as the artifact level in a dataset decreases. (ii) values of most DQI sub-components do not change significantly (<25% of the time) after adding samples in both categories. However, it changes considerably (>60% of the time) across two sub-components: Intra-sample word overlap and word similarity, both of which belong to the fifth component of DQI. This can

[6] See Supplemental Material: Evaluation, for details across all DQI components, hyperparameter tuning, and analyses.

again be explained by AFLite's sensitivity towards word overlap (Mishra et al., 2020).

6 Evaluation

We evaluate VAIDA's efficacy at providing real time feedback to educate crowdworkers during benchmark creation using expert review and a user study. We also evaluate model performance (BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (fewshot) (Brown et al., 2020)) on data created with VAIDA during the user study.

6.1 Expert Review

We present an initial prototype of our tool, to a set of three researchers with expertise in NLP and knowledge of data visualization, in order to judge the interface design. For each expert, the crowdworker interface and then analyst interfaces were demoed. Participants could ask questions and make interaction/navigation decisions to facilitate a natural user experience. All the experts appreciated the easily interpretable traffic-signal color scheme (and further suggested that alternates be provided to account for color blindness) and found the organization of the interfaces—providing separate detailed views within the analyst workflow– a way to prevent cognitive overload (too much information on one screen) while allowing multi-granular analysis; this would help in classifying samples of middling quality as benchmark size increases with relative ease.

6.2 User Study

Setup: We approach several software developers, testing managers, and undergraduate/graduate students. Based on their domain familiarity (in NLP and visualization, rated from 1:novice-5:expert), we split them into 23 crowdworkers and 8 analysts for constructing NLI samples, given premises. There are 100 high quality samples in the system at the time each participant participates in each ablation round. Their experience is evaluated using NASA Task Load Index (Hart, 2006)[7] (NASA TLX), where each task is scored in a 100-points range, with 5-point steps. To conduct an ablation study, we introduce modules one at a time (and finally the complete system) to all user classes as follows: (i) Crowdworkers— conventional crowdsourcing, traffic signal feedback, AutoFix, all, and (ii) Analysts— conventional analysis, traffic signal feedback, visualizations, TextFooler, all. For both types of users, a preliminary walkthrough of panels using 2 fixed samples– chosen randomly from the SNLI subset used in Section 5– is conducted for each round of the study (Figure 5).

Crowdworker

Mode	Time Allotted (minutes)	Minimum Questions Required
Conventional Crowdsourcing	10	3
W/ Traffic Signal Feedback	10	3
W/ AutoFix	10	3
Full System	15	3

Analyst

Mode	Time Allotted (minutes)	Minimum Questions Required
Conventional Analysis	10	3
W/ Traffic Signal Feedback	10	3
W/ Visualization	15	3
W/ TextFooler	10	3
Full System	20	3

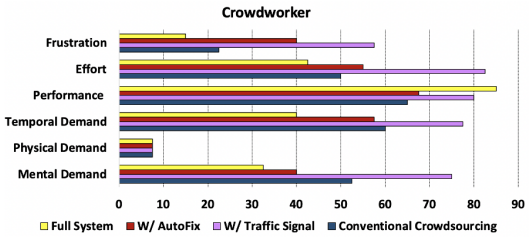
Figure 5: User Study Setup– describes the timeframe and requirements of the user study over ablation rounds.

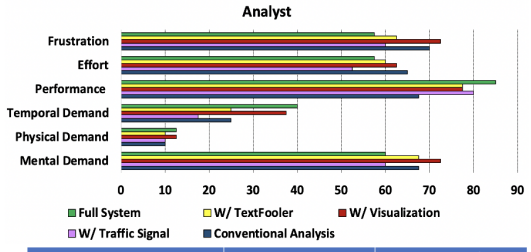
Analysis: Figure 6 summarizes study results, averaged over all user responses. The users are presented with system modules in the order listed, and are asked to report scores for NASA TLX dimensions relative to the original score they assign to the conventional crowdsourcing/analysis approaches; at the end of each round, they are also asked for their comments[7].

Crowdworkers: Traffic signal feedback initially increases time (+25%) and effort (+60%) required to create high quality samples, as users have to correct them. However they are more confident (performance– +27%) of sample quality. AutoFix usage causes an unexpected increase in effort (+5%) and frustration (+88.8%), as users do not fully trust recommendations without visual feedback. The drastic improvement over all aspects (frustration– -44.4%, mental demand– -38.1%, temporal demand– -29.1%, effort– -20%, average decrease in difficulty– -31.1%, performance– +34.6%) in the case of using the full system is in line with this observation. The number of questions created per round (traffic signal– -8.3%, AutoFix– +25%, full system– +83.3%) as well as system scores (traffic signal– +27.3%, AutoFix– +13.6%, full system– +54.5%) also follows this trend, across all types of crowdworkers.

Analysts: Analysts find the task easier (effort– -19.3%, performance– +26.9%) with traffic signal feedback, as quality is clearly marked. When analysts are shown the visualization interfaces, they are explicitly taught to differentiate how the traffic

[7] See Supplemental Material: User Study for more details. We do aggregated analysis of comments, full quotes of comments are present in the Supplemental Material. We also have IRB approval to conduct this user study.





Crowdworker

Mode	Number of Questions Created	System Score (/100)
Conventional Crowdsourcing	3	50
W/ Traffic Signal Feedback	2.75	72.5
W/ AutoFix	3.5	62.5
Full System	5.25	85

Analyst

Mode	Number of Questions Analyzed	System Score (/100)
Conventional Analysis	3	40
W/ Traffic Signal Feedback	5	67.5
W/ Visualization	2.75	72.5
W/ TextFooler	3.75	65
Full System	5.5	82.5

Figure 6: User Study Results– averaged across responses from crowdworkers and analysts respectively for each ablation round.

signal colors in the visualizations indicate a sample's effect on the overall dataset quality. Analysts can toggle between the states of original dataset and new sample addition. We find that analysts initially find toggling more difficult to do (mental demand– +15.4%, temporal demand– +36.4%, frustration– 3.5%), though they agree that it improves their judgement of quality (performance– +15.9%). Analysts' average behavior on TextFooler models the conventional approach quite closely, as analysts are seen to have a tendency to send all samples that are unclear to TextFooler immediately. With the full system, analysts also report improvement in all aspects (average decrease in difficulty– -14.3%), particularly mental demand (-19.2%) and performance (+30.8%), considering that the system increases the likelihood of a low hypothesis baseline. The visualization usage also improves, as analysts learn component relationships. Altogether, sample evaluation by analysts increases (full system– +83.3%), following this trend, and analysts are more assured of their performance (full system score– +94.1%).

Learning Curve: At the end of the study, all users are asked the following: *"What do you think high quality means?"* We find that users are able to distinguish certain patterns that promote higher quality, such as keeping sentence length appropriate and uniform across labels (not too long/short), using complex phrasing ('not bad')/gender information/modifiers across labels, decreasing premise-hypothesis word overlap; they also do not display undesirable behavior like tweaking previously submitted samples just to create more. We also find an overall decrease of 45.8% in the level of artifacts of created samples, across all rounds of ablation.

User Education: We also conduct a variation of the study where a subset of participants (7 crowdworkers and 2 analysts) agreed to create/ analyze samples, for varying numbers of pre-accepted samples (Figure 7), in the full system condition. In general, as the number of samples increases, the proportion of red or mixed samples also increases, and those green decreases. We find that when beginning from the cold start condition, as the sample number increases, due to their familiarity with the system, both crowdworkers and analysts are able to leverage the system better to avoid red samples. However, when participants are directly started in situations with $>$ 500 samples in the system, their unfamiliarity with the system initially causes a steepening of the learning curve compared to the cold start condition; this also tapers and saturates more slowly than cold start as the users gain experience. In the case of cold start, we find that users who create ∼50 samples report lesser reliance on AutoFix as they get better at creating higher quality samples; those who analyze ∼75 samples use TextFooler more efficiently as they understand how to deal with samples of middling quality better.

6.3 Model Performance Results

We evaluate BERT and RoBERTa (trained on the full SNLI dataset), and GPT-3 (in fewshot setting) against the data created during the ablation rounds of the user study. Figure 8 shows the results for samples over each round of ablation (totally 345 samples[8]). In the case of TextFooler, samples are created using the 'full system' condition and then further modified using TextFooler by the analyst. The other sample sets are not modified by the analyst, and are directly accepted after evaluation. We find that across all models, performance is lower when explicit quality feedback (via the traffic signal scheme) is provided, compared to the regular crowdsourcing condition. Performance further de-

[8]The dataset is included in the Supplemental Material.

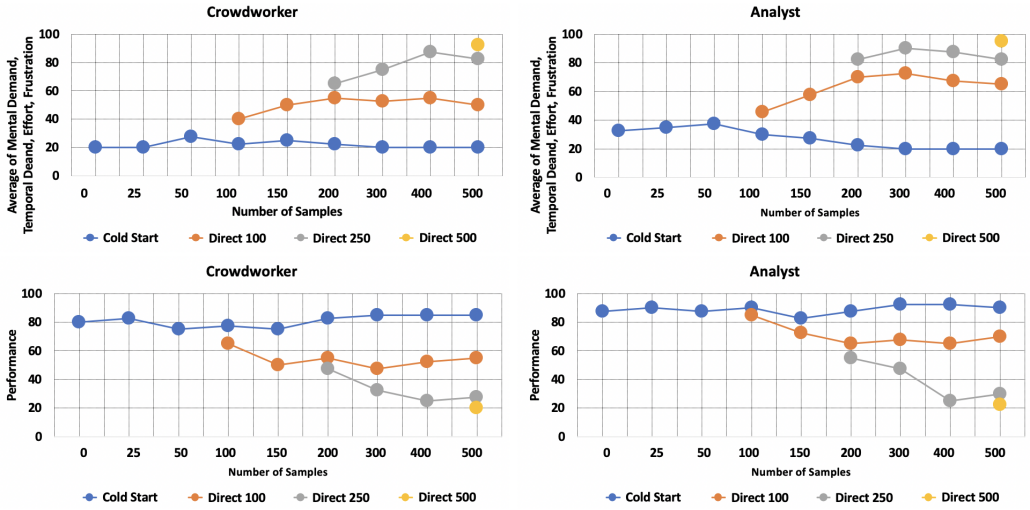


Figure 7: User education curves. Cold start has no pre-existing samples, and direct-n has n pre-existing samples. Mental Demand, Temporal Demand, Frustration, and Effort are averaged, Physical Demand is ignored. Performance is plotted separately as it shows differing behavior than the others.

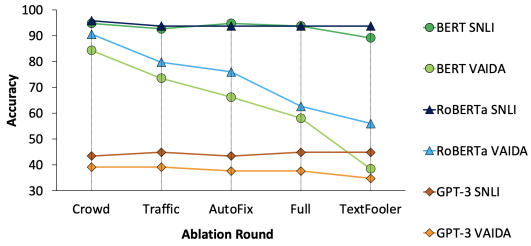


Figure 8: Model performance results for samples created during each ablation round of the user study.

creases when AutoFix is implemented, indicating the effectiveness of this module in seeding suggestions for sample improvement. A significant decrease is seen in the full system and TextFooler conditions. This indicates that crowdworkers and analysts are able to utilize VAIDA's affordances to create more robust text samples.

7 Concluding Remarks

We propose VAIDA, a paradigm to address benchmark artifacts, by integrating human-in-the-loop sensemaking with continuous visual feedback. We design complementary workflows for both crowdworkers and analysts, to create new samples, evaluate them for the existence of artifacts, and review/repair samples to ensure the overall benchmark quality. VAIDA uses several visualization interfaces to analyze quality considerations (based on artifact levels) at multiple granularities. In our evaluation, we see that users report greater satisfaction (crowdworkers: +34.6%, analysts: +30.8%) and lower difficulty (crowdworkers: -31.1%, analysts: -14.3%) with their work and system experience; this implies possible higher crowdworker retention and engagement. We intend to integrate VAIDA with an actual crowdsourcing framework, and run a full scale data creation study to create a high quality benchmark. Expanding to such a set up will require additional back end engineering, to ensure that accurate feedback continues to be provided in real-time to crowdworkers.

Additionally during ablation, we introduce modules in a fixed order to users, as per the patterns of usage preferred by the experts and see that users effectively identify and avoid artifact patterns during sample creation (-45.8% in artifacts). We also see that reliance on AutoFix and TextFooler reduces over time, decreasing the possibility of templated/artificial sample occurrence (though overusage might initially affect the sample quality and learning curve for lay users in a full scale crowdsourcing set up). In future work, we will compare our setup directly with the effect of in-depth user training (Roit et al., 2019) prior to crowdsourcing to analyze if/how user strategy and performance changes during VAIDA usage.

Samples created with VAIDA are not only of higher quality than achieved with conventional crowdsourcing, but are also adversarial across models, with performance decreases of -31.3% (BERT), -22.5% (RoBERTa), -14.98% (GPT-3 fewshot). VAIDA hence demonstrates a novel, dynamic approach for building benchmarks and mitigating artifacts, and serves as a starting point for the next generation of benchmarks in machine learning.

References

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724, Melbourne, Australia. Association for Computational Linguistics.

Joshua Attenberg, Panos Ipeirotis, and Foster Provost. 2015. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)*, 6(1):1–17.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? *arXiv preprint arXiv:1805.11818*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

Peter Clark. 2018. What knowledge is needed to solve the rte5 textual entailment challenge? *arXiv preprint arXiv:1806.03561*.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to" overinformative" referring expressions. *Psychological review*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. *arXiv preprint arXiv:1904.02868*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Sandra G Hart. 2006. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K Kummerfeld, Parker Hill, Michael A Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019a. Outlier detection for improved data quality and diversity in dialog systems. *arXiv preprint arXiv:1904.03122*.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019b. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K Kummerfeld. 2020. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528.

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*.

Nitika Mathur, Tim Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Marjorie McShane and Petr Babkin. 2016. Resolving difficult referring expressions. *Advances in Cognitive Systems*, 4:247–263.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Singh Sachdeva, Chris Bryan, and Chitta Baral. 2020. Dqi: A guide to benchmark evaluation. *arXiv: Computation and Language*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Tiantian Qin, Judee Burgoon, and Jay F Nunamaker. 2004. An exploratory study on promising cues in deception detection and application of decision tree. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 23–32. IEEE.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Kyle Richardson and Ashish Sabharwal. 2019. What does my qa model know? devising controlled probes using expert knowledge. *arXiv preprint arXiv:1912.13337*.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2019. Crowdsourcing a high-quality gold standard for qa-srl. *arXiv preprint arXiv:1911.03243*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017a. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017b. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

David Schuff, Karen Corral, and Ozgur Turetken. 2011. Comparing the understandability of alternative data warehouse schemas: An empirical study. *Decision support systems*, 52(1):9–20.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Katherine Stasaski, Grace Hui Yang, and Marti A Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4958–4968.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? *arXiv preprint arXiv:1808.09384*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Shawn Tan, Yikang Shen, Chin-wei Huang, and Aaron Courville. 2019. Investigating biases in textual entailment datasets. *arXiv preprint arXiv:1906.09635*.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Colin Vandenhof. 2019. A hybrid approach to identifying unknown unknowns of predictive models. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 180–187.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. Vibe: A tool for measuring and mitigating bias in image datasets. *arXiv preprint arXiv:2004.07999*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Mohammad-Ali Yaghoub-Zadeh-Fard, Boualem Benatallah, Fabio Casati, Moshe Chai Barukh, and Shayan Zamanirad. 2020. Dynamic word recommendation to obtain diverse crowdsourced paraphrases of user utterances. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 55–66.

Maria Yancheva and Frank Rudzicz. 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–953.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

A Supplemental Material

The following information is included in the appendix.

- Infrastructure Used
- Run-time Estimations
- Hyper Parameter
- Related Work
- DQI Components
- Evaluation: Artifact Case Study
- Interface Design
- AutoFix and TextFooler Examples
- User Study
- Expert and User Comments

Please refer to the accompanying folder for:

- Video demos of VAIDA workflow
- Sample dataset generated during the ablation rounds of the user study
- DQI and Model Performance Results for User Study Samples

A.1 Infrastructure Used

In Section 3, we describe VAIDA's flow by high level workflow and back-end processes(DQI, AutoFix, and TextFooler). Further, as discussed in Subsection 3.2 DQI can be used for quantifying artifact presence for the: i) overall benchmark, and ii) impact of new samples. Depending on the task at hand we run our experiments in different hardware settings. The DQI calculations run mostly using CPU, for new samples as well as overall samples. The AutoFix procedure, as explained in Subsection 3.2, gives the user assistance in improving quality on a per submission basis. Therefore that does not require high GPU intensive systems; we have provisions to shift execution to a GPU as well if necessary to speed up the process. For TextFooler the fine tuning of the model is run on "TeslaV100-SXM2-16GB"; CPU cores per node 20; CPU memory per node: 95,142 MB; CPU memory per core: 4,757 MB– this is not a necessity as code has been tested on lower configuration GPUs as well but we have run our experiments in this setting. The attack part of the TextFooler requires more memory and we run that code on "Tesla V100-SXM2-32GB" com-pute Capability: 7.0 core Clock: 1.53GHz, coreCount: 80, device Memory Size: 31.75GiB device Memory Bandwidth: 836.37GiB/s.

A.2 Run-time Estimations

The DQI calculations run on CPU (for real life setting purposes); for the approximate estimate of the time taken, we run experiments for fixed data size of 10K samples. If the DQI calculations are done to calculate the impact of individual new samples it take a couple of seconds. On the other hand, If we take the whole 10k size dataset it takes around 48 hours to complete the process on CPU. This whole process can be run in parallel to reduce the time taken to 16 hours.

The TextFooler part consists of two steps– the fine tuning part and attack part– for generating adversaries. For fine tuning models we use "TeslaV100-SXM2-16GB", and it takes 20-30 minutes to complete the process. For the attack part we use "Tesla V100-SXM2-32GB", which takes 2-3 hrs for completing 20k data samples. This estimate requires the cosine similarity matrix for word embeddings to be calculated before hand which takes around 1-2 hrs, but this step has to be done only if the word embeddings are modified. This is a rare task so we have kept this separated.

A.3 Hyper Parameters

to look at the estimations of DQI and its variations, we have kept basic hyper-parameters fixed in the experiments. We keep the learning rate to 1e-5, the number of epochs during the experiments are varied from 2-3, the per gpu train batch and eval batch sizes vary from 8-64 samples (the results shown are with respect to a batch size of 8), adam epsilon is set to 1e-8, weight decay is set to 0, maximum gradient normalisation is set to 1, and maximum sequence length is set to 128. For TextFooler the the semantic similarity is fixed to 0.5 uniformly for all the experiments shown in this paper.

Additionally, the variations and range in the DQI parameters are dataset specific, i.e., hyperparameters depend on the application task. (Mishra et al.,

2020) design DQI as a generic metric to evaluate diverse benchmarks. However, the definitions of what constitutes high and low quality will vary depending on the application. For example, BiomedicaNLP might have lower tolerance levels for spurious bias than General NLP. Another case is in water quality– cited as an inspiration for DQI by (Mishra et al., 2020)– where the quality of water needed for irrigation is different than that of drinking or medicine. We can therefore say that the hyper-parameters in the form of boundaries separating high and low quality data (i.e., inductive and spurious bias) are dependent on applications.

A.4 Related Work

A.4.1 Sample Quality and Artifacts

Data Shapley (Ghorbani and Zou, 2019) has been proposed as a metric to quantify the value of each training datum to the predictor performance. However, the metric might not signify bias content, as the value of training datum is quantified based on predictor performance, and biases might favor the predictor. Moreover, this approach is model and task-dependent. VAIDA uses DQI (Data Quality Index), proposed by (Mishra et al., 2020), to: (i) compute the overall data quality for a benchmark with n data samples, and (ii) compute the impact of a new $(n+1)^{th}$ data sample. (Wang et al., 2020) concurrently propose a tool for measuring and mitigating artifacts in image datasets.

Data Shapley (Ghorbani and Zou, 2019) has been proposed as a metric to quantify the value of each training datum to the predictor performance. However, this approach is model and task dependent. More importantly, the metric might not signify bias content, as the value of training datum is quantified based on predictor performance, and biases might favor the predictor. VAIDA uses DQI (data quality index), proposed by (Mishra et al., 2020), to: (i) compute the overall data quality for a benchmark with n data samples, and (ii) compute the impact of a new $(n+1)^{th}$ data sample. The quality of individual features (aspects) of samples are evaluated based on decreasing presence of artifacts and increasing generalization capability. In a concurrent work (Wang et al., 2020), a tool for measuring and mitigating bias in image datasets has also been proposed. DQI estimates artifact presence by calculating seven component values corresponding to a set of language properties;, along with their interpretation in VAIDA.

A.4.2 Crowdsourcing Pipelines

Adversarial Sample Creation: Pipelines such as Quizbowl(Wallace et al., 2019) and Dynabench(Kiela et al., 2021), highlight portions of text from input samples during crowdsourcing, based on how important they are for model prediction; this prompts users to alter their samples, and produce samples that can fool the model being used for evaluation. While these provide more focused feedback compared to adversarial pipelines like ANLI (Nie et al., 2019), which do not provide explicit feedback on text features, adversarial sample creation is contingent on performance against a specific model (Quizbowl for instance is evaluated against IR and RNN models, and may therefore not see significant performance drops against more powerful models). Additionally, such sample creation might introduce new artifacts over time into the dataset and doesn't always correlate with high quality– for instance, a new entity introduced to fool a model in an adversarial sample might be the result of insufficient inductive bias, though reducing the level of spurious bias.

A similar diagnostic approach is followed for unknown unknown identification– i.e., instances for which a model makes a high confidence prediction that is incorrect. (Attenberg et al., 2015) and (Vandenhof, 2019) propose techniques to identify UUs, in order to discover specific areas of failure in model generalization through crowdsourcing. The detection of these instances is however, model-dependent; VAIDA addresses the occurrence of such instances by comparing sample characteristics between different labels to identify (and resolve) potential artifacts and/or under-represented features in created data.

Promoting Sample Diversity: Approaches focusing on improving sample diversity have been proposed, in order to promote model generalization. (Yaghoub-Zadeh-Fard et al., 2020) use a probablistic model to generate word recommendations for crowdworker paraphrasing. (Larson et al., 2019a) propose retaining only the top k% of paraphrase samples that are the greatest distance away from the mean sentence embedding representation of all collected data. These 'outlier' samples are then used to seed the next round of paraphrasing. (Larson et al., 2020) iteratively constrain crowdworker writing by using a taboo list of words, that prevents the repetition of over-represented words, which are also a source of spurious bias. Additionally,

(Stasaski et al., 2020) assess the new sample's contribution to the diversity of the entire sub-corpus.

DQI encompasses the aspects of artifacts studied by the aforementioned works; it further quantifies the presence of many more inter and intra-sample artifacts, and provides a one stop solution to address artifact impact on multiple fronts. VAIDA leverages DQI to identify artifacts, and further focuses on educating crowdworkers on exactly 'why' an artifact is undesirable, as well as the impact its presence will have on the overall corpus. This is in contrast to the implicit feedback provided by word recommendation and/or highlighting in prior works– VAIDA facilitates the elimination of artifacts without the unintentional creation of new artifacts, something that has hitherto remained unaddressed.

A.4.3 Task Selection and Controlled Dataset Creation

In this work, we demonstrate VAIDA for a natural language inference task (though it is task-independent), and mimic the SNLI dataset creation and validation processes. Elicited annotation has been found to lead to social bias in SNLI using probablistic mutual information (PMI) (Rudinger et al., 2017a). Visual feedback is provided based on DQI (which takes PMI into account) to explicitly correct this bias, and discourage the creation of such samples. Also, human annotation of machine-generated sentences/sentences pulled from existing texts instead of elicitation has been suggested to reduce such bias (Zhang et al., 2017). However, machine-generated text might look artificial, and work has shown that text generation has its own set of quality issues (Mathur et al., 2020). While we use AutoFix and TextFooler as modules to automatically transform samples, they are designed to be used in parallel with human sample creation. Their results can also be further modified by humans prior to submission. We see less reliance on these tools over the course of our user study, as discussed in Subsection 6.2. Additionally, previous work (Roit et al., 2019) in controlled dataset creation trains crowdworkers, and selects a subset of the best-performing crowdworkers for actual corpus creation. Each crowdworker's work is reviewed by another crowdworker, who acts as an analyst (as per our framework) of their samples. However, in real-world dataset creation, such training and selection phases might not be possible. Additionally, the absence of a metric-in-the-loop basis for feedback provided during training can potentially bias (through trainers) the created samples.

A.5 DQI Components

DQI shows the (i) the overall data quality and (ii) the impact of new data created on the overall quality. In this paper, higher quality implies lower artifact presence and higher generalization capability. DQI clubs artifacts into seven broad aspects of text, which cover the space of various possible interactions between samples in an NLP dataset. Below, we provide more details of each of the 63 parameters that pertain to bias addressed in DQI, along with examples for better illustration.

A.5.1 Vocabulary

This bin deals with leads related to the vocabulary of a dataset. Specifically, the language used in the dataset in terms of its ambiguity and diversity is analyzed.

Vocabulary Magnitude: A dataset of size of 100k samples and 30k unique words will have a vocabulary magnitude of 0.3. (Poliak et al., 2018; Gururangan et al., 2018)

Language Perturbation: The substitution of words like 'and' or 'by' with fillers such as 'blah' helps check if the original words are being used as a part of the reasoning context or not.(Talmor et al., 2019)

Semantic Adverb Resolution: There is a difference in the contexts created by 'always', 'sometimes', 'often', and 'never.'(Talmor et al., 2019)

Domain Specific Vocabulary: The names of countries such as Syria, Canada, Mexico, etc., and nationalities, such as Indian, Swiss, etc. are not recognized by language models, and performance on instances containing these words is low.(Poliak et al., 2018; Gururangan et al., 2018; Glockner et al., 2018)

A.5.2 Inter-sample N-gram Frequency and Relation

This bin looks at leads that concern n-grams individually or in relation to other n-grams. Replacement based methods seem to provide a viable way to dilute the influence of these leads on bias.

Maximal Word Distance: A dataset that covers the scientific domain will have words dissimilar

to more commonly used language. (Poliak et al., 2018; Gururangan et al., 2018)

POS Tag Replacement: Consider the word 'Jordan' in vocabulary, where the context is that Jordan refers to the country. An equivalent country name (of the same POS tag) like 'Russia' can be used for replacement. Jordan could also refer to a person's name, such as 'Michael Jordan'. In this case, on replacement, 'Michael Russia' will be generated. This case does not add an example that makes sense. So such samples are discarded based on the count of the bigrams generated on replacement. In TextFooler, consider the input "The characters, cast in impossibly contrived situations, are totally estranged from reality." The output might be: "The characters, cast in impossibly engineered circumstances, are fully estranged from reality." (**?**Zhao et al., 2017; Glockner et al., 2018; Jin et al., 2019; Li et al., 2016)

Consecutive Verb Frequency: It has been observed that on translation from English to German and back, sentences such as 'She was cooking dressed for a wedding' drop the second verb on retranslation, and becoming 'She was cooking for a wedding.' (Zhao et al., 2017)

Anonymization of Entities: (Hermann et al., 2015; Li et al., 2018) Original Version: Content: 'The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host.' Question: Who hosts Top Gear? Answer:Jeremy Clarkson

Anonymized Version: Content: 'The $ent1$ producer allegedly struck by $ent2$ will not press charges against the "Top Gear" host.' Question: Who hosts Top Gear? Answer: $ent2$

Metonymy: 'If we don't get these papers in today, the suits will be after us.' Here, suits refers to business people. (Clark, 2018)

Stereotypes: Word associations like 'cook' or 'dolls' with 'girls', or 'temples' with 'India' are a source of bias. (Rudinger et al., 2017b)

Out of Distributions in Range: 'Sheila and I' and 'Sheila or I' have different contextual meanings which can't be solved by pattern correlation. 'Jim, John and Bob are 14, 12, and 18. Who is the second oldest?' returns the correct answer. But if their ages are '1997', '2001', and '2010', then the system returns the wrong answer. (Talmor et al., 2019)

Unnatural Language: The sentence: 'She was [MASK] fast, she was rapid,' has different meanings if you substitute 'not' or 'very' in it. (Talmor et al., 2019)

Broad Referring Expressions: Generic terms like 'this', 'the', 'that', or 'it' can be used to refer to objects on different occasions. These must be resolved to remove ambiguity. (Gundel et al., 1993; McShane and Babkin, 2016; Degen et al., 2020)

A.5.3 Inter-sample STS

This bin deals with leads that can create and dilute bias as a consequence of a new sample's introduction in terms of sentence similarity. Syntactic, semantic, and pragmatic properties of sentences are considered.

Sentence Structure: If a majority of sentence structures follow passive voice, an active voice sentence won't be easily parsed. (Poliak et al., 2018)

Multistep Reasoning: 'When comparing a 23, a 38 and a 31 year old, the [MASK] is oldest A. second B. first C. third.' (Talmor et al., 2019; Naik et al., 2018)

Inter-Sentence Antithesis: 'It was [MASK] hot, it was really cold . A. not B. really.'(Naik et al., 2018)

Sentence Length Variation: Sentences with less detail are shorter, and therefore more likely to be classified as entailment. (Gururangan et al., 2018)

Start Tokens: The candidate answer resolution is restricted by starting "wh-" and "how many" expressions. (Sugawara et al., 2018)

Ellipsis Resolution: 'I went to the mall on Monday, and she on Sunday' can be unrolled as 'I went to the mall on Monday, and she went to the mall on Sunday.' (Clark, 2018)

A.5.4 Intra-sample Word Similarity

This bin concerns intra-sample bias, in the form of word similarities. Specifically, bias seen within the premise and/or within the hypothesis statements of a sample is dealt with.

Presupposition and Query: 'This ban is the first ban for YouTube in China.' Here, the statement asuumes that there is a ban, and the model must reason on whether the ban was the first, not on the existence of the ban. (Clark, 2018)

Coreference Resolution: 'Tom said that he would get it done.'Here, he refers to Tom. (Gururangan et al., 2018; Cirik et al., 2018)

Taxonomy Trees: 'Horse and crow' are grouped as animal, but 'crow and horse' are grouped as birds. This is because 'crow' is closer to 'bird' on the taxonomy tree than 'animal.' (Talmor et al., 2019)

A.5.5 Intra-sample STS

This bin is concerned with another aspect of intra-sample bias, i.e., that which is seen between the premise and hypothesis statements.

Overlap: 'The dog sat on the mat' and 'The dog did not sit on the chair' contain significant overlap and hence can easily be solved. In HANS, consider the premise 'The judges heard the actors resigned' and 'The judges heard the actors'. If a model relied on overlap, it would mark this sample as entailment, even though the gold label is neutral. (Naik et al., 2018; McCoy et al., 2019)

A.5.6 N-gram Frequency per Label

This bin contains leads that reflect the dominating causes of bias introduced due to the influence of existing labels on the new sample's label. Leads are shortlisted in terms of bias originating from (i) premise, (ii) hypothesis, and (iii)both.

Erasure: Consider the sample 'I took my daughter and her step sister to see a show at Webster hall . It is so overpriced I'm in awe.' Using a BI-LSTM, the minimal set of words identified for 'value' is 'It is so overpriced I'm in awe.' (Li et al., 2016)

Similarity: Similarity indicates overlapping detail. For example, 'The bird sang' and 'The robin warbled outside the window as it looked for breakfast' have less overlap due to the presence of more detail in the second sentence. (Naik et al., 2018; Clark, 2018)

Negation: 'She was pleased' and 'She could do nothing that did not please her' might be labeled as contradiction, due to the presence of negation terms. (Poliak et al., 2018)

Antonymy: Simple binary opposites are 'hot' and 'cold'. Less direct opposites are words like 'winter' and 'summer'. (Naik et al., 2018)

WL Mapping: 'Humans' and' instruments are found to be indicators of entailment, 'tall' and 'win' that of neutral, and 'sleep' and 'no' of contradiction. (Poliak et al., 2018) $P(l/w) = \frac{p(w,l)}{p(w)\cdot p(l)}$

PL Mapping: For the phrase 'x was sentient. . . .' ; by identifying the nature of 'x', a model can infer the label without looking at the rest of the sentence . Such lexical semantic exploitation indicates that context is not used in solving such samples. (Poliak et al., 2018) $P(l/p) = \frac{p(p,l)}{p(p)\cdot p(l)}$

Vocabulary Score: Consider the word 'move' in the entailment, neutral, and contradiction classes, with counts 200, 345, and 126 respectively. Then, the score vector would be [3 200 345 126]. (Poliak et al., 2018)

Overlap Rate: $OverlapRate = \frac{numberofoverlapwords}{numberofwordsinsample}$ (Dasgupta et al., 2018)

Copying: Copy all possible subset of words from the premise to the hypothesis iteratively, and check when the label changes. (Gu et al., 2016; See et al., 2017; Gorman and Bedrick, 2019; Aharoni and Goldberg, 2018; Merity et al., 2016)

Hypothesis Only Prediction: The sample: 'People raise dogs because they are obedient' and 'People raise dogs because dogs are obedient', benefits from considering hypothesis only as there is no coreference to be resolved. (Tan et al., 2019)

Cue Influence: Let k be a cue, T_j be the set of tokens in the warrant for data point i with label j, and n be the total number of data samples. (Niven and Kao, 2019)

Applicability: number of data points a cue occurs with one label but not the other $\alpha_k = \sum_{i=1}^{n}[\exists j, k \in T_j^{(i)} k \notin T_{\neg j}^{(i)}]$ Productivity: proportion of applicable data points for which a cue predicts the correct answer $\pi_k = \frac{\sum_{i=1}^{n} 1[\exists j, k \in T_j^{(i)} \wedge k \notin T_{\neg j}^{(i)} \wedge y_i = j]}{\alpha_k}$ Coverage: proportion of applicable cases of a cue over the total number of data points $\xi_k = \frac{\alpha_k}{n}$

Length Mismatch: The sample: 'She was happy with her bonus' and 'She decided to celebrate her raise at work by eating out,' is more likely to be labelled as neutral. (Poliak et al., 2018; Gururangan et al., 2018; Naik et al., 2018)

Grammaticality: Consider the sample': She has no option' and She has no way than the oth-

ers'. This is more likely to be classified as 'non-entailment.' (Poliak et al., 2018)

PMI: $PMI(word, label) = \log \frac{p(word,label)}{p(word)\cdot p(label)}$ (Naik et al., 2019; Gururangan et al., 2018)

Scripts: Consider the sample: 'Canada's plans to launch a satellite, but U.S. officials say the launch is a disguised long-range missile test' and 'The U.S. fears that the Canadian satellite is a ruse to hide the testing of a missile.' There is a familiar script at play here. Countries want to test military equipment, but don't want to be seen as testing them, so may try and hide or cover up the test. Other countries are worried about this form of deceit, and may try and put political pressure on the testing country in order to prevent deceit. (Clark, 2018)

Numerical Reasoning: 'There were two major bombings in less than a week, with 10 people killed by a car bomb south of Baghdad and more than 30 dead when a suicide bomber blew himself up in the capital.' Requires a sum of 30+10 to be calculated to address the hypothesis: 'In less than a week there were 2 major bombings in Iraq, killing more than 40 people.' (Naik et al., 2018; Gururangan et al., 2018)

Gender: Using terms like 'woman' and 'boy' instead of 'person' or 'child' are indicative of non-entailment. (Gururangan et al., 2018)

Hypernyms and Hyponyms: (i) Words like 'wolf' and 'dog' are both animals, but confusion may occur during hyponym resolution as a wolf is a wild animal. (ii) A chair might serve as a superset for its legs, which is not a true hypernym. (Glockner et al., 2018; Richardson and Sabharwal, 2019; Levy et al., 2015)

Modifiers and Superlatives: Words like 'tall' or 'popular' and 'best' or 'first' are indicative of neutral label. (Gururangan et al., 2018)

Causal Phrases: Sentences that contain causal words like 'due to', 'because of', 'consequently', etc. are indicative of neutral label. (Gururangan et al., 2018)

Absence Indicators: The word 'sleep' indicates the absence of activity, and hence is used as an indicator of contradiction. (Gururangan et al., 2018)

Ambiguity: 'She had a black bat' requires context and knowledge to decide if 'bat' refers to an animal, or sports equipment. (Naik et al., 2018)

Bigram Entropy: (Tan et al., 2019; Li et al., 2018) Object bias: For example, 'playing piano' is the only class depicting pianos. This can be inferred by searching for 'piano' or 'music'.

Scene bias: For example, 'soccer juggling' can be resolved by searching for words like 'goal', 'net', or 'ball'.

Person bias: For example, 'military marching' can be resolved by matching to words like 'army' or 'parade'.

Paraphrasing: 'Same' and 'replica' are paraphrases, but 'same' and 'about same' are not. (Clark, 2018; Sugawara et al., 2018; Zhang et al., 2019)

PAWS: Word Swapping: 'Can a bad person become good? : 'Can a good person become bad?'

PAWS: Back Translation: 'The team also toured in Australia in 1953.' : 'In 1953, the team also toured in Australia.'

Multiple Cases: Context: [...] This plot of land is scheduled to house the permanent United Airlines Flight 93 memorial. [...] Question: What was the name of the flight? Answer: 93 Possible answers: United Airlines Flight 93, Flight 93 Here, multiple choices have the correct span of 93 (**?**Sugawara et al., 2018).

Modality and Belief: Epistemic: Agatha must be the murderer. (necessity:neutral)

Deontic: Agatha must go to jail. (obligatory:neutral)

Circumstantial: Agatha must sneeze. (possibility:entailment)

Belief for the above case is true/false in order to label them. (Bowman et al., 2015; Williams et al., 2017)

Shuffling Premises: It is a method of iteratively substituting premises to check word correlation.(Tan et al., 2019)

Concatenative Adversaries: Add distractor words at the end of hypotheses such as negation, superlatives, etc. to test the model's operation over the original samples. (Naik et al., 2018; Jia and Liang, 2017)

Crowdsource Setting: The length of a contradiction hypothesis is generally shorter than that of the original premise, and it uses simpler language. (Schwartz et al., 2017; Qin et al., 2004; Yancheva and Rudzicz, 2013; Newman et al., 2003; Mostafazadeh et al., 2016)

Sample Perturbation: Counterfactual Sample:

P: A young dark-haired woman crouches on the banks of a river while washing dishes.

OH: A woman washes dishes in the river while camping. (Neutral)

NH: A woman washes dishes in the river. (Entailment)

Contrast Set Sample:

Original Text: Two similarly-colored and similarly-posed cats are face to face in one image.

New Text: Two differently-colored but similarly-posed chow dogs are face to face in one image. (Kaushik et al., 2019; Gardner et al., 2020)

A.5.7 Inter-split STS

This bin talks about the necessity of optimal dissimilarity between training and test sets. All the leads of the previous groups must be optimized within each spilt as well.

Variation of Split: Different split variations are required for proper benchmarking, to ensure a true accuracy increase. (Tan et al., 2019; Gorman and Bedrick, 2019) $\widehat{\delta} = M(G_{test}, S_1) - M(G_{test}, S_2)$

Accuracy difference: $\widehat{\delta}$

Model: M

Test Set: G_{test}

Systems 1 and 2: S_1, S_2

Innoculation Cost: (Richardson and Sabharwal, 2019; Nie et al., 2019)

Adversarial NLI:

Premise:A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term "melee" originates in the 1640s from the French word, which refers to hand-to-hand combat, a close quarters battle, a brawl, a con- fused fight, etc. Melee weapons can be broadly divided into three categories

Hypothesis: Melee weapons are good for ranged and hand-to-hand combat.

Disagreement: A particular annotator overuses the label of entailment, and marks very few samples as neutral. This pattern can be used as a bias by a model. (Reidsma and Carletta, 2008)

A.6 Evaluation: Artifact Case Study

Test cases have been developed to show the efficacy of DQI in our proposed data creation paradigm, with varying numbers of preexisting samples. We tune the hyperparameters proportionally, based on the dataset size. The value ranges for the DQI component colors are also set accordingly. DQI has been calculated for the following cases:

(i) No Preexisting Samples

(ii) 100 Preexisting Samples from the Good Split of the SNLI Test Set (random sampling done 10 times for a fair comparison)

In case (i), DQI of the new sample is calculated. In case (ii), first, DQI for the preexisting sample set is computed, as x_1. Then, the new sample is added and DQI is recalculated for the updated sample set, as x_2. The new samples, shown in Table 5, have been taken from a recent work on adversarial filtering, AFLite.

Then, the difference $\Delta x = x_1 - x_2$ is calculated. On the main interface, the crowd source worker views the colors of DQI components corresponding to Δx. The analyst views Δx as 'Sample' and x_2 as 'Dataset' component colors on the visualizations.

Dataset	Sample ID	Split	Label	DQI Color
SNLI	S7	Good	Entailment	
	S8			
	S9		Neutral	
	S10			
	S11		Contradiction	
	S12			
	S5	Bad	Entailment	
	S6			
	S3		Neutral	
	S4			
	S1		Contradiction	
	S2			
Story CLOZE	S1	Good	True	
	S2			
	S3		False	
	S4			
	S5	Bad	True	
	S6			
	S7		False	
	S8			
SQUAD 2.0	S1	Good	True	
	S2			
	S3		False	
	S4			
	S5	Bad	True	
	S6			
	S7		False	
	S8			
MNLI	S1	Good	Entailment	
	S2			
	S3		Neutral	
	S4			
	S5		Contradiction	
	S6			
	S7	Bad	Entailment	
	S8			
	S9		Neutral	
	S10			
	S11		Contradiction	
	S12			

Table 4: Evaluating VAIDA over the most sensitive DQI component, Intra-Sample Word Similarity. Successes: green/orange for good split, red/orange for bad split. Failures: red for good split, green for bad split.

We use DQI to identify artifacts over

four datasets: SNLI (Bowman et al., 2015), MNLI(Williams et al., 2017), SQUAD 2.0 (Rajpurkar et al., 2018), and Story CLOZE Task (Schwartz et al., 2017). In the case of SQUAD 2.0 and Story CLOZE, we split each sample into multiple samples– for e.g., in Story CLOZE there are two ending choices per sample and so we make two samples, with label *True* for the sample with the correct ending and *False* for the sample with the incorrect ending. The presence of a large number of artifacts has been shown in several studies on SNLI (Gururangan et al., 2018) and Story CLOZE Task (Schwartz et al., 2017). MNLI and SQUAD 2.0 have been shown to have a relatively smaller number of artifacts (Gururangan et al., 2018; Kaushik and Lipton, 2018), and therefore ensure adversarial evaluation of VAIDA. We evaluate each dataset using its test sets, or if unavailable, on its dev sets.

For fair comparison, we have taken illustrative samples from the AFLite paper (Bras et al., 2020) for SNLI (Table 5). We randomly sample for other datasets (Tables 6, 7 8, 9) as corresponding examples were not illustrated in those papers. There exist two hyperparameters separating the boundary between red, yellow, and green flags. We tune hyperparameters on 0.01% of data manually in a supervised manner (Mishra et al., 2020). This is analogous to how humans learn quickly from few samples.

Results: DQI component colors across settings are correctly predicted according to AFLite categorization of good and bad splits on an average of 10/12 times in SNLI, 5/8 times in SQUAD 2.0 and Story CLOZE, and 7/12 times in MNLI [6] as illustrated in Table 4. We convert SQUAD 2.0 and Story CLOZE into NLI format, with *answer* and *ending* corresponding to *hypothesis*, and *context* and *story* corresponding to *premise*, respectively.

Analysis: False positives and false negatives can be attributed to the limitation of AFLite in incorrectly classifying samples (Mishra et al., 2020). Additionally, we have two observations: (i) VAIDA's prediction accuracy decreases as the artifact level in a dataset decreases. (ii) The values of most DQI sub-components do not change significantly (<25% of the time) after adding samples in both categories. However, it changes considerably (>60% of the time) across two sub-components: Intra-sample word overlap and word similarity, both of which belong to the fifth component of DQI. This can again be explained by AFLite's sensitivity towards word overlap (Mishra et al., 2020).

A.6.1 Case(i) - Addressing Cold Start

Case (i) addresses the situation of cold-start for DQI. Unlike adversarial filtering algorithms, DQI can be used even with low data levels. In the situation of cold start, the component initialization (shown for SNLI samples from Table 5) is as follows:

Vocabulary: The first term is scaled appropriately as it takes the size of the dataset into account. The second term returns the standard deviation between the premise and hypothesis lengths. Since the third term defines upper and lower bounds on sentence length, it takes a value of one as long as the lengths of both the premise and hypothesis statements exceed three words, and zero if it is three words or less, as seen for sample 5 in Table 10.

Sample	Terms			DQI C1
	T1	T2	T3	
S1	0.0693	2.121	1.0000	2.1906
S2	0.0396	0.7071	1.0000	0.7467
S3	0.1089	2.1213	1.0000	2.2302
S4	0.1188	7.7781	1.0000	7.8969
S5	0.06930	5.6568	0.0000	0.0693
S6	0.1188	11.3137	1.0000	11.4325
S7	0.0594	0.0000	1.0000	0.0594
S8	0.0792	4.9497	1.0000	5.0289
S9	0.0693	1.4142	1.0000	1.4835
S10	0.0891	4.9497	1.0000	5.0388
S11	0.0990	2.8284	1.0000	2.9274
S12	0.1089	2.8284	1.0000	2.9373

Table 10: DQI_{C1} for Case (i)

Inter-sample N-gram Frequency and Relation: Term 1 captures the inverse of standard deviation, and hence yields infinity in the case of POS tags, when a word with that POS tag does not occur at all, or only occurs once as standard deviation tends to zero. In some cases, the standard deviation can be zero, as seen in Table 18 for trigrams, as each trigram occurs an equal number of times. High non-infinite values for term one are seen for bigrams and trigrams due to their balanced distributions in a sample, as in Table 21.

Sample ID	Premise	Hypothesis	Label	Split
S1	A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill.	contradiction	Dev-Bad
S2	The dog is catching a treat.	The cat is not catching a treat.	contradiction	Dev-Bad
S3	Three young men are watching a tennis match on a large screen outdoors.	Three young men watching a tennis match on a screen outdoors, because their brother is playing.	neutral	Dev-Bad
S4	A girl dressed in a pink shirt, jeans, and flip-flops sitting down playing with a lollipop machine.	A funny person in a shirt.	neutral	Dev-Bad
S5	A man in a green apron smiles behind a food stand.	A man smiles.	entailment	Dev-Bad
S6	A little girl with a hat sits between a woman's feet in the sand in front of a pair of colorful tents.	The girl is wearing a hat.	entailment	Dev-Bad
S7	People are throwing tomatoes at each other.	The people are having a food fight.	entailment	Dev-Good
S8	A man poses for a photo in front of a Chinese building by jumping.	The man is prepared for his photo.	entailment	Dev-Good
S9	An older gentleman speaking at a podium.	A man giving a speech.	neutral	Dev-Good
S10	A man poses for a photo in front of a Chinese building by jumping.	The man has experience in taking photos.	neutral	Dev-Good
S11	People are waiting in line by a food vendor.	People sit and wait for their orders at a nice sit down restaurant.	contradiction	Dev-Good
S12	Number 13 kicks a soccer ball towards the goal during children's soccer game.	A player passing the ball in a soccer game.	contradiction	Dev-Good

Table 5: SNLI Samples used for Test Cases

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	7	13.0958	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	4.0000	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	15	32.7698	0.1578	0
Trigrams	16	64.0000	0.7647	0

Table 11: $DQI_{C2}$and DQI_{C6} (contradiction) for S1, Case (i)

Sentences are seen to differ across samples in terms of the language used, and their length. Therefore, when setting the upper and lower bounds of granularities for Term 2, standardizing the bounds for cold start fails in the case of POS tags, particularly adverbs, as in seen Tables 11 - 22. These bounds therefore need to be reset at cold start particular to the sample's language.

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	4	6.9282	1.0000	0
Adjectives	0	nan	nan	0
Adverbs	0	nan	nan	0
Verbs	1	inf	1.0000	0
Nouns	3	6.3639	1.0000	0
Bigrams	9	20.4101	0.2727	0
Trigrams	8	22.6274	0.5555	0

Table 12: $DQI_{C2}$and DQI_{C6} (contradiction) for S2, Case (i)

Sample ID	Premise	Hypothesis	Label	Split
S1	To their good fortune, he's proving them right.	He is showing that they guessed correctly.	entailment	Dev-Good
S2	Strange as it may seem to the typical household, capital gains on its existing assets do not contribute to saving as measured in NIPA.	The increased equity of a house may not be considered as savings by NIPA.	entailment	Dev-Good
S3	Among runners-up is Boston solo Eleanor Newhoff.	Eleanor Newhoff had trained hard for the Olympic triathlon.	neutral	Dev-Good
S4	This was used for ceremonial purposes, allowing statues of the gods to be carried to the river for journeys to the west bank, or to the Luxor sanctuary.	Statues were moved to Luxor for funerals and other ceremonies.	neutral	Dev-Good
S5	Or just a philosophy of any weapon to hand?	They don't allow any weapon.	contradiction	Dev-Good
S6	Diets for men in their prime	A plan to keep men fat.	contradiction	Dev-Good
S7	Justice Kennedy does not care what law librarians across the country Reporters from 1790 through 1998.	Justice Kennedy doesn't care if do with all the Supreme Court the Supreme Court Reporters from 1790 to 1998 are thrown away.	entailment	Dev-Bad
S8	are you originally from uh Texas	You're originally from Texas?	entailment	Dev-Bad
S9	Click here for Finkelstein's explanation of why this logic is expedient.	Click here for Finkelstein's explanation of why this logic is expedient due to philosophical constraints.	neutral	Dev-Bad
S10	Two, most other productive operations are easier to study and understand, since few firms have 40,000 locations and a large proportion of their workforce working outdoors.	The productivity of the operations is directly related to the workforce that's based outdoors.	neutral	Dev-Bad
S11	Treat yourself and bill it to Si.	Don't treat yourself, Si has to pay for that.	contradiction	Dev-Bad
S12	Eh! Monsieur Lawrence, called Poirot.	Poirot did not call upon Monsieur Lawrence.	contradiction	Dev-Bad

Table 6: MNLI Samples used for Test Cases

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	11	23.5495	1.0000	0
Adjectives	3	6.3639	1.0000	0
Adverbs	0	6.3639	nan	0
Verbs	2	4.0000	1.0000	0
Nouns	5	12.5000	1.0000	0
Bigrams	19	37.4563	-0.1851	0
Trigrams	20	45.0185	0.2000	0

Table 13: $DQI_{C2}$and DQI_{C6} (neutral) for S3, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	12	41.5692	1.0000	0
Adjectives	3	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	4	inf	1.0000	0
Nouns	5	12.5000	1.0000	0
Bigrams	20	89.4427	0.8095	0
Trigrams	19	4.6757e+16	1.0000	0

Table 14: $DQI_{C2}$and DQI_{C6} (neutral) for S4, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	7	14.3457	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	11	36.4828	0.6667	0
Trigrams	10	6.8359e+16	1.0000	0

Table 15: $DQI_{C2}$and DQI_{C6} (entailment) for S5, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	12	30.8285	1.0000	0
Adjectives	3	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	7	20.0041	1.0000	0
Bigrams	25	125.0000	0.8461	0
Trigrams	24	7.0540e+16	1.0000	0

Table 16: $DQI_{C2}$and DQI_{C6} (entailment) for S6, Case (i)

Sample ID	Question	Context	Answer	impossible	Split
S1	By how many kilometers are shear waves separated when measuring the crust?	Seismologists can use the arrival times of seismic waves in reverse to image the interior of the Earth. Early advances in this field showed the existence of a liquid outer core (where shear waves were not able to propagate) and a dense solid inner core. These advances led to the development of a layered model of the Earth, with a crust and lithosphere on top, the mantle below (separated within itself by seismic discontinuities at 410 and 660 kilometers), and the outer core and inner core below that. More recently, seismologists have been able to create detailed images of wave speeds inside the earth in the same way a doctor images a body in a CT scan. These images have led to a much more detailed view of the interior of the Earth, and have replaced the simplified layered model with a much more dynamic model.	at 410 and 660 kilometers	True	Dev-Good
S2	Where is Geoffrey Parker from?	The plague repeatedly returned to haunt Europe and the Mediterranean throughout the 14th to 17th centuries. According to Biraben, the plague was present somewhere in Europe in every year between 1346 and 1671. The Second Pandemic was particularly widespread in the following years: 1360–63; 1374; 1400; 1438–39; 1456–57; 1464–66; 1481–85; 1500–03; 1518–31; 1544–48; 1563–66; 1573–88; 1596–99; 1602–11; 1623–40; 1644–54; and 1664–67. Subsequent outbreaks, though severe, marked the retreat from most of Europe (18th century) and northern Africa (19th century). According to Geoffrey Parker, "France alone lost almost a million people to the plague in the epidemic of 1628–31."	France	True	Dev-Good
S3	When was the European Convention on Human Rights established?	None of the original treaties establishing the European Union mention protection for fundamental rights. It was not envisaged for European Union measures, that is legislative and administrative actions by European Union institutions, to be subject to human rights. At the time the only concern was that member states should be prevented from violating human rights, hence the establishment of the European Convention on Human Rights in 1950 and the establishment of the European Court of Human Rights. The European Court of Justice recognised fundamental rights as general principle of European Union law as the need to ensure that European Union measures are compatible with the human rights enshrined in member states' constitution became ever more apparent. In 1999 the European Council set up a body tasked with drafting a European Charter of Human Rights, which could form the constitutional basis for the European Union and as such tailored specifically to apply to the European Union and its institutions. The Charter of Fundamental Rights of the European Union draws a list of fundamental rights from the European Convention on Human Rights and Fundamental Freedoms, the Declaration on Fundamental Rights produced by the European Parliament in 1989 and European Union Treaties.	1950	False	Dev-Good
S4	What did Lavoisier perceive the air had lost as much as the tin had gained?	In one experiment, Lavoisier observed that there was no overall increase in weight when tin and air were heated in a closed container. He noted that air rushed in when he opened the container, which indicated that part of the trapped air had been consumed. He also noted that the tin had increased in weight and that increase was the same as the weight of the air that rushed back in. This and other experiments on combustion were documented in his book Sur la combustion en général, which was published in 1777. In that work, he proved that air is a mixture of two gases; 'vital air', which is essential to combustion and respiration, and azote ("lifeless"), which did not support either. Azote later became nitrogen in English, although it has kept the name in French and several other European languages.	weight	False	Dev-Good

Table 7: SQUAD 2.0 Test Cases - Dev Good

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	6	14.6969	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	4	9.2376	1.0000	0
Bigrams	11	36.4828	0.6667	0
Trigrams	10	6.8359e+16	1.0000	0

Table 17: $DQI_{C2}$and DQI_{C6} (entailment) for S7, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	8	17.2819	1.0000	0
Adjectives	2	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	inf	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	19	4.6757e+16	1.0000	0
Trigrams	17	inf	1.0000	0

Table 18: $DQI_{C2}$and DQI_{C6} (entailment) for S8, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	7	3.3356e+16	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	inf	1.0000	0
Nouns	4	inf	1.0000	0
Bigrams	10	6.8359e+16	1.0000	0
Trigrams	8	inf	1.0000	0

Table 19: $DQI_{C2}$and DQI_{C6} (neutral) for S9, Case (i)

Sample ID	Question	Context	Answer	impossible	Split
S5	Why are normal body cells attacked by NK cells?	Natural killer cells, or NK cells, are a component of the innate immune system which does not directly attack invading microbes. Rather, NK cells destroy compromised host cells, such as tumor cells or virus-infected cells, recognizing such cells by a condition known as "missing self." This term describes cells with low levels of a cell-surface marker called MHC I (major histocompatibility complex) – a situation that can arise in viral infections of host cells. They were named "natural killer" because of the initial notion that they do not require activation in order to kill cells that are "missing self." For many years it was unclear how NK cells recognize tumor cells and infected cells. It is now known that the MHC makeup on the surface of those cells is altered and the NK cells become activated through recognition of "missing self". Normal body cells are not recognized and attacked by NK cells because they express intact self MHC antigens. Those MHC antigens are recognized by killer cell immunoglobulin receptors (KIR) which essentially put the brakes on NK cells.	express intact self MHC antigens	True	Dev-Bad
S6	What did higher material living standards lead to for most of human history?	For most of human history higher material living standards – full stomachs, access to clean water and warmth from fuel – led to better health and longer lives. This pattern of higher incomes-longer lives still holds among poorer countries, where life expectancy increases rapidly as per capita income increases, but in recent decades it has slowed down among middle income countries and plateaued among the richest thirty or so countries in the world. Americans live no longer on average (about 77 years in 2004) than Greeks (78 years) or New Zealanders (78), though the USA has a higher GDP per capita. Life expectancy in Sweden (80 years) and Japan (82) – where income was more equally distributed – was longer.	better health and longer lives	True	Dev-Bad
S7	What happens as they build phase 1?	The owner produces a list of requirements for a project, giving an overall view of the project's goals. Several D&B contractors present different ideas about how to accomplish these goals. The owner selects the ideas he or she likes best and hires the appropriate contractor. Often, it is not just one contractor, but a consortium of several contractors working together. Once these have been hired, they begin building the first phase of the project. As they build phase 1, they design phase 2. This is in contrast to a design-bid-build contract, where the project is completely designed by the owner, then bid on, then completed.	they design phase 2	False	Dev-Bad
S8	When was the Third Assessment Report published?	Another example of scientific research which suggests that previous estimates by the IPCC, far from overstating dangers and risks, have actually understated them is a study on projected rises in sea levels. When the researchers' analysis was "applied to the possible scenarios outlined by the Intergovernmental Panel on Climate Change (IPCC), the researchers found that in 2100 sea levels would be 0.5–1.4 m [50–140 cm] above 1990 levels. These values are much greater than the 9–88 cm as projected by the IPCC itself in its Third Assessment Report, published in 2001". This may have been due, in part, to the expanding human understanding of climate.	2001	False	Dev-Bad

Table 8: SQUAD 2.0 Test Cases - Dev Bad

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	9	20.4100	1.0000	0
Adjectives	3	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	inf	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	19	4.6757e+16	1.0000	0
Trigrams	17	4.6757e+16	1.0000	0

Table 20: $DQI_{C2}$and DQI_{C6} (neutral) for S10, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	10	23.7170	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	8	18.4752	1.0000	0
Bigrams	20	1.4046e+17	1.0000	0
Trigrams	18	7.0027e+16	1.0000	0

Table 21: $DQI_{C2}$and DQI_{C6} (contradiction) for S11, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	11	16.3156	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	8	11.3137	1.0000	0
Bigrams	18	55.6619	0.6000	0
Trigrams	18	7.0027e+16	1.0000	0

Table 22: $DQI_{C2}$and DQI_{C6} (contradiction) for S12, Case (i)

Inter-sample STS: The first term focuses on the standard deviation of similarity values that cross a threshold between all sentences. Since there is only one similarity value calculated, the value of Term 1, as in Table 23, is set to that similarity value to prevent it from becoming infinity. The second term is always taken to have a value of 2, as there is no definite set threshold for taking a maximum.

Sample ID	Story	Ending	Label	Split
S1	Fred receives a specialty coffee maker for Christmas. He finally opens it after leaving it in its box for a few weeks.Fred decides to make himself a cappuccino.To his surprise, it tastes just as good as the ones he buys outside.	Frank will save about $25 a week making coffee himself.	True	Dev-Good
S2	My family is sharing a bowl of popcorn.Mom is reading a book and eating one piece at a time.Dad and I are playing iPad games and eating handfuls at a time.We have played this game before!	Dad and I love popcorn.	True	Dev-Good
S3	I got a job as a shopping mall Santa last December. The hours were long.The pay was bad.But I found interacting with the kids to be completely amazing.	I found that playing Santa was not worth my time off.	False	Dev-Good
S4	Carry has been short her whole life.She could never reach the top shelf at the store.Greg saw her struggling to reach.He went over and helped her.	She refused his help and walked away.	False	Dev-Good
S5	Lou was on a diet.She was eating very little.But she still struggled to lose weight!Then she added an exercise regimen.	Lou was finally able to lose weight.	True	Dev-Bad
S6	Kim had been working extra hard for weeks.She learned of a promotion up for grabs at her company.It came with a new office and great benefits.Finally all her work paid off and she was offered the promotion.	She was happy to get the promotion.	True	Dev-Bad
S7	James has just started working at a company with a ping pong table.He has always wanted to play ping pong with a coworker.One day after work, his friend challenges him to a game.James plays very well, but eventually loses the game.	James was worried because he beat his boss at ping pong.	False	Dev-Bad
S8	Dan loves the sport of bowling.His dad taught him how to play when he was little.The use to compete in tournaments together.His dad has since passed away.	Dan never liked to bowl anyway.	False	Dev-Bad

Table 9: Story CLOZE Test Cases

Intra-sample Word Simlarity: The fourth component scales appropriately, as it takes the size of the dataset into account and can therefore be directly computed, as in Table 23.

Sample	DQI C3 - T1	DQI C3 - T2	DQI C4
S1	0.8938	2.0	0.9896
S2	0.9060	2.0	0.7779
S3	0.8722	2.0	1.3180
S4	0.6512	2.0	0.9093
S5	0.6982	2.0	0.0848
S6	0.6806	2.0	1.1088
S7	0.7443	2.0	0.6826
S8	0.7672	2.0	1.0860
S9	0.8219	2.0	0.5084
S10	0.7750	2.0	0.9601
S11	0.7616	2.0	1.1597
S12	0.8255	2.0	1.2076

Table 23: T1 and T2 for DQI_{C3}, DQI_{C4}, Case (i)

Intra-sample STS: The first term, in Table 24, deals with whether the Premise-Hypothesis similarity crosses a threshold. This scales as it takes dataset size into account, and can be calculated for different threshold values. The second and third terms, Table 25, involve the calculation of the mean and standard deviation of length difference between the premise and hypothesis. Therefore, the second term is directly computed, while the third is always zero, since only one value is present. The fourth term's value, in Table 25, also uses standard deviation and is directly taken to be the similarity between the premise and hypothesis, as only one value is calculated. The fifth and sixth terms look at word overlap and word similarity levels between the premise and hypothesis, and can be directly calculated. These are represented in Tables 27 - 30.

Sample Set	Terms T1		
	ISIM=0.5	ISIM=0.6	ISIM=0.7
+S1	2.53901172	3.40305015	5.15852057
+S2	2.46282325	3.26756734	4.85347200
+S3	2.68605483	3.67251159	5.80405898
+S4	6.61292347	19.5239860	20.4998054
+S5	5.04523160	10.1825780	557.710874
+S6	5.53586344	12.4007484	51.6536766
+S7	4.09274400	6.92833358	22.5556185
+S8	3.74140198	5.97801932	14.8633715
+S9	3.10654715	4.50651832	8.20339191
+S10	3.6359872	5.71335622	13.3282739
+S11	3.8217013	6.18568557	16.2170311
+S12	3.0714259	4.43298421	7.96294530

Table 24: T1 for DQI_{C5}, Case (i)

Sample	DQI C5 -T2,C6 - T3	DQI C5 - T3,C6 - T4	DQI C5 - T4
S1	0.2500	nan	0.8938
S2	0.5000	nan	0.9060
S3	0.2500	nan	0.8722
S4	0.0830	nan	0.6512
S5	0.1111	nan	0.6982
S6	0.0588	nan	0.6806
S7	1.0000	nan	0.7443
S8	0.1250	nan	0.7672
S9	0.3333	nan	0.8219
S10	0.1250	nan	0.7750
S11	0.2000	nan	0.7616
S12	0.2000	nan	0.8255

Table 25: T2/3 and T3/4 for DQI_{C5}/DQI_{C6}, T4 for DQI_{C5} , Case (i)

N-gram Frequency per Label: Since cold start only involves the text data of a single sample, the label of that sample is the only one with initialized values in DQI_{C6}. Table 24 has Terms 1 and 2 of DQI_{C6}, as they are equivalent to the terms of DQI_{C2} for the label of the new sample. These terms are set to zero for the other two labels. Table 25 has Terms 3 and 4, which are the same as terms 2 and 3 of DQI_{C5}, and are only computed for the label of the new sample. Also, since the counts of all granularities are only initialized for a single label, the fifth term is set to zero for all samples.

Inter-split STS: Since DQI_{C7} is calculated on the basis of the most similar training sample for every test set sample, it is not applicable to the case of cold start, as there is only one sample. Hence, its value is taken as zero.

Sample	DQI C1	DQI C2	DQI C3	DQI C4	DQI C5 (ISIM=0.5)	DQI C6	DQI C7
S1	2.1906	80.2076	2.8938	0.9896	12.3961	80.4576	0
S2	0.7467	32.4274	2.9060	0.7779	9.7696	32.9274	0
S3	2.2302	49.4839	2.8722	1.3180	15.0742	49.7339	0
S4	7.8969	4.6757E+16	2.6512	0.9093	18.2884	4.6757E+16	0
S5	0.0693	6.8359E+16	2.6982	0.0848	16.3837	6.8359E+16	0
S6	11.4325	7.0540E+16	2.6806	1.1088	23.0456	7.054E+16	0
S7	0.0594	6.8359E+16	2.7443	0.6826	16.4604	6.8359E+16	0
S8	5.0289	4.6757E+16	2.7672	1.0860	15.8438	4.6757E+16	0
S9	1.4835	1.0171E+17	2.8219	0.5084	77.4403	1.01715E+17	0
S10	5.0388	9.3514E+16	2.7750	0.9601	16.2461	9.3514E+16	0
S11	2.9274	2.1048E+17	2.7616	1.1597	20.1601	2.10487E+17	0
S12	2.9373	7.0027E+16	2.8255	1.2076	16.6541	7.0027E+16	0

Table 26: DQI Terms, Case (i)

Sample	Overlap Count	length(hypothesis) / Overlap Count
S1	3	**2.0000**
S2	2	**1.5000**
S3	8	**1.1250**
S4	1	10.0000
S5	2	**3.5000**
S6	2	**5.5000**
S7	1	**4.0000**
S8	2	3.5000
S9	0	**40.0000**
S10	2	3.5000
S11	1	**5.0000**
S12	3	3.0000

Table 27: Word Overlap, Red: $<$ 3.9375, Yellow: 3.9375-9.8333 Green: $>$ 9.8333

Sample	Overlap Count	length(hypothesis+premise) / Overlap Count
S1	3	**3.3333**
S2	2	**3.0000**
S3	8	**2.3750**
S4	1	**13.0000**
S5	2	**4.5000**
S6	2	**7.0000**
S7	1	**7.0000**
S8	2	5.0000
S9	0	**70.0000**
S10	2	5.5000
S11	1	**11.0000**
S12	3	4.6667

Table 28: Word Overlap, Red: $<$ 5.5347, Yellow: 5.5347-17.1944 Green: $>$ 17.1944

Sample	Premise Word Count	Hypothesis Word Count	Sum of Word Similarities
S1	10	9	5.4753
S2	6	7	2.7865
S3	12	15	**8.9008**
S4	15	6	**9.8715**
S5	9	3	6.5202
S6	17	6	**29.0358**
S7	7	6	**3.6143**
S8	12	7	**6.5335**
S9	7	5	**3.6679**
S10	127	7	**6.0583**
S11	9	12	**4.3558**
S12	12	9	28.5806

Table 29: Word Similarity With Stop Words, Red: $>$ 10.4317, Yellow: 8.8017-10.4317 Green: $<$ 8.8017

Sample	Premise Word Count	Hypothesis Word Count	Sum of Word Similarities
S1	6	4	**5.3800**
S2	3	3	2.9008
S3	10	9	**8.8910**
S4	10	3	**7.9413**
S5	7	2	**6.0292**
S6	11	3	**9.7704**
S7	4	3	**3.6234**
S8	7	3	**6.2102**
S9	4	3	**3.1786**
S10	7	4	**6.2102**
S11	5	6	**4.3768**
S12	9	5	7.8905

Table 30: Word Similarity Without Stop Words, Red: $>$ 6.8188, Yellow: 5.2483-6.8188 Green: $<$ 5.2483

A.6.2 Case(ii)-Adding to the Test Good Split

A 100 samples are taken at random 10 times from the good split of the SNLI Test set and x_1 is calculated. Then the new sample is added to the dataset. x_2 and Δx are calculated. For all components, DQI values are calculated using the same hyperparameter values as those used for the full test set. The results, shown in Tables 31 - 46, indicate the need for hyperparameter scaling.

What requires Scaling? From tables 38 and 35-41, we find that hyperparameters used to set upper and lower bounds for POS tag frequencies across and within labels require significant scaling. Additionally, we find that sentence, bigram, and trigram terms should be omitted when calculating the DQI until their overall frequencies and variance reach a certain threshold. This is because terms inversely proportional to the standard deviation of the distributions of those granularities are found to explode for lesser numbers of samples.

A.6.3 Assigning Colors

The new sample set has six samples removed by AFLite, that from the bad split of the Dev set, and six that are retained, i.e.,from the good split of the Dev set. In both case (i) and case (ii), we find that on adding samples to the existing dataset, there is no significant difference in the term/component values except in the cases of word overlap and word similarity, seen in T5 and T6 of DQI_{C5}. We observe that DQI component colors are correctly predicted 10/12 times on an average. Also, the change in DQI_{C5} corresponding to word overlap and word similarity is as expected as per the findings of AFLite.

Sample Set	Terms			DQI C1
	T1	T2	T3	
Original	5.8200	6.6656	0.9300	12.0190
+S1	5.7921	6.6347	0.9307	11.9669
+S2	5.7822	6.6507	0.9307	11.9719
+S3	5.8020	6.6409	0.9307	11.9826
+S4	5.8119	6.6550	0.9307	12.0056
+S5	5.7723	6.6590	0.9208	11.9038
+S6	5.7822	6.6849	0.9307	12.0038
+S7	5.7822	6.6470	0.9307	11.9685
+S8	5.7921	6.6422	0.9307	11.9739
+S9	5.8020	6.6551	0.9307	11.9958
+S10	5.7921	6.6422	0.9307	11.9739
+S11	5.7921	6.6355	0.9307	11.9677
+S12	5.8317	6.6355	0.930	12.0073

Table 31: DQI_{C1} for Case (ii)

Sample Set	DQI C4
Original	0.00657581
+S1	0.00653241
+S2	0.00652070
+S3	0.00654317
+S4	0.00652860
+S5	0.00610259
+S6	0.00653705
+S7	0.00651307
+S8	0.00653624
+S9	0.00649185
+S10	0.00653108
+S11	0.00653874
+S12	0.00654020

Table 32: DQI_{C4} for Case (ii)

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	7.1303e+16	1.0000	1045.3358	2.0833	7.1303e+16	1.0000	92.8203
+S1	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485
+S2	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485
+S3	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S4	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S5	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S6	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S7	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S8	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S9	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S10	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S11	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485
+S12	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485

Table 33: Case (ii), Sentence Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	113.4748	0.5548	136.5557	0.6599	105.1059	0.5255	2.4416
+S1	113.4748	0.5548	136.5557	0.6599	103.7067	0.5219	2.4509
+S2	113.4748	0.5548	136.5557	0.6599	107.3208	0.5339	2.4325
+S3	113.4748	0.5548	137.7114	0.6182	105.1059	0.5255	2.3670
+S4	113.4748	0.5548	138.5993	0.6422	105.1059	0.5255	2.4336
+S5	109.7512	0.5298	136.5557	0.6599	105.1059	0.5255	2.4566
+S6	117.4812	0.5679	136.5557	0.6599	105.1059	0.5255	2.4518
+S7	115.2611	0.5520	136.5557	0.6599	105.1059	0.5255	2.4241
+S8	110.1518	0.5562	136.5557	0.6599	105.1059	0.5255	2.4491
+S9	113.4748	0.5548	136.5917	0.6604	105.1059	0.5255	2.4467
+S10	113.4748	0.5548	134.4891	0.6595	105.1059	0.5255	2.4267
+S11	113.4748	0.5548	136.5557	0.6599	110.1129	0.5304	2.4310
+S12	113.4748	0.5548	136.5557	0.6599	112.6038	0.5459	2.4524

Table 34: Case (ii), Word Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	65.4824	0.1935	48.9086	0.1130	44.8057	-0.2113	2.6514
+S1	74.6675	0.0909	50.8008	0.1500	57.0071	0.0164	2.8685
+S2	61.3138	-0.0588	52.7111	0.0815	51.3651	-0.1351	3.1961
+S3	76.2138	0.0588	46.8815	0.1339	60.6168	0.0476	3.0158
+S4	62.4955	-0.0423	58.8794	0.2480	52.4764	-0.1389	3.2262
+S5	71.8135	-0.0133	48.3257	0.1707	57.2251	0.0667	2.9149
+S6	71.5360	0.0571	50.7164	0.1897	49.4934	0.0000	2.5007
+S7	69.5736	0.1475	52.5575	0.0676	58.1186	0.0312	2.6028
+S8	73.1520	0.1250	45.2213	0.1000	51.0064	0.0149	2.7511
+S9	68.4000	0.0000	48.3109	0.0615	52.7210	0.0000	2.8224
+S10	72.3354	0.0684	48.7879	0.1147	53.0237	0.0667	3.0774
+S11	68.2115	-0.0410	47.9655	0.1355	50.9620	-0.0294	2.6320
+S12	74.7011	0.0000	51.4393	0.0518	45.1122	-0.1384	2.6840

Table 35: Case (ii), Adjective Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	18.4752	0.2000	21.4630	0.1765	6.3640	0.0000	5.1159
+S1	3.6029e+16	1.0000	16.4141	-0.0769	6.3640	0.0000	3.0036
+S2	10.0021	0.3333	13.4297	0.2632	9.2376	0.0000	2.9621
+S3	16.0997	0.4287	25.0000	0.3333	6.3640	0.0000	4.8231
+S4	inf	1.0000	20.8025	0.0000	9.2376	0.2000	3.4788
+S5	20.0042	0.5000	19.2428	0.1250	12.5	0.3333	4.2973
+S6	inf	1.0000	21.4630	0.1765	6.3639	0.0000	2.9468
+S7	28.6378	0.6000	19.0918	0.0000	6.3639	0.0000	3.5977
+S8	18.4752	0.2000	27.6955	0.4444	9.2376	0.2000	3.4223
+S9	21.6481	0.2727	28.6216	0.3000	6.3639	0.0000	5.3589
+S10	8.0632	-0.2307	19.2428	0.1250	9.6096	0.0000	4.3729
+S11	inf	1.0000	19.2428	0.1250	9.2376	0.2000	4.0262
+S12	inf	1.0000	23.7684	0.2222	6.3639	0.0000	4.1769

Table 36: Case (ii), Adverb Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	65.4824	0.1935	51.9736	-0.0598	35.1110	-0.1081	2.7836
+S1	40.3696	-0.2069	48.5430	-0.1525	29.9195	-0.2405	2.4728
+S2	43.9037	-0.2424	53.3506	-0.0093	30.1625	-0.0909	2.6133
+S3	37.4444	-0.3030	56.2047	-0.1057	27.3594	-0.2286	2.3308
+S4	42.1040	-0.3333	46.2161	-0.0973	31.2449	-0.1667	2.5586
+S5	38.3571	-0.3714	50.6384	-0.0182	24.4386	-0.2000	2.5610
+S6	41.7648	-0.2537	48.9552	-0.0280	28.8722	-0.1642	2.7063
+S7	46.5989	-0.2537	53.4887	-0.1260	31.1722	-0.2500	2.2977
+S8	35.4040	-0.3548	48.3655	-0.0990	26.0207	-0.2615	2.7680
+S9	40.6156	-0.2000	53.4014	-0.1056	32.0340	-0.2307	2.5957
+S10	41.3657	-0.3230	53.0775	-0.0847	29.1653	-0.2876	2.2606
+S11	42.3999	-0.2187	46.3814	-0.1452	33.3842	-0.1267	2.6794
+S12	37.5858	-0.2258	49.7109	-0.1071	26.0396	-0.0667	2.6669

Table 37: Case (ii), Verb Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	42.7808	-0.3056	53.6301	0.2841	38.7466	-0.2050	2.3372
+S1	38.3026	-0.3659	52.7785	0.2989	39.4878	-0.2601	2.4916
+S2	35.9868	-0.2752	51.9745	0.3097	41.0652	-0.2558	2.3264
+S3	36.7162	-0.3247	52.4598	0.2667	41.5999	-0.2485	2.3551
+S4	36.7565	-0.2617	53.2731	0.2570	37.4839	-0.2075	2.3918
+S5	33.0670	-0.2752	54.0598	0.3030	44.1367	-0.2817	2.3645
+S6	38.3611	-0.3250	54.9709	0.3040	42.2864	-0.2528	2.5035
+S7	37.7188	-0.3414	51.8644	0.2844	37.6200	-0.2327	2.6013
+S8	38.9773	-0.3254	55.4119	0.3028	41.6562	-0.2441	2.4018
+S9	35.4958	-0.3200	50.3967	0.3313	39.9118	-0.2121	2.4067
+S10	32.9868	-0.2765	52.1225	0.2954	38.6028	-0.2484	2.4450
+S11	36.0093	-0.3333	55.2239	0.3352	42.8904	-0.2402	2.4570
+S12	34.8526	-0.3509	50.4304	0.3113	51.0263	-0.2448	2.5026

Table 41: Case (ii), Noun Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	497.2044	0.8411	620.1037	0.9075	415.2737	0.8610	0.7924
+S1	497.2043	0.8411	620.1037	0.9075	403.4774	0.8206	0.7928
+S2	497.2043	0.8411	620.1037	0.9075	427.4754	0.8636	0.7917
+S3	497.2043	0.8411	625.7171	0.8873	415.2737	0.8610	0.7694
+S4	497.2043	0.8411	616.7056	0.9055	415.2737	0.8610	0.7864
+S5	473.5139	0.8528	620.1037	0.9075	415.2737	0.8610	0.8045
+S6	518.7792	0.8684	620.1037	0.9075	415.2737	0.8610	0.8088
+S7	503.1652	0.8648	620.1037	0.9075	415.2737	0.8610	0.7960
+S8	491.4631	0.8588	620.1037	0.9075	415.2737	0.8610	0.8069
+S9	497.2043	0.8411	617.3021	0.9064	415.2737	0.8610	0.7986
+S10	497.2043	0.8411	619.8558	0.9072	415.2737	0.8610	0.7936
+S11	497.2043	0.8411	620.1037	0.9075	437.4726	0.8657	0.8003
+S12	497.2043	0.8411	620.1037	0.9075	427.2611	0.8623	0.7915

Table 42: Case (ii), Bigram Granularity Terms in DQI_{C6}

Sample Set	Terms						
	entailment		neutral		contradiction		
	T1	T2	T1	T2	T1	T2	T5
Original	1567.0110	0.7652	2174.6543	0.7302	1135.1086	0.7193	1.7297
+S1	1567.0110	0.7652	2174.6543	0.7302	1154.0280	0.7094	1.7212
+S2	1567.0110	0.7652	2174.6543	0.7302	1157.8255	0.8636	1.7298
+S3	1567.0110	0.7652	2215.9640	0.7163	1135.1086	0.7193	1.6799
+S4	1567.0110	0.7652	2245.9485	0.7355	1135.1086	0.7193	1.7383
+S5	1517.6459	0.7571	2174.6543	0.7302	1135.1086	0.7193	1.7468
+S6	1642.3849	0.7601	2174.6543	0.7302	1135.1086	0.7193	1.7383
+S7	1593.6394	0.7615	2174.6543	0.7302	1135.1086	0.7193	1.7406
+S8	1529.5108	0.7521	2174.6543	0.7302	1135.1086	0.7193	1.7470
+S9	1567.0110	0.7652	2204.5792	0.7324	1135.1086	0.7193	1.7470
+S10	1567.0110	0.7652	2190.9585	0.7245	1135.1086	0.7193	1.7235
+S11	1567.0110	0.7652	2174.6543	0.7302	1199.7393	0.7288	1.7470
+S12	1567.0110	0.7652	2174.6543	0.7302	1199.7393	0.7288	1.7383

Table 43: Case (ii), Trigram Granularity Terms in DQI_{C6}

Sample Set	Terms					
	entailment		neutral		contradiction	
	T3	T4	T3	T4	T3	T4
Original	0.1846	0.2003	0.1465	0.1226	0.1008	0.3662
+S1	0.1846	0.2003	0.1465	0.1226	0.1037	0.3485
+S2	0.1846	0.2003	0.1465	0.1226	0.1046	0.3514
+S3	0.1846	0.2003	0.1480	0.1195	0.1008	0.3662
+S4	0.1846	0.2003	0.1448	0.1195	0.1008	0.3662
+S5	0.1811	0.1894	0.1465	0.1226	0.1008	0.3662
+S6	0.1712	0.2065	0.1465	0.1226	0.1008	0.3662
+S7	0.1923	0.1931	0.1465	0.1226	0.1008	0.3662
+S8	0.1824	0.1887	0.1465	0.1226	0.1008	0.3662
+S9	0.1846	0.2003	0.1484	0.1197	0.1008	0.3662
+S10	0.1846	0.2003	0.1464	0.1191	0.1008	0.3662
+S11	0.1846	0.2003	0.1465	0.1226	0.1033	0.3473
+S12	0.1846	0.2003	0.1465	0.1226	0.1033	0.3473

Table 44: Terms 3 and 4 in DQI_{C6} for Case (ii)

Sample Set	Sentences		Words		Adjectives		Adverbs		Verbs		Nouns		Bigrams		Trigrams		DQI C2
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	
Original	2807.2405	0.9800	137.2755	0.6371	52.0534	0.3111	20.0385	-0.04	46.8398	-0.025	54.2786	0.3888	707.8112	0.8852	2723.6406	0.8910	5927.1970
+S1	2849.6668	0.9802	137.0171	0.6368	55.6705	0.3065	21.7786	-0.1111	50.8642	-0.0356	49.5464	0.3452	697.9764	0.8815	2706.4317	0.8857	5922.7847
+S2	2849.6668	0.9802	137.0171	0.6368	55.6705	0.3065	21.7789	-0.1111	50.8642	-0.0356	49.5464	0.3452	697.9764	0.8815	2706.4317	0.8857	5922.7847
+S3	2849.6668	0.9802	137.9140	0.6393	52.6620	0.2414	17.4592	0.0833	43.8252	-0.0661	55.2815	0.3505	712.9377	0.8847	2763.8091	0.8924	6009.2173
+S4	2849.6668	0.9802	138.3361	0.6392	54.2001	0.2576	24.9929	0.1250	48.5320	-0.0313	50.1523	0.3498	706.9163	0.9043	2765.4396	0.8921	6021.0912
+S5	2849.6668	0.9802	135.4295	0.6365	49.2904	0.2619	23.3950	0.0000	49.0989	-0.0840	52.0959	0.3432	697.8102	0.9029	2649.2411	0.8895	5892.6612
+S6	2849.6668	0.9802	137.1086	0.6379	53.9239	0.3609	20.0385	-0.0400	48.0375	-0.0538	52.8044	0.3463	711.5407	0.9064	2723.0651	0.8903	5984.3517
+S7	2849.6668	0.9802	137.4205	0.6359	48.4367	0.2015	35.9211	0.1538	45.0502	-0.0361	54.6786	0.4303	710.2298	0.9058	2739.3807	0.8916	6003.5736
+S8	2849.6668	0.9802	136.2514	0.6368	49.6075	0.2268	57.0399	0.3846	49.9798	-0.0445	52.5582	0.3432	705.7911	0.9052	2693.8612	0.8888	5962.1966
+S9	2849.6668	0.9802	137.6593	0.6375	58.2917	0.3388	24.5189	-0.0244	52.4063	0.0041	50.5623	0.3237	707.6845	0.9048	2742.9126	0.8915	6002.3536
+S10	2849.6668	0.9802	136.2477	0.6371	56.5772	0.2511	29.8974	-0.1034	51.6379	-0.0206	51.8621	0.3484	708.3581	0.9052	2718.4279	0.8899	5968.5017
+S11	2849.6668	0.9802	137.7623	0.6373	49.6725	0.2197	20.5196	-0.0667	47.5031	-0.0370	54.6531	0.3741	717.2547	0.9062	2767.0664	0.8921	6027.7480
+S12	2849.6668	0.9802	139.5281	0.6413	59.9832	0.3101	15.2008	-0.2727	52.8410	0.0723	50.6446	0.3174	713.8007	0.9052	2763.0228	0.8920	6027.8220

Table 38: DQI_{C2} for Case (ii)

Sample Set	Terms					DQI C3 (e=0.5)			
	T1			T2 (SIM=0.5)					
	SIM=0.5	SIM=0.6	SIM=0.7	e=0.25	e=0.33	e=0.5	SIM=0.5	SIM=0.6	SIM=0.7
Original	14.1194	4.9647	4.2968	200.0000	200.0000	198.4692	212.5886	203.4339	202.766
+S1	14.0959	4.9880	4.2882	202.0000	202.0000	199.9066	214.0025	204.8946	204.1948
+S2	14.2729	4.8939	4.3000	202.0000	202.0000	200.9450	215.2179	205.8389	205.245
+S3	14.1055	4.9749	4.2710	202.0000	202.0000	199.9066	214.0121	204.8815	204.1776
+S4	14.1285	4.9797	4.3134	202.0000	202.0000	200.4539	214.5824	205.4336	204.7673
+S5	14.1522	4.9797	4.3072	202.0000	202.0000	200.4539	214.6061	205.4336	204.7611
+S6	14.1961	4.9827	4.3041	202.0000	202.0000	200.4539	214.65	205.4366	204.758
+S7	14.1656	4.9842	4.3197	202.0000	202.0000	200.4539	214.6195	205.4381	204.7736
+S8	14.2711	4.9873	4.3015	202.0000	202.0000	200.9450	215.2161	205.9323	205.2465
+S9	14.2321	4.9836	4.3214	202.0000	202.0000	200.9450	215.1771	205.9286	205.2664
+S10	14.2859	4.9888	4.2944	202.0000	202.0000	200.9450	215.2309	205.9338	205.2394
+S11	14.1403	4.9720	4.3122	202.0000	202.0000	200.4539	214.5942	205.4259	204.7661
+S12	14.1707	4.9874	4.3211	202.0000	202.0000	199.9066	214.0773	204.894	204.2277

Table 39: DQI_{C3} for Case (ii)

Sample Set	DQI C6
Original	228.3537
+S1	202.4647
+S2	197.6054
+S3	196.3454
+S4	196.1489
+S5	200.7986
+S6	213.8920
+S7	202.4102
+S8	202.2893
+S9	198.4766
+S10	202.7345
+S11	200.9509
+S12	197.8010

Table 45: DQI_{C6} for Case (ii)

Sample Set	DQI C7		
	SSIM=0.2	SSIM=0.3	SSIM=0.4
Original	0.00304989	0.00421324	0.00629840
+S1	0.00189475	0.00229266	0.00290212
+S2	0.00216703	0.00270372	0.00359374
+S3	0.00186796	0.00225356	0.00283975
+S4	0.00196072	0.00238996	0.00305981
+S5	0.00188903	0.00228429	0.00288872
+S6	0.00190351	0.00230549	0.00292271
+S7	0.00201427	0.00247000	0.00319224
+S8	0.00187124	0.00225832	0.00284732
+S9	0.00197442	0.00241034	0.00309330
+S10	0.001886216	0.00228017	0.00288214
+S11	0.002048964	0.00252237	0.00328026
+S12	0.002076182	0.00256374	0.00335058

Table 46: DQI_{C7} for Case (ii)

A.6.4 Results Across Datasets

The following tables contain DQI component values across the sets of samples from Tables 5-9 in SNLI, MNLI, SQUAD 2.0, and Story CLOZE. Here, 'Good' denotes samples present in the 'Good' split of AFLite and 'Bad' denotes samples present in the 'Bad' Split of AFLite respectively.

Parameter 1: The following tables contain values for Parameter 1 across SNLI, MNLI, SQUAD

Sample Set	Terms								DQI C5 (ISIM=0.5)
	T1			T2	T3	T4	T5	T6	
	ISIM=0.5	ISIM=0.6	ISIM=0.7						
Original	3.79338794	5.79942751	9.64213607	0.13869626	0.06846071	0.00106449	19.2658	0.08669236	4.00160940
+S1	3.77492292	5.75927311	9.55986754	0.13950276	0.06756993	0.00105670	19.1081	0.08686184	3.98305231
+S2	3.77320467	5.75527455	9.54885537	0.13988920	0.06771915	0.00105824	19.1048	0.08711365	3.98187126
+S3	3.77796738	5.76636257	9.57941700	0.13950276	0.06756993	0.00105429	19.0986	0.08666733	3.98609436
+S4	3.80946946	5.84007436	9.69296631	0.13797814	0.06754694	0.00105432	19.2038	0.08661618	4.01604886
+S5	3.80273001	5.82425011	9.73687404	0.13854595	0.06744772	0.00105055	19.1196	0.08696758	4.00977423
+S6	3.80524680	5.83015604	9.72041244	0.13704206	0.06799806	0.00105172	19.1444	0.08642433	4.01133864
+S7	3.79613706	5.80879868	9.69710399	0.14008322	0.06781511	0.00104881	19.1444	0.08708462	4.00508420
+S8	3.79286615	5.80114342	9.67578885	0.13873626	0.06744340	0.00104868	19.1246	0.08673365	4.00009449
+S9	3.78510214	5.78300049	9.62542175	0.13969571	0.06763740	0.00105033	19.7681	0.08710369	3.99348558
+S10	3.79176275	5.79856261	9.66861134	0.13873626	0.06744340	0.00104875	19.1295	0.08675259	3.99899116
+S11	3.79366621	5.80301526	9.68099727	0.13931034	0.06751676	0.00104867	19.1840	0.08695819	4.00154198
+S12	3.78458008	5.78178193	9.62204642	0.13931034	0.06751676	0.00105054	19.1213	0.08674638	3.99245772

Table 40: DQI_{C5} for Case (ii)

2.0, and Story CLOZE.

Term	T1	T2	T3	DQI C1
Good	**1.8996**	**6.0409**	**0.9532**	**7.6578**
Bad	0.6416	5.8135	0.9494	6.1609

Table 47: SNLI Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	**1.8996**	**6.0409**	**0.9532**	**7.6578**
Bad	0.6416	5.8135	0.9494	6.1609

Table 48: SNLI Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	1.6177	**104.6542**	**0.7550**	**80.6316**
Bad	**7.4100**	14.1068	0.6020	15.9023

Table 49: MNLI Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	1.7715	71.3947	-0.0023	1.6073
Bad	**11.1550**	**73.3092**	**-0.001**	**11.1476**

Table 50: SQUAD 2.0 Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	3.3010	13.4569	0.2772	7.0313
Bad	**4.7675**	**13.4895**	**0.2839**	**8.5972**

Table 51: Story-CLOZE Sub-Component and Overall Values for DQI_{c1}

.

Granularity	Split	T1	T2	Contribution
Words	**Good**	**121.9512**	**0.7269**	**88.6463**
	Bad	52.3560	0.6500	34.0314
Adjectives	**Good**	**31.7460**	**0.2966**	**9.4159**
	Bad	16.9205	0.3590	6.0745
Adverbs	**Good**	**21.0970**	**0.1847**	**3.8966**
	Bad	10.7875	0.1732	1.8684
Verbs	**Good**	**43.6681**	**0.2349**	**10.2576**
	Bad	16.5289	0.1893	3.1289
Nouns	**Good**	**49.2611**	**0.4351**	**21.4335**
	Bad	21.0084	0.3685	7.7416
Bigrams	**Good**	**1296.3443**	**0.9374**	**1215.1931**
	Bad	873.2862	0.9355	816.9592
Trigrams	**Good**	**7686.3951**	**0.9546**	**7337.4328**
	Bad	6119.9510	0.9422	5766.2178
Sentences	**Good**	9070.7819	**0.6607**	**5993.0656**
	Bad	**14537.0541**	0.2705	3932.2731
Sentences (Not Normalized)	**Good**	**3.0656**	**0.6607**	**3.7263**
	Bad	1.2655	0.2705	1.0607
DQIC2	**Good**	-	-	**8668.3012**
	Bad	-	-	6636.3641

Table 52: SNLI Sub-Component and Overall Values for DQI_{c2}, Good Split

Granularity	Split	T1	T2	Contribution
Words	**Good**	299.2489	0.9223	275.9972
	Bad	**1026.2828**	**1.0000**	**1026.2828**
Adjectives	**Good**	147.7382	**1.0000**	147.7382
	Bad	**333.8001**	**1.0000**	**333.8001**
Adverbs	**Good**	14.9467	0.5166	7.7214
	Bad	**54.2488**	**0.7318**	**39.6992**
Verbs	**Good**	76.0906	0.6893	52.4492
	Bad	**182.7695**	**0.7130**	**130.3146**
Nouns	**Good**	225.1162	**0.9726**	218.9480
	Bad	**477.5051**	0.9704	**463.3709**
Bigrams	**Good**	4394.8945	**1.0000**	4394.8945
	Bad	**5615.4581**	**1.0000**	**5615.4581**
Trigrams	**Good**	16628.8816	0.9907	16474.2330
	Bad	**35285.2261**	**0.9735**	**34350.1676**
Sentences	**Good**	**15197.5684**	0.0049	74.4680
	Bad	11085.6756	**0.9680**	**10730.9339**
Sentences (Not Normalized)	**Good**	1.2314	0.0049	0.0060
	Bad	**11.1732**	**0.9680**	**10.8156**
DQIC2	**Good**	-	-	21646.4558
	Bad	-	-	**52700.84312**

Table 53: MNLI Sub-Component and Overall Values for DQI_{c2}, Good Split

Granularity	Split	T1	T2	Contribution
Words	**Good**	138.6878	**0.6744**	93.5310
	Bad	**615.0626**	0.6224	**382.8149**
Adjectives	**Good**	37.0775	**1.0000**	37.0775
	Bad	**161.0191**	**1.0000**	**161.0191**
Adverbs	**Good**	4.0080	0.7473	2.9951
	Bad	**18.7378**	**0.7610**	**14.2594**
Verbs	**Good**	30.1469	0.9051	27.2859
	Bad	**152.9500**	**0.9372**	**143.3447**
Nouns	**Good**	58.5576	**1.0000**	58.5576
	Bad	**255.8677**	1.0000	**255.8677**
Bigrams	**Good**	1665.8142	**0.9763**	1626.3344
	Bad	**4563.8191**	0.9755	**4452.0055**
Trigrams	**Good**	20526.6346	**1.0000**	20526.6346
	Bad	**39155.8925**	0.9821	**38455.0020**
Sentences	**Good**	**4811.1347**	-0.0013	-6.2544
	Bad	1996.9248	**0.2460**	**491.2435**
Sentences	**Good**	0.3991	-0.0013	-0.0005
(Not Normalized)	**Bad**	**1.3043**	**0.2460**	**0.3208**
DQIC2	**Good**	-	-	22366.1613
	Bad	-	-	**44355.87788**

Table 54: SQUAD 2.0 Sub-Component and Overall Values for DQI_{c2}, Good Split

Granularity	Split	T1	T2	Contribution
Words	**Good**	**396.9190**	**0.3661**	**145.3120**
	Bad	52.3560	0.3239	16.9581
Adjectives	**Good**	**77.3987**	**0.8307**	**64.2951**
	Bad	70.2610	0.8020	56.3493
Adverbs	**Good**	17.3230	0.4292	7.4350
	Bad	**27.8482**	**0.6178**	**17.2046**
Verbs	**Good**	59.4638	**0.5936**	**35.2977**
	Bad	**63.3871**	0.5511	34.9326
Nouns	**Good**	**270.8688**	0.8953	**242.5088**
	Bad	250.9358	**0.9289**	233.0942
Bigrams	**Good**	**4116.6448**	**1.0000**	**4116.6448**
	Bad	2991.6306	**1.0000**	2991.6306
Trigrams	**Good**	**30424.4890**	**1.0000**	**30424.4890**
	Bad	17757.2356	0.9383	16661.6141
Sentences	**Good**	**8161.7926**	-0.0015	-12.2426
	Bad	2544.5235	**0.0000**	**0.0000**
Sentences	**Good**	2.1199	-0.0015	-0.0031
(Not Normalized)	**Bad**	**2.1204**	**0.0000**	**0.0000**
DQIC2	**Good**	-	-	**35023.73666**
	Bad	-	-	20011.78371

Table 55: Story CLOZE Sub-Component and Overall Values for DQI_{c2}, Good Split

Parameter 2: Tables 52-55 contain values for Parameter 2 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Parameter 3: The following tables contain values for Parameter 3 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	9.1320	11.3955	14.3267
Bad	**10.3842**	**13.1062**	**16.6390**

Table 56: SNLI Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	**0.0468**	**0.0244**	**0.0103**
Bad	0.0404	0.0216	0.0094

Table 57: SNLI Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	9.4123	11.4508	14.3370
Bad	**10.3936**	**13.1156**	**16.7024**

Table 58: SNLI DQI_{C3}

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	**334.2154**	**695.0772**	**1040.5142**
Bad	312.4684	643.3308	953.5445

Table 59: MNLI Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	**0.0148**	**0.0108**	**0.0067**
Bad	**0.0111**	**0.0084**	**0.0056**

Table 60: MNLI Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	**334.2221**	**695.0839**	**1040.5209**
Bad	312.474	643.3364	953.5501

Table 61: MNLI DQI_{C3}

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	**129.8631**	**171.7117**	**228.9109**
Bad	88.9812	110.6097	141.2737

Table 62: SQUAD 2.0 Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	0.0051	0.0039	0.0026
Bad	**0.0055**	**0.0042**	**0.0094**

Table 63: SQUAD 2.0 Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	**129.8657**	**171.7143**	**228.9135**
Bad	88.984	110.6125	141.2765

Table 64: SQUAD 2.0 DQI_{C3}

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	**285.1348**	**513.1720**	**820.2516**
Bad	209.0823	368.5646	594.0969

Table 65: Story CLOZE Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	**0.0069**	**0.0053**	**0.0036**
Bad	**0.0069**	**0.0053**	**0.0036**

Table 66: Story CLOZE Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	**285.1384**	**513.1756**	**820.2552**
Bad	209.0859	368.5682	594.1005

Table 67: Story CLOZE DQI_{C3}

Parameter 4: The following tables contain values for Parameter 4 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split	DQIC4
Good	**0.0004**
Bad	0.0001

Table 68: SNLI DQI_{c4}

Split	DQIC4
Good	**0.0197**
Bad	0.0011

Table 69: MNLI DQI_{c4}

Split	DQIC4
Good	**5.2208**
Bad	0.4577

Table 70: SQUAD 2.0 DQI_{c4}

Split	DQIC4
Good	**0.0025**
Bad	0.0008

Table 71: Story CLOZE DQI_{c4}

Parameter 5: The following tables contain values for Parameter 5 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split	ISIM=0.3	ISIM=0.4	ISIM=0.5	ISIM=0.6
Good	**2.2349**	**2.8763**	**4.0125**	**6.3065**
Bad	2.2215	2.8558	3.9784	6.2237

Table 72: SNLI Term 1 for DQI_{c5}

Split	T2	T3	T4	T5	T6
Good	**0.1439**	**0.0038**	**6.4064e-05**	**20.3518**	**0.0903**
Bad	0.1430	0.0007	1.2711e-05	19.9288	0.0900

Table 73: SNLI Terms 2,3,4,5,6 for DQI_{c5}

Split	DQI C5
Good	**24.6024**
Bad	24.1409

Table 74: SNLI DQI_{c5}, with ISIM=0.5

Split	ISIM=0.3	ISIM=0.4	ISIM=0.5	ISIM=0.6
Good	**2.2233**	**2.8585**	**3.9884**	**6.3364**
Bad	2.1256	2.6986	3.6843	5.5845

Table 75: MNLI Term 1 for DQI_{c5}

Split	T2	T3	T4	T5	T6
Good	**0.0791**	0.0162	1.1073E-05	**15.3835**	14.7547
Bad	0.0741	**0.0307**	**20.9407E-05**	12.3932	**17.6181**

Table 76: MNLI Terms 2,3,4,5,6 for DQI_{c5}

Split	DQI C5
Good	**34.2219**
Bad	33.8006

Table 77: MNLI DQI_{c5}, with ISIM=0.5

Split	ISIM=0.3	ISIM=0.4	ISIM=0.5	ISIM=0.6
Good	2.5073	3.3460	5.0031	9.1300
Bad	**2.5379**	**3.4012**	**5.1352**	**9.6189**

Table 78: SQUAD 2.0 Term 1 for DQI_{c5}

Split	T2	T3	T4	T5	T6
Good	**0.0085**	0.0052	7.3081E-06	22.9314	**102.9990**
Bad	0.0079	**0.0524**	**7.4403E-05**	**27.0966**	88.8872

Table 79: SQUAD 2.0 Terms 2,3,4,5,6 for DQI_{c5}

Split	DQI C5
Good	**130.9472**
Bad	121.1793

Table 80: SQUAD 2.0 DQI_{c5}, with ISIM=0.5

Split	ISIM=0.3	ISIM=0.4	ISIM=0.5	ISIM=0.6
Good	**3.1103**	**4.5013**	**7.7337**	14.4898
Bad	3.0639	4.4163	7.5943	**14.7772**

Table 81: Story CLOZE Term 1 for DQI_{c5}

Split	T2	T3	T4	T5	T6
Good	**0.0400**	0.0027	3.1939E-05	**0.0400**	2.6196e-06
Bad	0.0398	**0.0084**	**9.7664E-05**	0.0398	**7.6306e-06**

Table 82: Story CLOZE Terms 2,3,4,5,6 for DQI_{c5}

Split	DQI C5
Good	**7.8164**
Bad	7.6824

Table 83: Story CLOZE DQI_{c5}, with ISIM=0.5

Parameter 6: The following tables contain values for Parameter 6 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split/Label	Entailment	Neutral	Contradiction
Good	1110	1430	708
Bad	5626	5008	6118

Table 84: SNLI Sample counts for Splits across Labels

Split-Label	T1	T2
Good-Entailment	8829.2425	**0.9387**
Bad-Entailment	**21655.2868**	0.8571
Good-Neutral	7467.5349	0.8699
Bad-Neutral	**31616.2545**	**0.9141**
Good-Contradiction	4932.7421	**0.9210**
Bad-Contradiction	**29145.0957**	0.8783

Table 85: SNLI Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-Entailment	**142.8571**	**0.7277**
Bad-Entailment	81.9672	0.6110
Good-Neutral	**153.8462**	**0.9118**
Bad-Neutral	117.6471	0.7071
Good-Contradiction	**163.9344**	**0.6764**
Bad-Contradiction	101.0101	0.6088

Table 86: SNLI Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-Entailment	**42.1230**	**0.34114**
Bad-Entailment	26.4201	0.30551
Good-Neutral	**48.8998**	0.46865
Bad-Neutral	38.1534	**0.47497**
Good-Contradiction	**43.1593**	0.31019
Bad-Contradiction	29.2826	**0.32385**

Table 87: SNLI Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-Entailment	**18.4128**	0.056911
Bad-Entailment	11.0963	**0.05816**
Good-Neutral	8.6798	0.09709
Bad-Neutral	**14.6135**	**0.43124**
Good-Contradiction	**37.9795**	**0.34286**
Bad-Contradiction	23.7192	0.21583

Table 88: SNLI Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-Entailment	**41.7885**	**0.16091**
Bad-Entailment	22.9410	0.05348
Good-Neutral	**48.9476**	0.17946
Bad-Neutral	38.9105	**0.20192**
Good-Contradiction	**53.5045**	**0.20000**
Bad-Contradiction	34.6380	0.13589

Table 89: SNLI Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-Entailment	**59.2768**	**0.49650**
Bad-Entailment	34.3643	0.38238
Good-Neutral	**62.7353**	**0.44534**
Bad-Neutral	46.4253	0.40586
Good-Contradiction	**66.3570**	**0.45653**
Bad-Contradiction	39.9202	0.37431

Table 90: SNLI Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-Entailment	1131.7133	**0.93307**
Bad-Entailment	**1173.5409**	0.93206
Good-Neutral	1261.2663	0.93783
Bad-Neutral	**1598.1514**	**0.94117**
Good-Contradiction	1100.8597	**0.94325**
Bad-Contradiction	**1369.0528**	0.93387

Table 91: SNLI Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-Entailment	5921.2942	**0.94672**
Bad-Entailment	**7757.5306**	0.93496
Good-Neutral	6414.8208	0.94517
Bad-Neutral	**10229.7186**	**0.95015**
Good-Contradiction	5478.1014	**0.95359**
Bad-Contradiction	**8984.3224**	0.94430

Table 92: SNLI Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Repetition	1	2	3	4	5	6
Good-Entailment	0.9844	0.0155	0	0	0	0
Bad-Entailment	0.9659	0.0308	0.001849	0	0.0007	0.0005
Good-Neutral	0.9667	0.0325	0.0007	0	0	0
Bad-Neutral	0.9785	0.0204	0.0010	0	0	0
Good-Contradiction	0.9798	0.0201	0	0	0	0
Bad-Contradiction	0.9785	0.0204	0.0010	0	0	0

Table 93: SNLI Sentence Granularity Repetitions

Split-Label	T3
Good-Entailment	**0.1457**
Bad-Entailment	0.1330
Good-Neutral	0.1496
Bad-Neutral	**0.1571**
Good-Contradiction	0.1313
Bad-Contradiction	**0.1434**

Table 94: SNLI T3 for DQI_{c6}

Split-Label	T4
Good-Entailment	**0.0100**
Bad-Entailment	0.0021
Good-Neutral	**0.0084**
Bad-Neutral	0.0022
Good-Contradiction	0.0197
Bad-Contradiction	**0.0020**

Table 95: SNLI T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	**15.3475**	11.6614
Words	**0.9313**	0.6596
Adjectives	**1.2190**	0.9185
Adverbs	**1.5708**	1.1850
Verbs	**0.9667**	0.7001
Nouns	**1.0623**	0.7358
Bigrams	0.3646	**0.4893**
Trigrams	0.1860	**0.2760**

Table 96: SNLI T5 for DQI_{c6}

Split-Label	DQI C6
Good	**556.6914**
Bad	320.2893

Table 97: SNLI DQI_{c6}

Split/Label	Entailment	Neutral	Contradiction
Good	6150	6098	6082
Bad	700	60	240

Table 98: MNLI Sample counts for Splits across Labels

Split-Label	T1	T2
Good-Entailment	2.69E+04	0.8133
Bad-Entailment	6.47E+03	0.9542
Good-Neutral	2.78E+04	0.8465
Bad-Neutral	4.76E+16	1.0000
Good-Contradiction	4.62E+04	0.9378
Bad-Contradiction	1.05E+17	1.0000

Table 99: MNLI Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-Entailment	5.67E+02	0.970607701
Bad-Entailment	9.48E+02	0.957116548
Good-Neutral	8.70E+02	0.488048002
Bad-Neutral	6.74E+02	0.794573643
Good-Contradiction	9.40E+02	0.965482191
Bad-Contradiction	7.01E+02	0.885763001

Table 100: MNLI Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-Entailment	1.16E+02	0.7834
Bad-Entailment	2.83E+02	1.0000
Good-Neutral	2.86E+02	1.0000
Bad-Neutral	1.92E+02	0.8771
Good-Contradiction	3.47E+02	1.0000
Bad-Contradiction	2.67E+02	1.0000

Table 101: MNLI Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-Entailment	2.56E+01	0.4803
Bad-Entailment	5.20E+01	0.6531
Good-Neutral	3.61E+01	0.6091
Bad-Neutral	7.15E+01	0.6521
Good-Contradiction	3.43E+01	0.5017
Bad-Contradiction	5.19E+01	0.3939

Table 102: MNLI Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-Entailment	1.71E+02	0.7901
Bad-Entailment	1.61E+02	0.6620
Good-Neutral	1.43E+02	0.5911
Bad-Neutral	1.69E+02	0.3061
Good-Contradiction	1.79E+02	0.7271
Bad-Contradiction	1.30E+02	0.5636

Table 103: MNLI Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-Entailment	2.61E+02	0.8994
Bad-Entailment	4.52E+02	0.9447
Good-Neutral	4.68E+02	1.0000
Bad-Neutral	2.61E+02	0.7235
Good-Contradiction	4.84E+02	1.0000
Bad-Contradiction	2.80E+02	0.9287

Table 104: MNLI Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-Entailment	3.38E+03	0.9361
Bad-Entailment	4.83E+03	1.0000
Good-Neutral	9.21E+03	1.0000
Bad-Neutral	1.91E+03	1.0000
Good-Contradiction	1.04E+04	1.0000
Bad-Contradiction	2.97E+03	1.0000

Table 105: MNLI Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-Entailment	9.27E+03	0.9573
Bad-Entailment	2.93E+04	1.0000
Good-Neutral	4.54E+04	0.9913
Bad-Neutral	4.61E+03	0.8822
Good-Contradiction	1.04E+05	1.0000
Bad-Contradiction	6.96E+03	0.9937

Table 106: MNLI Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Repetition	1	2	3
Good-Entailment	0.9512	0.0484	0.0003
Bad-Entailment	0.9884	0.0115	0.0000
Good-Neutral	0.9612	0.0363	0.0024
Bad-Neutral	1.0000	0.0000	0.0000
Good-Contradiction	0.9844	0.0150	0.0005
Bad-Contradiction	1.0000	0.0000	0.0000

Table 107: MNLI Sentence Granularity Repetitions

Split-Label	T3
Good-Entailment	0.0647
Bad-Entailment	0.0860
Good-Neutral	0.0926
Bad-Neutral	0.0590
Good-Contradiction	0.1000
Bad-Contradiction	0.2290

Table 108: MNLI T3 for DQI_{c6}

Split-Label	T4
Good-Entailment	0.0803
Bad-Entailment	0.0202
Good-Neutral	0.0041
Bad-Neutral	0.0484
Good-Contradiction	0.2018
Bad-Contradiction	0.0326

Table 109: MNLI T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	14.6049	72.1687
Words	1.2571	0.8533
Adjectives	1.4557	1.7959
Adverbs	0.7319	0.9429
Verbs	1.0382	1.0345
Nouns	1.7755	1.5836
Bigrams	0.4008	0.3561
Trigrams	0.6547	0.9724

Table 110: MNLI T5 for DQI_{c6}

Split-Label	DQI C6
Good	2.74E+05
Bad	1.53E+17

Table 111: MNLI DQI_{c6}

Split/Label	True	False
Good	10946	10770
Bad	914	1086

Table 112: SQUAD 2.0 Sample counts for Splits across Labels

Split-Label	T1	T2
Good-True	4431.2159	0.0007
Bad-True	1921.2260	0.5448
Good-False	4412.2037	0.0014
Bad-False	1853.6963	0.5009

Table 113: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-True	263.6776	1.0000
Bad-True	954.5225	1.0000
Good-False	259.3381	0.3105
Bad-False	776.2031	1.0000

Table 114: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-True	75.3820	1.0000
Bad-True	244.8719	1.0000
Good-False	70.8210	1.0000
Bad-False	222.5754	1.0000

Table 115: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-True	6.31677	0.6666
Bad-True	27.6740	0.6494
Good-False	6.4805	0.6632
Bad-False	24.6482	0.6878

Table 116: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-True	58.2850	0.8789
Bad-True	219.8726	0.8851
Good-False	59.0344	0.9066
Bad-False	208.3846	0.9113

Table 117: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-True	110.8118	1.0000
Bad-True	415.9473	1.0000
Good-False	109.7139	1.0000
Bad-False	307.1137	1.0000

Table 118: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-True	2923.9305	0.9768
Bad-True	5800.9793	0.9762
Good-False	2834.7978	0.9758
Bad-False	5157.4516	0.9749

Table 119: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-True	35363.3144	1.0000
Bad-True	49074.7258	1.0000
Good-False	34076.1381	1.0000
Bad-False	40854.1931	1.0000

Table 120: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Label	T3
Good-True	0.0085
Bad-True	0.00852
Good-False	0.0079
Bad-False	0.0078

Table 121: SQUAD 2.0 T3 for DQI_{c6}

Split-Label	T4
Good-True	0.0104
Bad-True	0.0106
Good-False	0.1165
Bad-False	0.0954

Table 122: SQUAD 2.0 T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	20.5287	9.6533
Words	0.0711	0.0682
Adjectives	0.6497	1.1487
Adverbs	0.4012	0.6832
Verbs	0.4918	0.8153
Nouns	0.5183	0.9957
Bigrams	0.1262	0.05600
Trigrams	0.1366	0.09422

Table 123: SQUAD 2.0 T5 for DQI_{c6}

Split-Label	DQI C6
Good	75918.2760
Bad	105949.3404

Table 124: SQUAD 2.0 DQI_{c6}

Split/Label	True	False
Good	2568	2568
Bad	800	800

Table 125: Story CLOZE Sample counts for Splits across Labels

Split-Label	T1	T2
Good-True	1.30E+05	0.9984
Bad-True	5.06E+16	1.0000
Good-False	1.30E+05	0.9984
Bad-False	5.06E+16	1.0000

Table 126: Story CLOZE Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-True	5.47E+05	0.9792
Bad-True	5.22E+05	0.8618
Good-False	5.47E+05	0.5316
Bad-False	4.96E+05	0.8537

Table 127: Story CLOZE Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-True	129.1883	0.7800
Bad-True	133.5904	0.7711
Good-False	121.0435	0.7459
Bad-False	128.3632	0.8014

Table 128: Story CLOZE Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-True	41.1600	0.5959
Bad-True	49.9482	0.5368
Good-False	36.9653	0.6145
Bad-False	54.7544	0.6194

Table 129: Story CLOZE Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-True	103.8261	0.5285
Bad-True	115.6968	0.5828
Good-False	112.3307	0.5946
Bad-False	113.4481	0.5155

Table 130: Story CLOZE Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-True	551.3272	0.8898
Bad-True	458.9138	0.8862
Good-False	520.3204	0.9047
Bad-False	462.2876	0.9252

Table 131: Story CLOZE Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-True7139.05776	1.0000	
Bad-True5158.2473	1.0000	
Good-False6941.1989	1.0000	
Bad-False5006.1656	1.0000	

Table 132: Story CLOZE Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-True54497.5504	1.0000	
Bad-True33876.9502	1.0000	
Good-False50906.0915	1.0000	
Bad-False33618.6103	1.0000	

Table 133: Story CLOZE Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Label	T3
Good-True	0.0085
Bad-True	0.0079
Good-False	0.0085
Bad-False	0.0078

Table 134: Story CLOZE 2.0 T3 for DQI_{c6}

Split-Label	T4
Good-True	0.0104
Bad-True	0.1165
Good-False	0.0106
Bad-False	0.0954

Table 135: Story CLOZE 2.0 T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	382.2842	2262.7417
Words	1.0447	1.0192
Adjectives	3.9910	5.0527
Adverbs	1.7714	3.1284
Verbs	2.2377	3.5188
Nouns	5.8841	7.3696
Bigrams	1.6522	1.9489
Trigrams	4.9660	6.8154

Table 136: Story CLOZE T5 for DQI_{c6}

Split-Label	DQI C6
Good	1.01E+17
Bad	1.01E+17

Table 137: Story CLOZE DQI_{c6}

Parameter 7: The following tables contain values for Parameter 7 across SNLI, MNLI, and SQUAD 2.0. Story CLOZE does not have a separate training set and is hence not evaluated.

Split	SSMIL=0.2	SSMIL=0.3	SSMIL=0.4
Good	**0.0031**	**0.0042**	**0.0063**
Bad	0.0029	0.0040	0.0057

Table 138: SNLI DQI_{c7}

Split	SSMIL=0.2	SSMIL=0.3	SSMIL=0.4
Good	0.0004	0.0005	0.0002
Bad	**0.0009**	**0.0011**	**0.0005**

Table 139: MNLI DQI_{c7}

Split	SSMIL=0.2	SSMIL=0.3	SSMIL=0.4
Good	**1.2500**	**1.4285**	**1.6666**
Bad	0.0029	0.0040	0.0057

Table 140: SQUAD 2.0 DQI_{c7}

A.7 Interface Design

Careful Selection of Visualizations Prior to the design of test cases and a user interface, data visualizations highlighting the effects of sample addition are built. Considering the complexity of the formulas for the components of empirical DQI, we carefully select visualizations to help illustrate and analyze the effect to which individual text properties are affected.

All DQI Component Values are Shown for Each Visualization: We show all DQI component values for each visualization, since the user needs to optimize across several dependent components while selecting the best quality data. All DQI component values are tracked across different visualizations using two separate panels present at the bottom of the screen. The first panel shows the component-wise values as colored circles for the overall dataset prior to adding the sample. The second panel is initially a set of grayscale circles. Once the new sample is added, both the panels are updated. The first panel may not show any color changes, as it represents the overall dataset. The second however, will now display colored circles based on the DQI component values of the individual new sample. The values of the components can be viewed with a tooltip.

Traffic Signal Color Scheme: The color combination of Red-Yellow-Green used in all the visualizations represents the quality of the component/property being observed/analyzed. Here, red represents an undesirable quality value, yellow a permissible value, and green an ideal value. The color scale follows a pattern of red-yellow-green-yellow-red unless otherwise specified, centered around the ideal value of a component.

A.7.1 Vocabulary

Which Characteristics of Data are Visualized? The contribution of samples to the size of the vocabulary is tracked using a dual axis bar chart. This displays the vocabulary size, along with the vocabulary magnitude, across the train, dev, and test splits for the dataset. Also, the distribution of sentence lengths is plotted as a histogram. Each sample

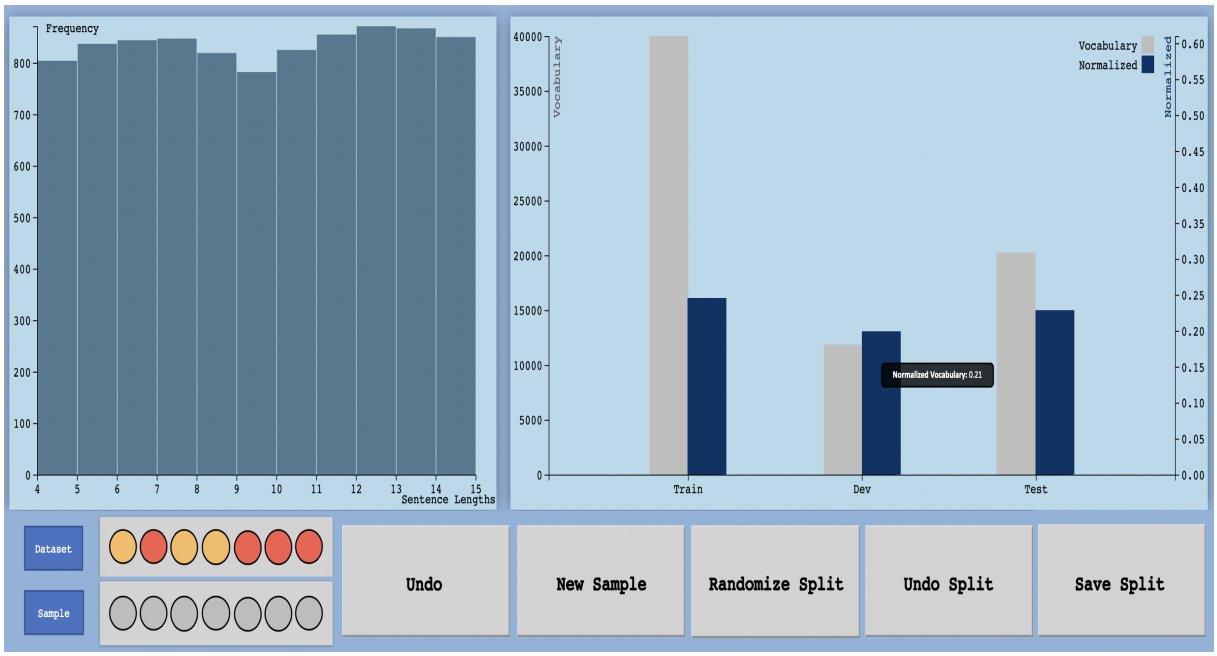


Figure 9: DQI_{c1} Visualization Prior to New Sample Addition

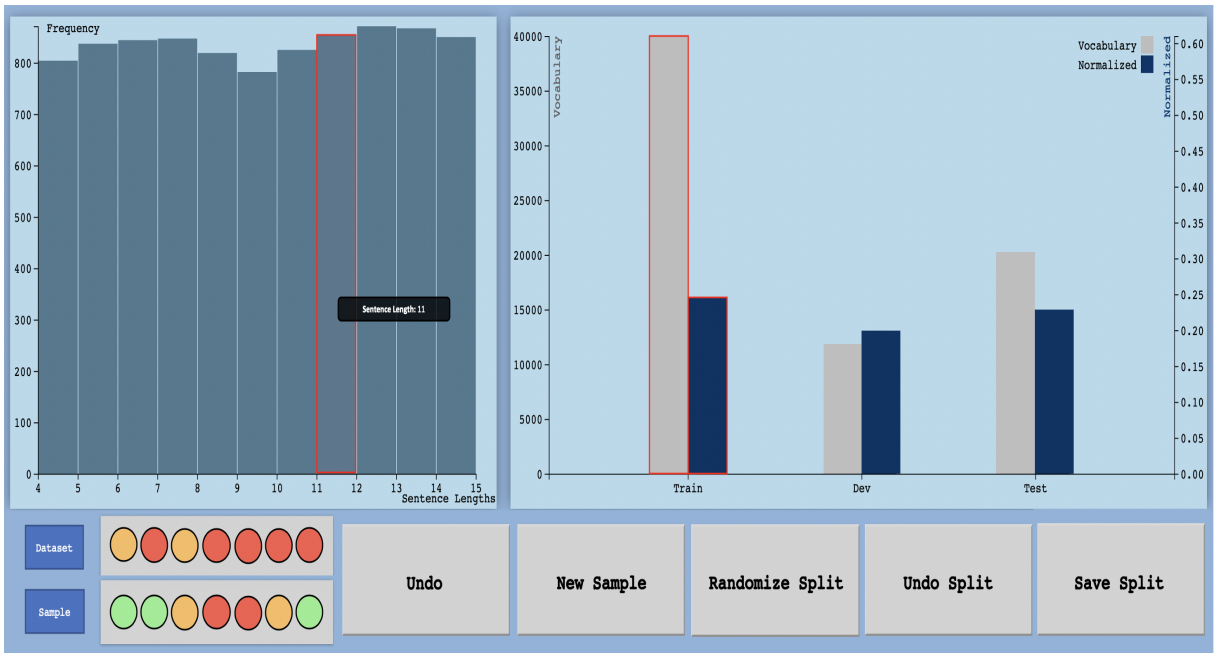


Figure 10: DQI_{c1} Visualization On New Sample Addition

contributes two sentences, i.e., the premise and hypothesis statements. Figure 9 illustrates this.

Interactions: Interactions are supported through a tooltip and buttons. The tooltip displays the quantities in both charts on mouseover, and the buttons are used to update the chart. There are five "interactions" supported:

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the train split by default. A script to calculate the new words this sample contributes to the vocabulary set is run, and the bar chart is accordingly updated. The sentence lengths of the premise and hypothesis statements are used to update the histogram. The updated portions of both the charts are highlighted, as shown in Figure 10. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

- **Removal of a New Sample *(Undo)*:** This

reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

- **Randomization of Split *(Randomize Split)*:** The samples are distributed randomly between the train, dev, and test splits, using a 70:10:20 split ratio. Once the split is randomized, the new sample cannot be removed from the split anymore, as it is not necessarily a part of the train set. In order to account for annotator bias, the annotator id of dataset samples is used to create mutually exclusive annotator sets across splits. Additionally, the split is designed such that if a premise has multiple hypothesis statements and is therefore repeated across samples, then all samples containing that premise belong to the same split. This split operation can be performed multiple times, as an attempt to understand the effect of data ordering on the DQI component values for the overall dataset. The previous state of the visualization is saved in a set of variables.

- **Undo Split *(Undo Split)*:** This reverses the operations of 'randomization of split' by using the saved state variables to restore the visualizations back to their original state. Only the latest randomization operation is reversed.

- **Save Split *(Save Split)*:** Once the split is satisfactory, this button can be used to freeze this split state for the remainder of the analysis. On addition of the next sample, this frozen state is used for the initialization of the visualizations.

A.7.2 Inter-sample N-gram Frequency and Relation

Which Characteristics of Data are Visualized? There are different granularities of samples that are used to calculate the values of this component, namely: words, POS tags, sentences, bigrams, and trigrams. The granularities' respective frequency distributions and standard deviations are utilized for this calculation.

Bubble Chart for visualizing the frequency distribution: A bubble chart is used to visualize the frequency distribution of the respective granularity. This design choice is made in order to clearly view the contribution made by a new sample when added to the existing dataset in terms of different granularities. The bubbles are colored according to the bounds set for frequencies by the hyperparameters, and sized based on the frequency of the elements they represent. Additionally, some insight into variance can be obtained from this chart, by observing the variation in bubble size.

Bullet Chart for impact of new sample: The impact of sample addition on standard deviation can be viewed using the bullet chart. The red-yellow-green color bands for each granularity represent the standard deviation bounds of that granularity. The vertical black line represents the ideal value of the standard deviation of that granularity. The two horizontal bars represent the value of standard deviation before and after the new sample's addition. Figure 11 illustrates the visualization.

Interactions: A tooltip, buttons, and a drop down are used for interactions. The tooltip displays the quantities in both charts on mouseover, and the buttons/drop down are used to update the chart. The following tasks are supported by the latter.

- **Changing Granularity *(Drop Down)*:** The drop down menu is used to select the granularity of the bubble chart displayed, as shown in Figure 11.

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and an updated bubble chart of the word frequency distribution is generated. The new words that are added/ existing words that are updated are highlighted with thick black outlines in the chart. The granularity of the view can be changed using the drop down. The additions/modifications in the frequency distribution are similarly highlighted across all granularities, as illustrated in Figure 12. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

A.7.3 Inter-sample STS

Which Characteristics of Data are Visualized? The main units used in this DQI component are

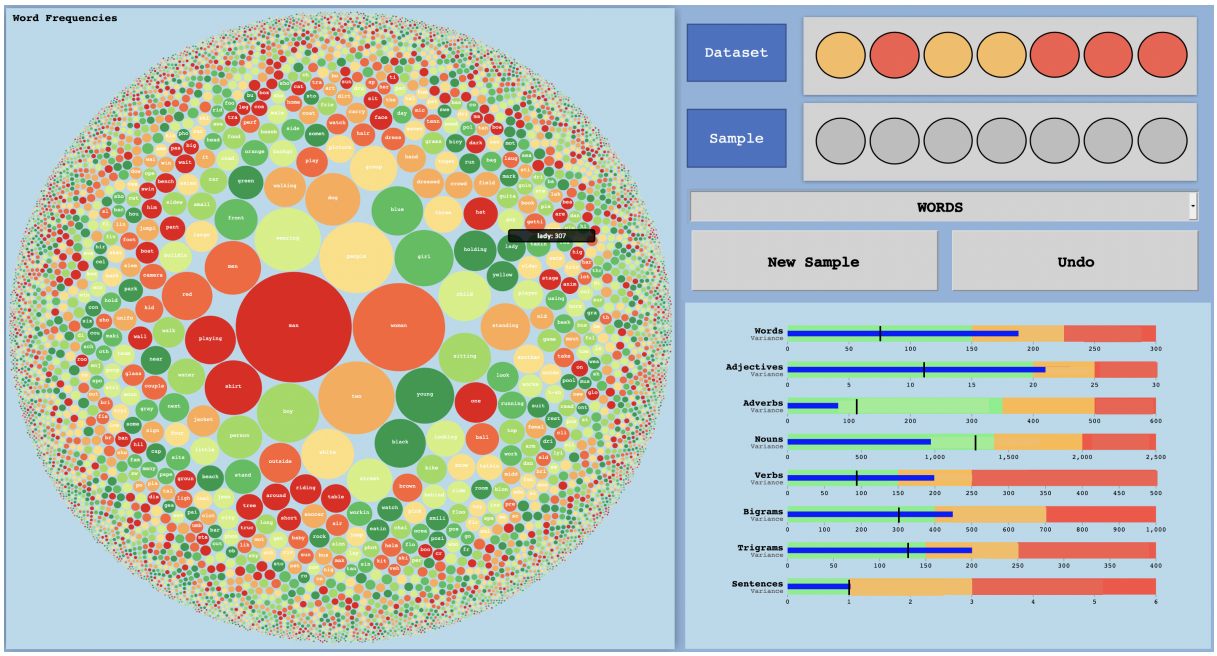


Figure 11: DQI_{c2} Visualization Prior to New Sample Addition

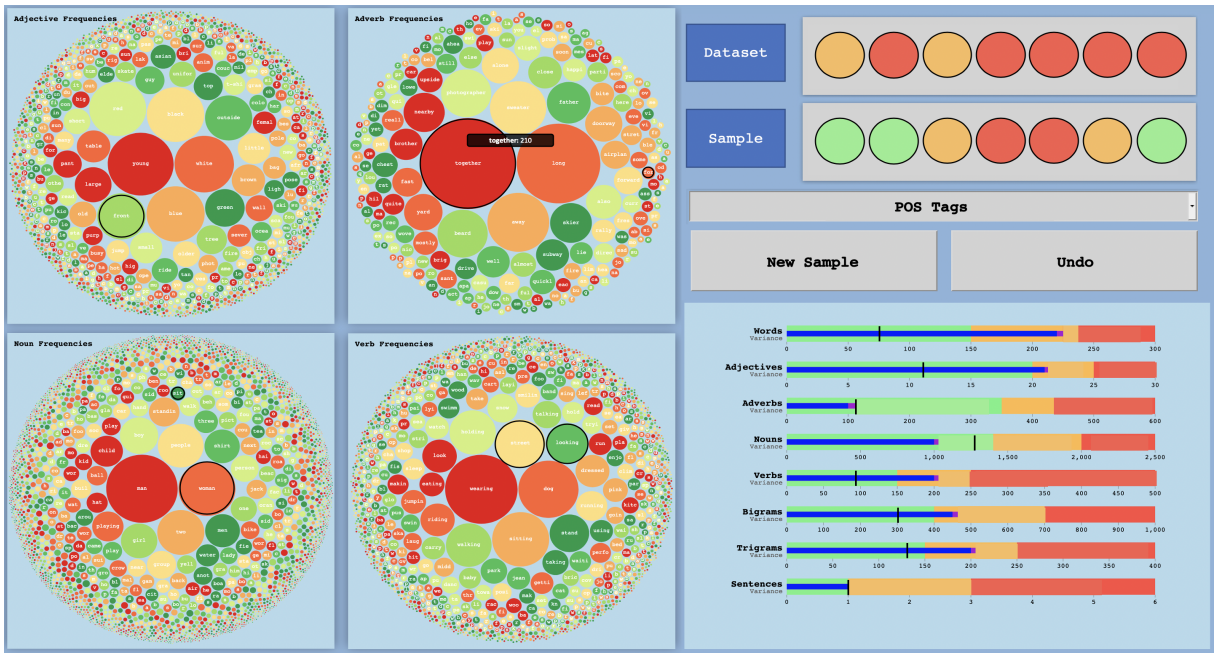


Figure 12: DQI_{c2} Visualization On New Sample Addition

the similarity values between sentences across the dataset. This refers to either premise or hypothesis statements, relative to all other premise/hypothesis statements. In order to understand the similarity relations of sentences, a force layout and horizontal bar chart are used. This is illustrated in Figure 13.

Force Layout for Similar Sentence Pairs In the force layout, those sentence pairs with a similarity value that meets the minimum threshold are connected. Each node represents a sentence. The thickness of the connecting line depends on how close the similarity value is to the threshold.

Horizontal Bar Chart for Most Similar Sentences In the horizontal bar chart, the sentences that are most similar to the given sentence are ordered in terms of their similarity value. The bar colors are centered around the threshold.

Interactions: Interactions via tooltip display the sentence id- i.e., the sample id, and whether the sentence is a premise/hypothesis of that sample-

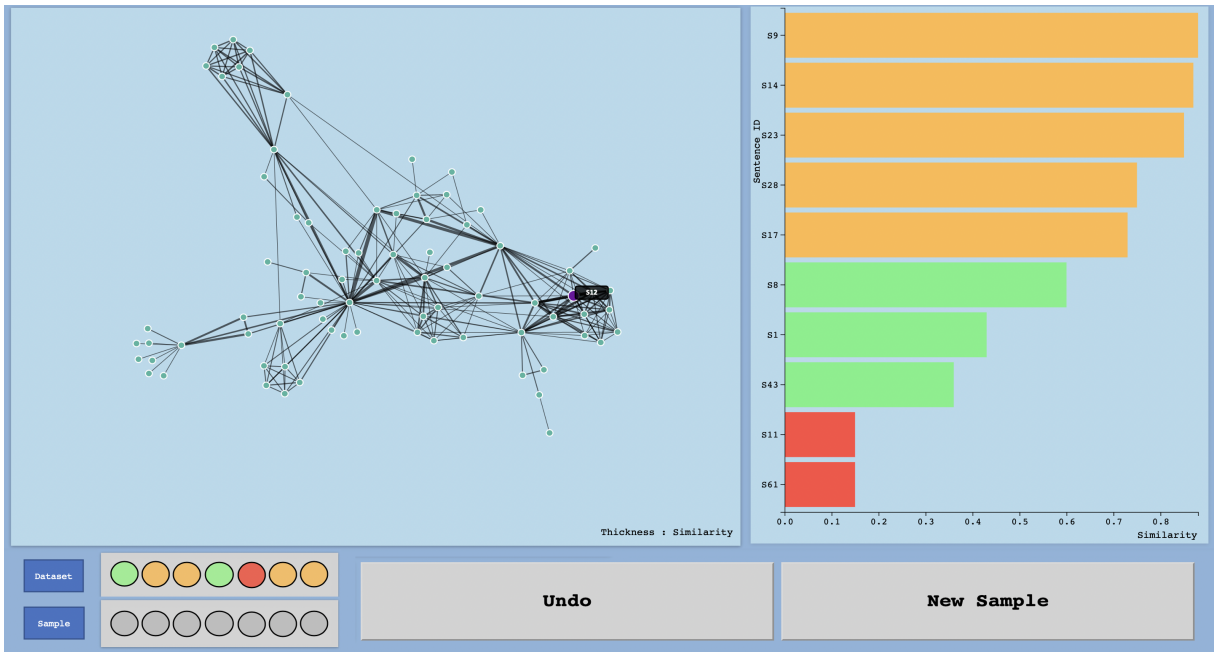


Figure 13: DQI_{c3} Visualization Prior to New Sample Addition

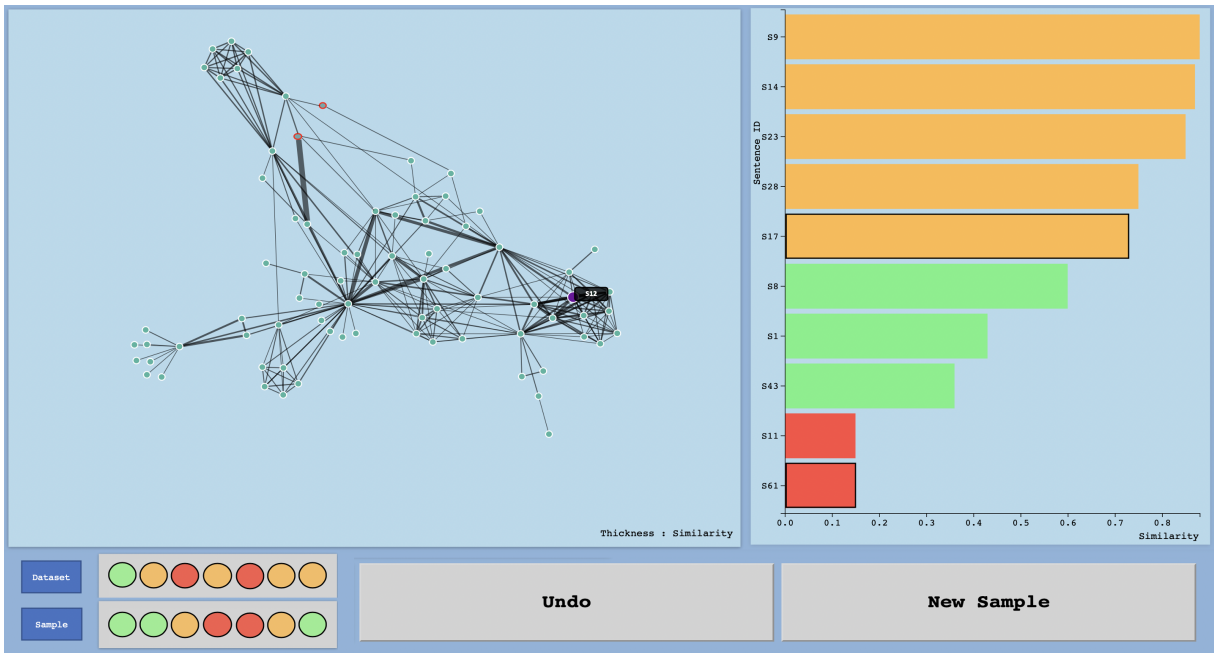


Figure 14: DQI_{c3} Visualization On New Sample Addition

and similarity value in case of both the charts. The two charts are also linked on click of a node in the force layout. Other interactions are fuelled by buttons. The complete set of tasks is as follows:

- **Displaying Horizontal Bar Chart *(on node click)*:** By selecting a node in the force layout, a horizontal bar chart is produced, that displays the ten most similar sentences to the sentence represented by the node. The benefits of the bar chart are two-fold. First, the bar chart accounts for sentence links not present in the force layout. It displays those sentences whose similarity value is below the minimum threshold. This can help if certain sentences are isolated without links in the force layout. Second, it enhances the readability of information present in the force layout by drilling down on a subset, if the dataset size is very large.

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and

two new nodes are created in the force layout. The outline of these two nodes is in black, and by default, the premise is auto-selected to generate the bar chart. If the new sample's sentences appear in the bar chart for any other sample, then the outline of those bars is in black, as illustrated in Figure 14. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

A.7.4 Intra-sample Word Similarity

Which Characteristics of Data are Visualized? In this section, A sample's word similarity is viewed in terms of premise-only, hypothesis-only, and both. The relationship between non-adjacent words in the sample's sentences is analyzed specifically.

Overview Chart for Average Word Similarities and Heatmap for Single Sample The overview chart that is used is a one-level tree map, which uses the average value of all word similarities per sample- i.e., concatenated premise and hypothesis- to color and group its components. This is illustrated in Figure 15 The detailed view is a heat map of all the words in a single sample, as shown in Figure 17.

Interactions: Tooltips display the sample id for the tree map, and the similarity value between words for the heat map. Other interactions include a drop down used to select the sentence to be viewed in the heat map, linking the heat map to the tree map on click, and buttons to modify the visualizations. The tasks are as follows:

- **Displaying Heat Map *(on Tree Map click)*:** By clicking on a box of the tree map, the user is shown the heat map of the clicked on sample.

- **Displaying the Tree Map *(on Heat Map click)*:** By clicking anywhere on the heat map, the user is taken back to the tree map view.

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and a new box is added to the tree map, with a black outline to highlight it, as illustrated in Figure 16. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

- **Change Heat Map View *(Drop Down)*:** Using the drop down, the heatmap can be changed to show word similarities for the (a) premise, (b) hypothesis, or (c) both sentences.

A.7.5 Intra-sample STS

Which Characteristics of Data are Visualized? Premise-Hypothesis similarity is analyzed on the basis of length variation, meeting a minimum threshold, and similarity distribution across the dataset. The first is addressed already in the vocabulary property by viewing the sentence length distribution. The other two are visualized using a histogram and kernel density estimation curve, as shown in Figure 18.

Histogram and Kernel Density Curve for Sample Distribution The histogram represents the distribution of the samples, and is colored by centering around the threshold as the ideal value. The number of bins can be changed, and therefore multi-level analysis can be conducted. The kernel density curve is used to check for the overall skew of the distribution.

Interactions: Tooltips on the histogram display the number of samples per bin. Buttons and a text box are used for implementing other interactions:

- **Re-binning Histogram *(textbox)*:** By filling a new value in the textbox, the number of bins in the histogram changes to that value.

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, the histogram and density plot are updated accordingly. The bar in the histogram to which the sample contributes is outlined in black across all histogram binnings, as illustrated in Figure 19. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

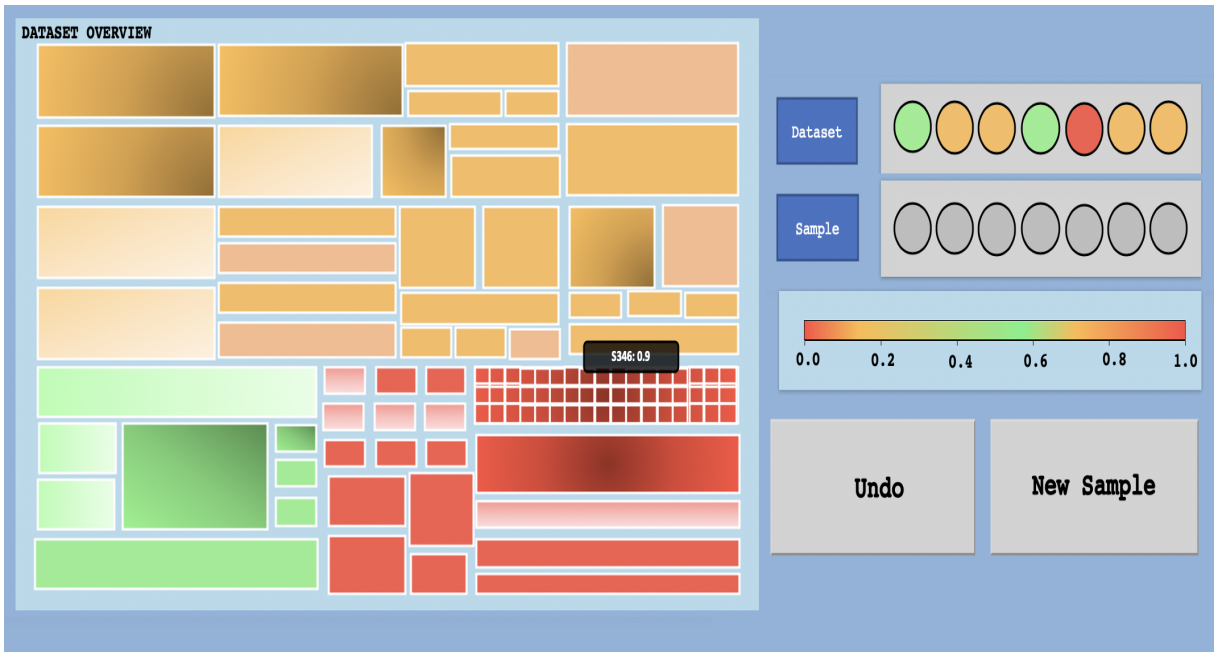


Figure 15: DQI_{c4} Visualization Prior to New Sample Addition

Figure 16: DQI_{c4} Visualization On New Sample Addition: Dataset View

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

A.7.6 N-Gram Frequency per Label

Which Characteristics of Data are Visualized? This component drills down on the second component, to view the patterns seen in granularities per label. There are two small multiples charts, divided based on label, used in this view- a violin plot and a box plot.

Violin plot and Kernel Density Curve for Skew of Distribution: The violin plots are structured to display both jittered points, according to their frequency distribution, as well as a kernel density curve to judge the skew of the distribution. The points each represent an element of the granularity.

Box Plots for More Information The box plots are used to garner more information about the distri-

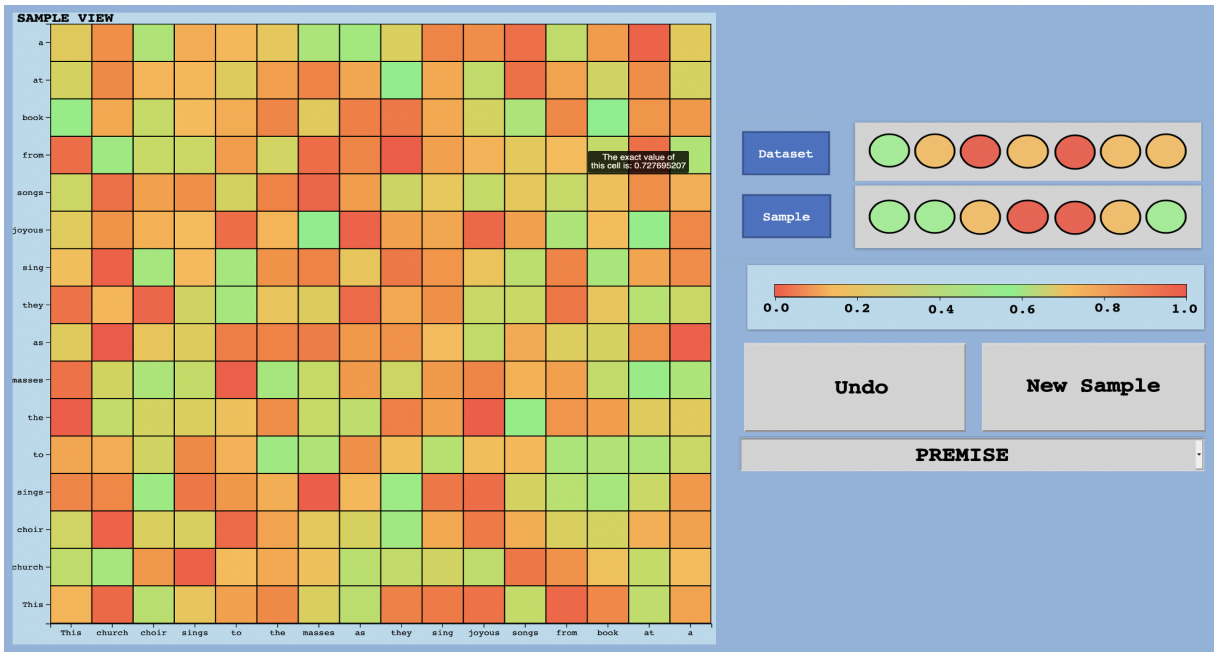


Figure 17: DQI_{c4} Visualization On New Sample Addition: Sample View

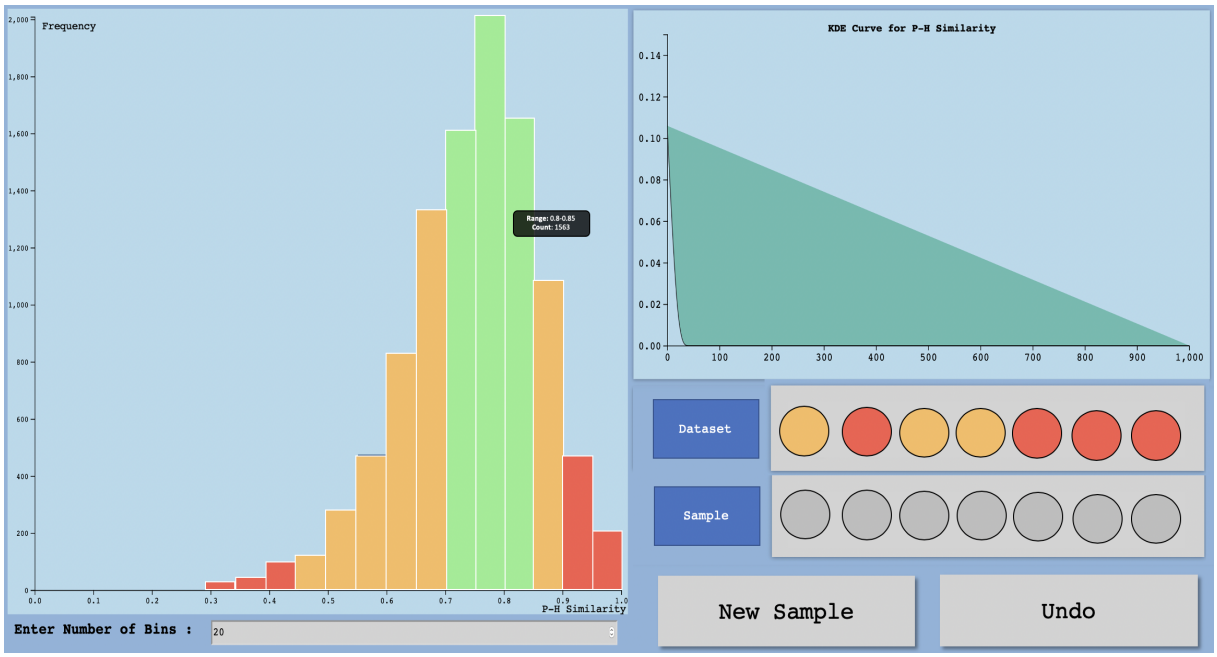


Figure 18: DQI_{c5} Visualization Prior to New Sample Addition

bution, in terms of its min, max, median, mean, and inter quartile range. These help further characterize the distribution, as well as provide a quantitative definition of the skew seen using density curves. Jittered points representing elements are present in this plot as well.

Interactions: On mouseover of a point in both visualizations, the element and its frequency are displayed in a tooltip. Other interactions are based on a dropdown and buttons as follows:

- **Changing Granularity *(Drop Down)*:** The drop down menu is used to select the granularity of the plots displayed, as shown in Figure 20. This granularity can be in terms of words, POS tags, bigrams, trigrams, or sentences.

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and updated plots of the word frequency distribution are generated. The new words that are added/ existing words that are updated

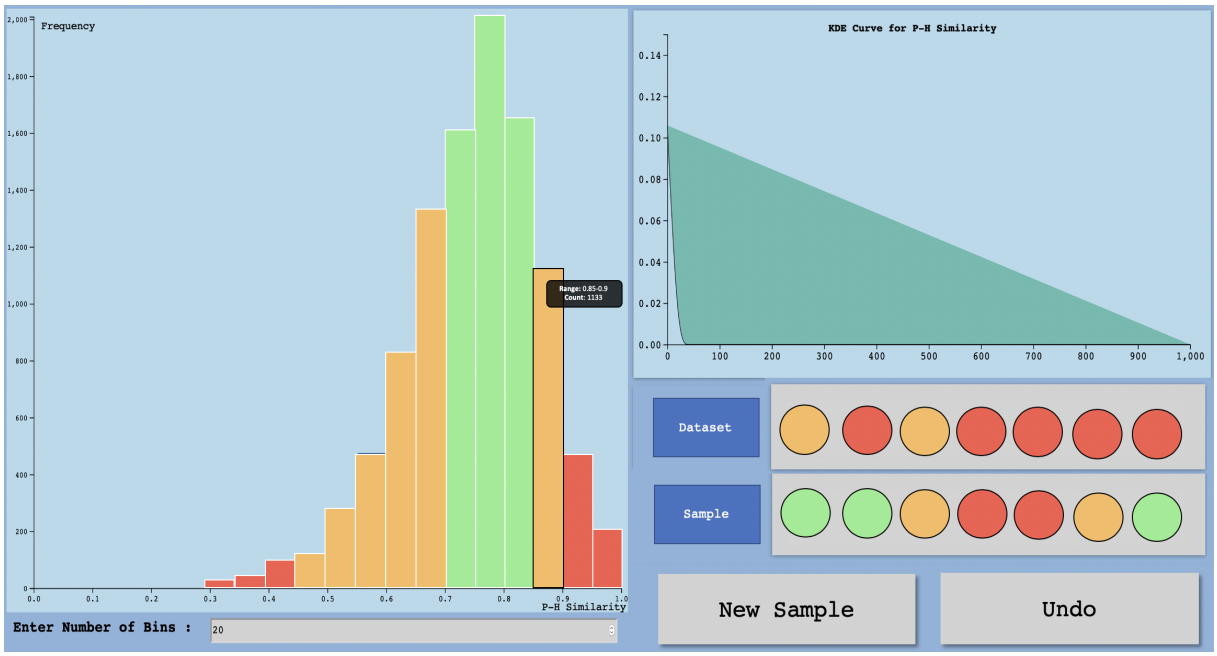


Figure 19: DQI_{c5} Visualization On New Sample Addition

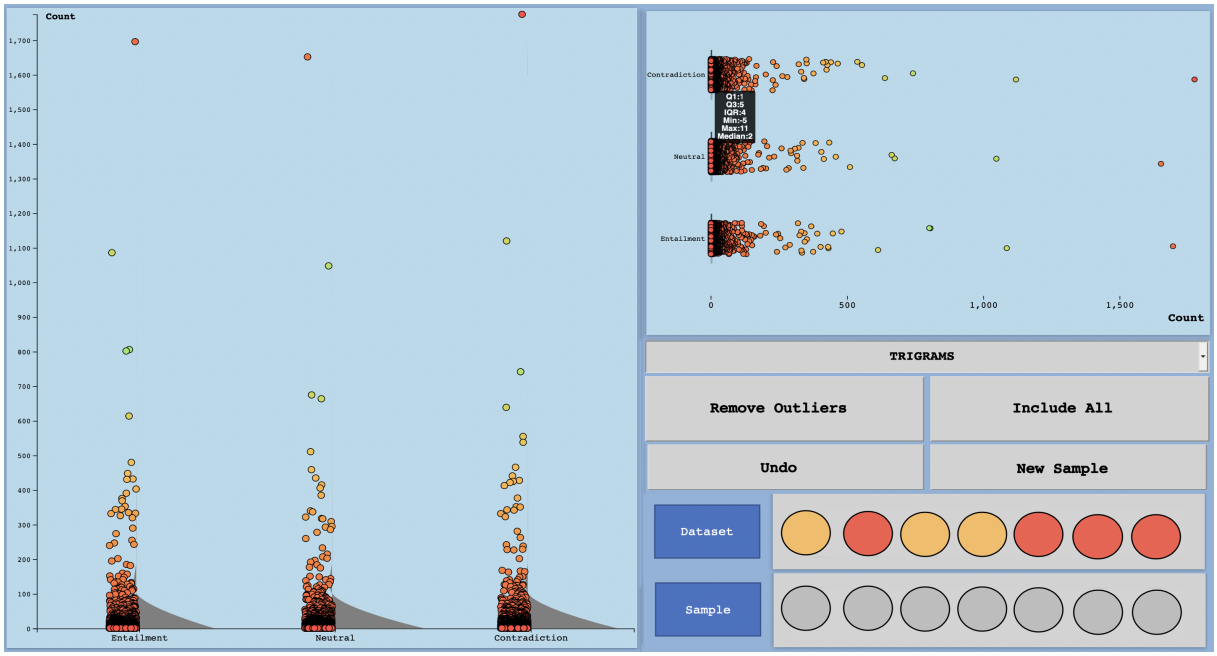


Figure 20: DQI_{c6} Visualization Prior to New Sample Addition

are highlighted with thick white outlines in the chart. The granularity of the view can be changed using the drop down. The additions/modifications in the frequency distribution are similarly highlighted across all granularities. This is shown in Figure 22 and 23 .The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

- **Outlier Handling *(Remove Outliers)*:** This removes elements with frequency counts less than the median to get a less skewed picture of the remainder of the distribution. The component value panels are updated as well, as illustrated in Figure 21. The previous state of

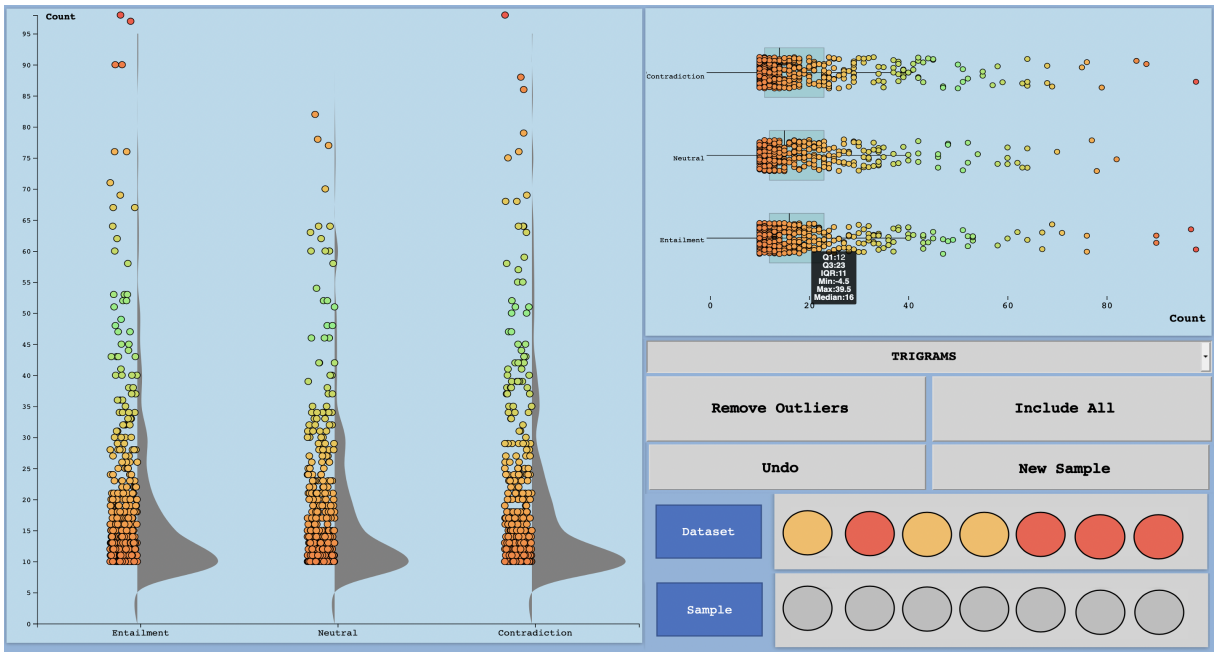


Figure 21: DQI_{c6} Visualization after removing outliers Prior to New Sample Addition

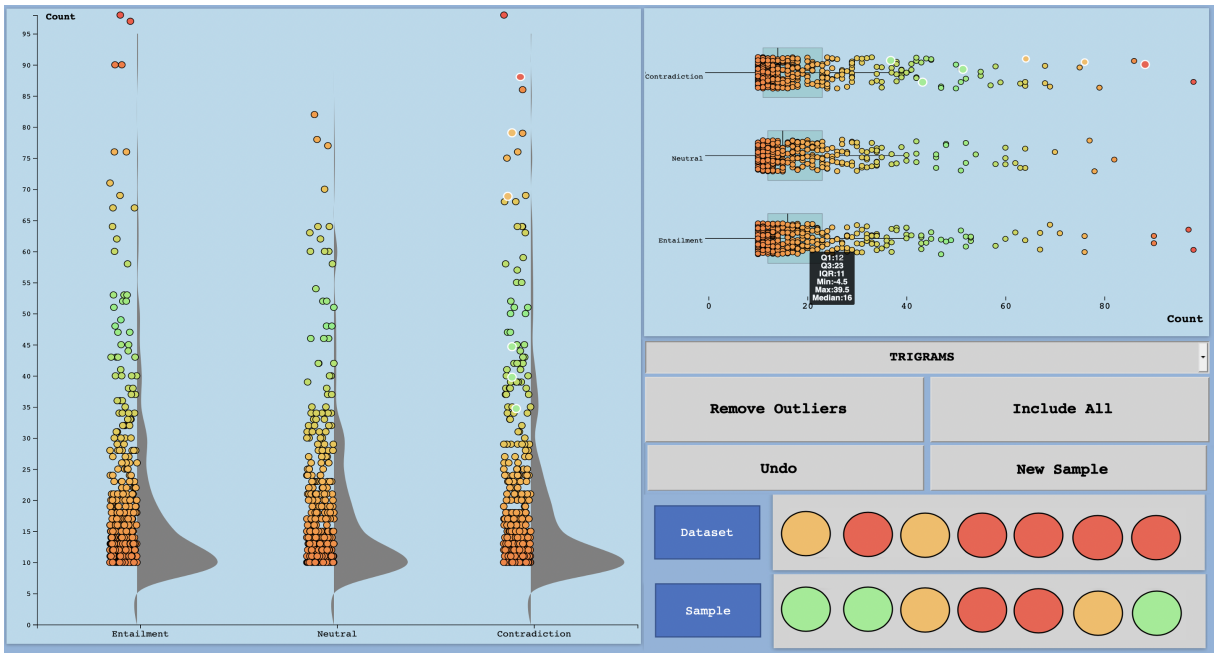


Figure 22: DQI_{c6} Visualization On New Sample Addition

the visualization is saved in a set of variables.

- **Full Distribution View *(Include All Samples)*:** This reverses the operations of 'outlier handling' by using the saved state variables to restore the visualizations back to their original state.

A.7.7 Inter-split STS

Which Characteristics of Data are Visualized? Train-Test similarity must be kept minimal to prevent data leakage. This component's main feature is finding the train split sample that is most similar to a given test split sample.

Parallel Coordinate Graph for Train-Test Similarity: A subset of test and train samples, all found to have close similarity within their respective splits, and significant similarity across the splits are plotted as a one step parallel coordinate graph, with test samples along one axis, and train samples along the other. This subset is seeded

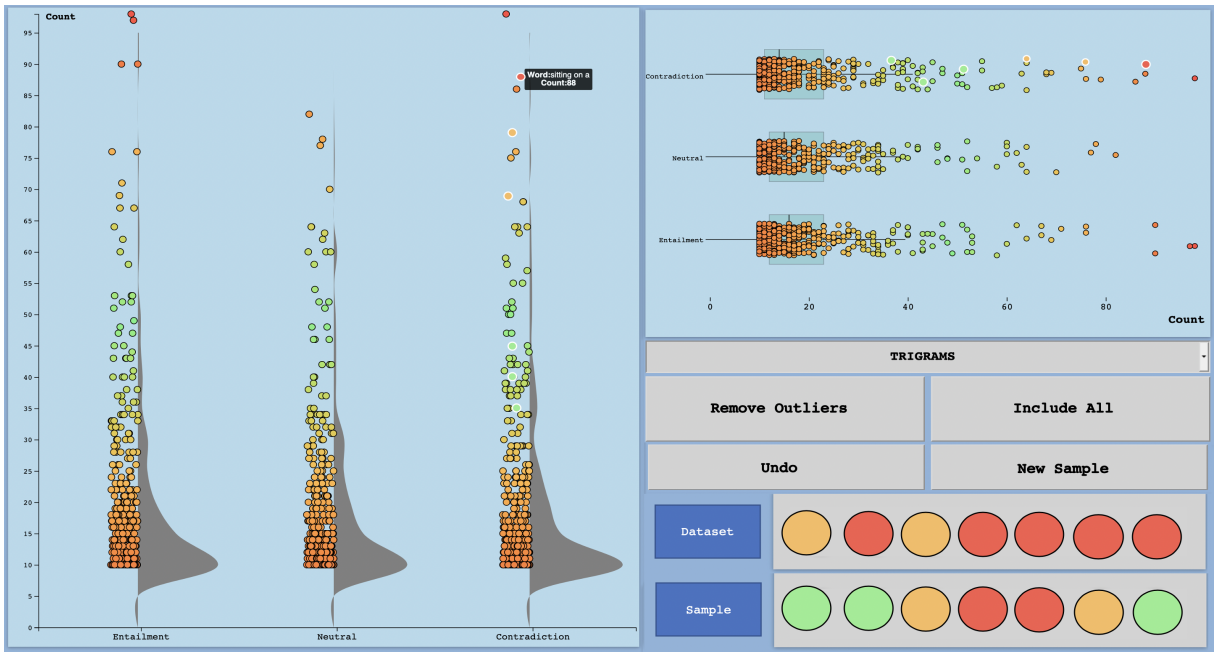


Figure 23: DQI_{c6} Visualization with mouseover On New Sample Addition

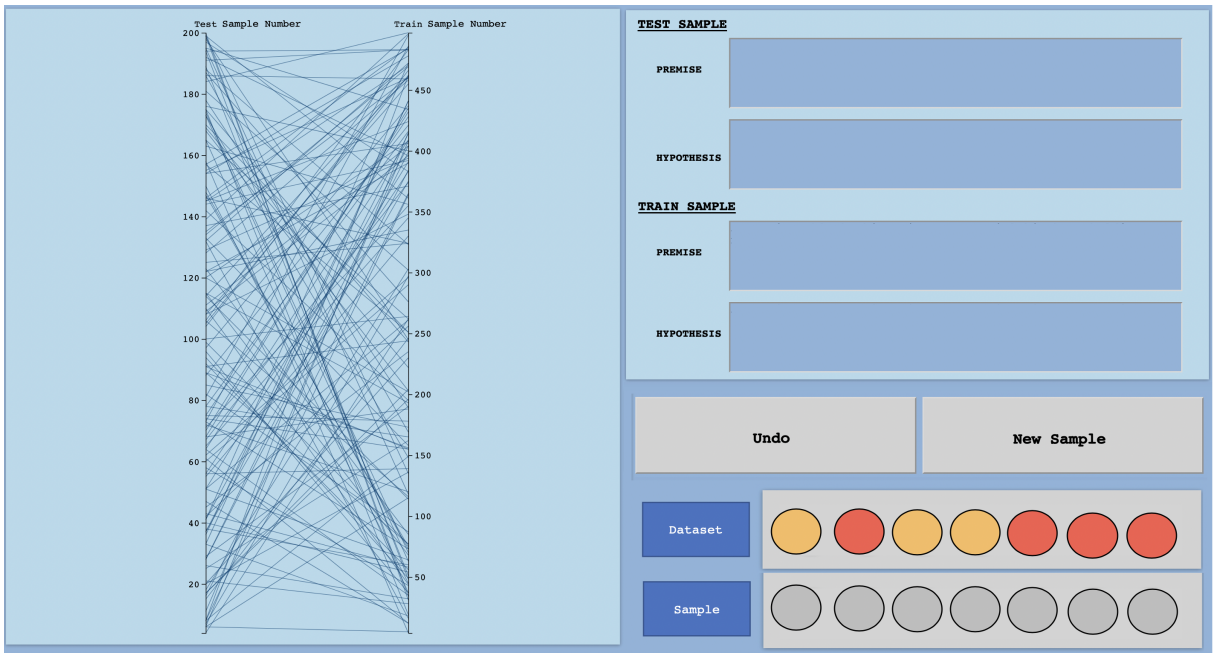


Figure 24: DQI_{c7} Visualization Prior to New Sample Addition

with those samples closest in similarity to the new sample to be introduced, based on the third component's visualization. The links connecting points on the two axes are drawn between the most similar matches across the split, as shown in Figure 24.

Interactions: Interactions include a tooltip that displays the sample ids connected on mouseover of a link, text boxes filled on click of a link, and other tasks by buttons:

- **Details of Linked Pair *(on click of link)*:** Clicking on a link causes the link to turn red, and the premises and hypotheses of the two samples are displayed in the text boxes on the screen. Clicking on another link changes the values of the textboxes, and highlights only the new link.

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and the sample is added to the axis of the parallel coordinates plot depending on the split that

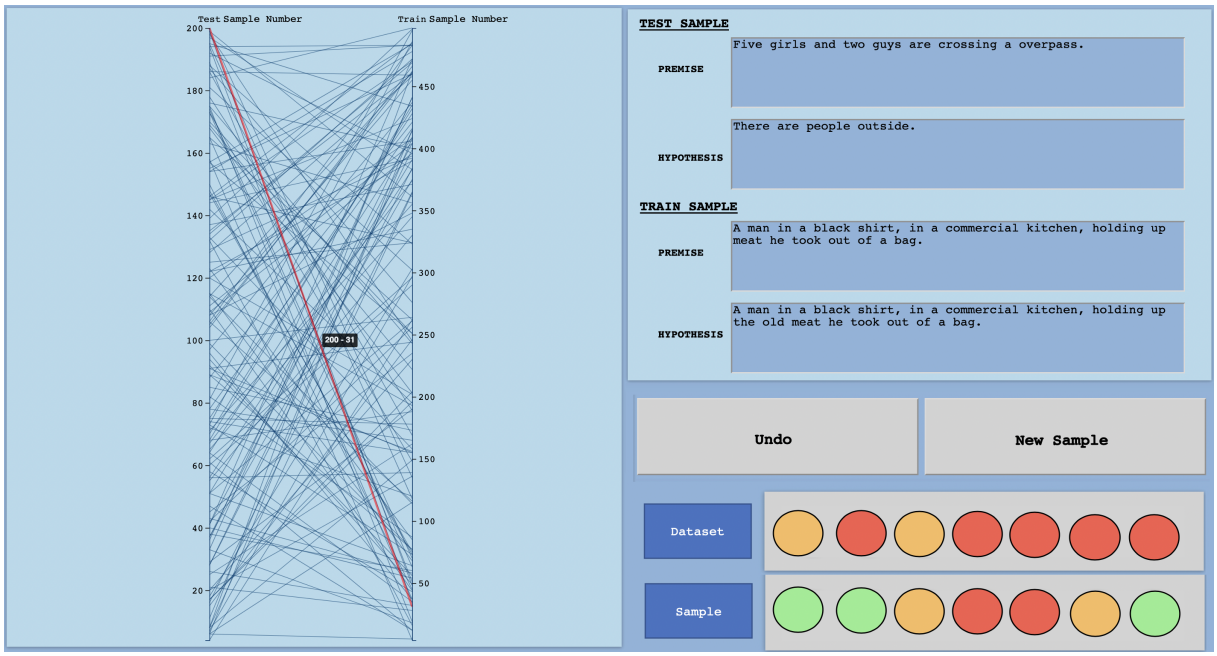


Figure 25: DQI_{c7} Visualization On New Sample Addition

it belongs to, as determined by the component one visualization. The sample's link is auto-selected and the textboxes are accordingly updated. The component value panels are updated as well, as illustrated in Figure 25. The previous state of the visualization is saved in a set of variables.

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

UI for Data Creation and Valiation: The UI design is two-fold. It targets two aspects of data creation- crowd source worker creation, and analyst review. The first phase uses colored flags to provide feedback to a crowd source worker about the quality of the sample they have created, so that they can fix it manually/with autofix assistance before submitting for higher return. The second phase uses the data visualizations to help the analyst determine if the sample should be added, rejected, or fixed.

A.7.8 Crowd-Source Worker:

The design choices made are heavily focused on the notion of providing simple, yet critical feedback to the crowd source worker, to enhance the quality of data created by means of minimizing spurious bias. The methods and principles used in building the interface used for SNLI's (Bowman et al., 2015) data collection process are the basis of our interface design. There are two types of feedback given in the UI, pre-submission and post-submission of the sample.

Instructions A sliding panel instruction tab is on the left corner of the screen. It consists of two sets of instructions. The first set goes over all general interface functionality descriptions, including post-submission user feedback. The second set specifically focuses on the pre-submission feedback loop.

Pre-Submission Feedback Loop: After reviewing the main instruction panel, the user can begin data creation. There is an instructions box displayed at all times on the main creation panel, which gives examples used in the original SNLI interface design, to make users understand the nature of the samples they are required to create. The premise field is auto-filled with captions from the Flickr30k corpus. This field can be changed to a fresh premise at any time by clicking on the 'new premise' button. The 3 types of hypothesis (entailment, neutral, and contradiction) must be entered in their respective fields.

DQI based on past history Following this, each hypothesis is evaluated individually with the premise. Henceforth, the use of the term sample denotes premise and only the hypothesis under consideration. The hypothesis under consideration can

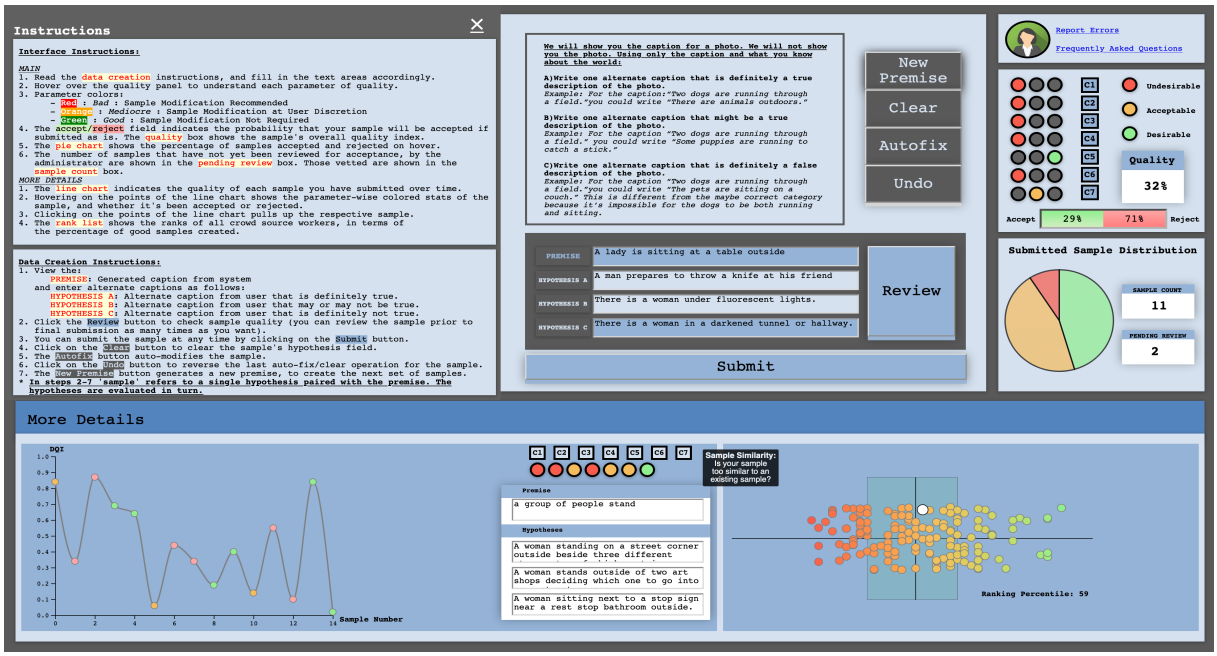


Figure 26: Crowd Source Worker View

be cleared at any time by clicking the 'clear' button. The user must click the 'Review' button at least once before submitting. The 'Review' button populates the DQI indication panel, which displays the values of the DQI components with respect to both the newly created sample and the existing set of accepted samples. The general aspect of data that is being analyzed by a component can be viewed on a tooltip, on mouseover of the component label. The messages displayed are as follows:

- Vocabulary: Does your sample contribute new words?
- Combinations: Does your sample contribute new combinations of words and phrases?
- Sentence Similarity: How similar is your hypothesis to all other premises or hypotheses?
- Word Similarity: How similar are all the words within your sample?
- PH Score: How similar is your hypothesis to the premise?
- Label Giveaway: Is your hypothesis too obvious for our system?
- Sample Similarity: Is your sample too similar to an existing sample?

Feedback Flags The values of the DQI components are indicated using a traffic signal analogy (red, yellow, and green), thereby indicating if a particular aspect of the data created might lead to bias. The colors respectively advise the user to stop, revise, and proceed in their sample creation tactics. The probability of the newly created sample being accepted/rejected is also displayed. Based on this feedback, the user can choose to: (i) manually fix their sample and review it again, (ii) 'auto-fix' the sample by paraphrasing it using concept net, (iii) submit the sample as is. Once the user is satisfied with the sample created, they can submit the sample. Once the sample has been submitted, the 'pending review' box is accordingly updated, as is the 'count' box for total number of submitted samples.

Post-Submission Feedback Loop: We retain the notion of a background expert reviewing samples to ensure that the sentences use appropriate ideas and language. Once the analyst reviews the sample and marks it as accepted/rejected (see section 8.2), the following updates occur on the crowdsource worker's UI [9] :

- The line chart on the secondary panel indicates the quality of the user's submitted samples over time. It is color coded according to

[9] these updates are only loaded at the start of each new user login session

whether the sample was accepted or rejected. On hovering over any one sample, the quality level of that sample are displayed on a tooltip. On click the sample appears in a text box.

- The 'pending review' box count on the main panel is decremented by one.
- The ranks are displayed using a box plot that calibrates ranks based on the percentage of accepted samples created by each user.
- The pie chart on the main panel is updated according to the accept/reject percentages.

Additional Communication Links: There are additional FAQ and Reporting Problem links present in the interface. The FAQs deal with data creation guidelines, and the Reporting Problems form is intended for technical issues only. This is in accordance with similar functionalities from the original SNLI interface. Figure 26 illustrates the crowdsource worker's UI.

A.7.9 Analyst:

Analysts' basic interface similar to crowdsource workers': The analyst interface is focused on the data validation process. The layout of the interface follows the same pattern as that of the crowd source workers interface. This is done so that the analyst understands the environment presented to the crowd source worker for data creation. The sliding panel for instructions, data entry boxes, DQI indication panel, and communication links are retained as is. The piechart, count box, pending review box, line chart, and rank box plot change depending on the annotator id associated with the sample being evaluated, as they represent the performance of that particular annotator.

Review Button The 'Next' buttons loads the next created sample set that must be reviewed. The text fields are filled with the premise and all hypotheses statements matching that premise. On clicking 'Review', the analyst reviews each hypothesis paired with the premise individually, as done in the crowdsource worker interface.

Buttons for Appropriate Visualizations: The DQI indication panel has buttons that link to each component's respective visualization. There are buttons present instead of labels for each component in this panel that can be used to navigate to each visualization in turn. The sample considered in the visualizations as the 'new sample' is the sample that is under review.

Data Validation The 'Accept' button can be used to accept the sample as is, and causes the piechart, pending review box, count box, rank box plot, and line chart for the annotator of the sample to be updated. The 'Reject' button is used mainly to discard samples that contain obscenities, have incoherent/ungrammatical hypothesis statements, and have hypothesis statements of length less than three words. If the sample has low quality, but can be converted to a higher quality adversarial sample with some modification and resubmitted, the 'Generate Adversarial Sample' button sends the sample to Text-Fooler. Samples that are auto-fixed at the analyst end in this manner are displayed as the yellow slice of the pie chart. Crowdsource workers receive lesser rewards for these samples. Figure 27 illustrates this.

A.8 AutoFix and TextFooler Examples

See Tables 142, 143.

A.9 User Study

AutoFix Suggestions: See Tables 144, 143.

NASA TLX: The NASA Task Load Index (NASA-TLX) is a subjective, multidimensional assessment tool that rates perceived workload in order to assess a task, system, or team's effectiveness or other aspects of performance (Hart, 2006).

NASA-TLX divides the total workload into six subjective subscales that are represented on a single page. There is a description for each of these subscales that the subject should read before rating. They rate each subscale within a 100-point range, with 5-point steps, as shown in Figure 28. Providing descriptions for each measurement can be found to help participants answer accurately (Schuff et al., 2011). The descriptions are as follows:

- **Mental Demand:** How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- **Physical Demand:** How much physical activity was required? Was the task easy or demanding, slack or strenuous?
- **Temporal Demand:** How much time pressure did you feel due to the pace at which the

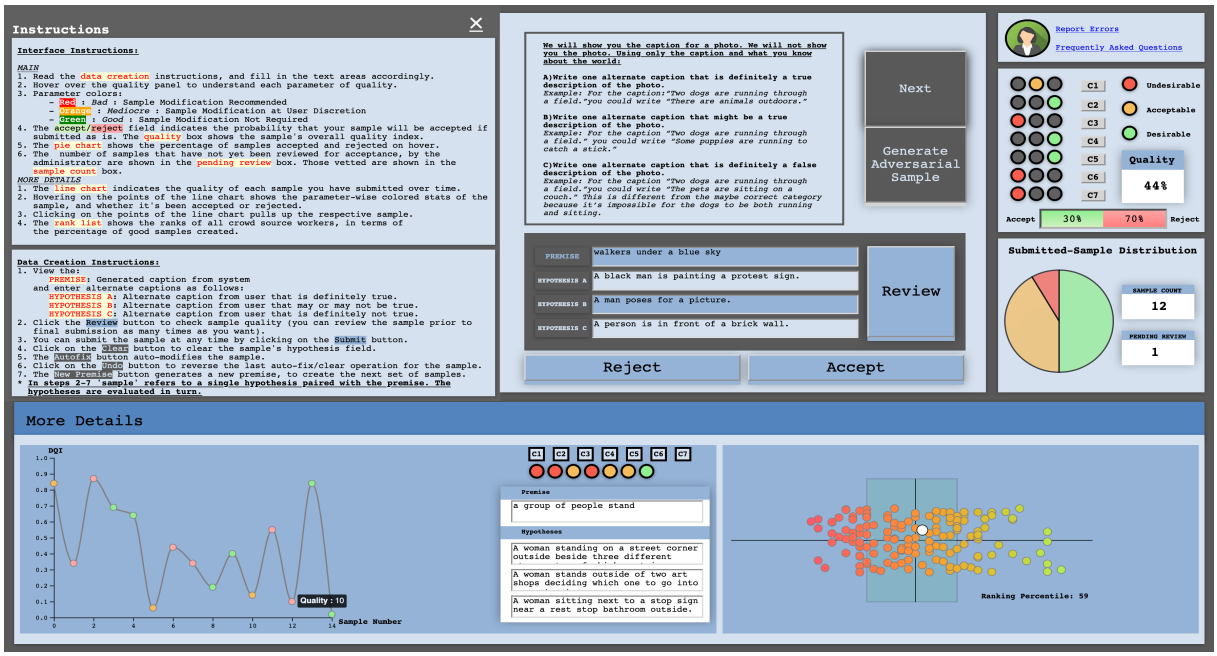


Figure 27: Analyst View

tasks or task elements occurred? Was the pace slow or rapid?

- **Performance:** How successful were you in performing the task? How satisfied were you with your performance?
- **Effort:** How hard did you have to work (mentally and physically) to accomplish your level of performance?
- **Frustration:** How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

We record participant demographics– age, gender, and occupation. We also ask participants to rate their familiarity with Visualization and NLP, on a scale of 1 (novice) to 5 (expert). Demographic information is shown in Figure **??**. Participants are asked to fill this form at the end of each round of the user study. We also record the number of questions participants successfully create, as well as a record of how often participants use each module in the full system round. At the end of the user study, participants are asked what their impression of data quality is, and their free response is recorded.

Subscale Wise Results: Individual results of the averaged subscales in Figure 7 are shown in Figures 30,31. Physical demand does not change significantly across user study rounds.

A.10 Expert and User Comments

Experts *(P)*: We present an initial prototype of our tool, to a set of three researchers with expertise in NLP and knowledge of data visualization, in order to judge the interface design. For each expert, the crowdworker interface and then analyst interfaces were demoed. Participants (P) could ask questions and make interaction/navigation decisions to facilitate a natural user experience. All the experts appreciated the easily interpretable traffic-signal color scheme and found the organization of the interfaces—providing separate detailed views within the analyst workflow– a way to prevent cognitive overload (too much information on one screen); P_2 said the latter *"...enhances readability for understanding the data at different granularities."*. P_1 suggested the inclusion of *"...a provenance module within the analyst views to show historical sample edits and overall data quality changes over time to understand how data quality evolves as the benchmark size increases... this would help with the bubble plot and tree map which will get more cluttered and complex as data size increases"*. Additionally P_3 remarked that *"The frequency of samples of middling quality should increase as benchmark size increases, but the initial exposure that analysts will have with higher or lower quality samples should lessen the learning curve as they are familiar enough with interface subtleties by the time they begin to encounter more*

Task	Description	Component
New Sample	Adds the sample under review to dataset and updates visualizations.	All
Undo	Removes sample under review from dataset and updates visualizations.	All
Randomize Split	Randomized re-sampling of data across splits in a 70:10:20 ratio.	Vocabulary
Undo Split	Reverses last random split generated.	Vocabulary
Save Split	Freezes split for the remainder of analysis.	Vocabulary
Changing Granularity	View granularity can be changed by selecting drop down option.	Inter-sample N-gram Frequency and Relation, N-Gram Frequency per Label
Change Heat Map View	Using the drop down, the heatmap shows word similarities for the (a) premise, (b) hypothesis, or (c) both sentences.	Intra-sample Word Similarity
Rebinning Histogram	By filling a new value in the textbox, the number of bins in the histogram changes to that value.	Intra-sample STS
Remove Outliers	Removes elements with frequency count less than median count of granularity being viewed.	N-Gram Frequency per Label
Include All Samples	Displays all elements for a granularity.	N-Gram Frequency per Label

Table 141: Task Descriptions for Visual Interfaces

Premise	Orig. Hypothesis	DQI	Suggested Words	New Hypothesis based on suggestions	New DQI
A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill	2.4650170	preparing,sleep	A woman is organizing to rest on a treadmill	2.5275722
The dog is catching a treat	The cat is not catching a treat	2.752542	catching	the cat is not getting a treat	3.6909140
Three young men are watching a tennis match on a large screen outdoors	Three young men watching a tennis match on a screen outdoors, because their brother is playing	2.6435402 891414217	young,watching, playing	Three youthful men observing a tennis match on a screen outdoors, because their brother is performing.	2.6787982

Table 142: A few samples for Autofix with Intra Sample STS in DQI

challenging cases."

Crowdworkers *(C)*: When presented with traffic signal feedback, crowdworkers report that the time and effort required to create high quality samples increases–*"You need to keep redoing the sample since when you see it's all red, you know it's probably not going to be accepted"*(C_3); however, they are more confident about their performance and sample quality *"...when there's green, I know I've done it right, and it cuts down on my having to create a lot of samples to get paid"* (C_{15}). We find that AutoFix usage [7] causes an unexpected increase in mental and temporal demand, as well as frustration; we attribute this to observed user behavior– *"I'm not sure how much I trust this recommendation without seeing the colors"*(C_{12}), and *"I'd prefer to change a couple of things since I can't see the feedback anymore*(C_{21}). The drastic improvement over all aspects (highest for frustration) in the case of using the full system is in line with this observation–*"This is so easy, I can create samples really fast, and I have a better chance of getting more accepted."*(C_8) and *"Now that I get the feedback along with the recommendation, I can see the quality improvement. So using the recommendation is now definitely faster."*(C_{12}). The number of questions created per round as well as system scores also follows this trend, across all types of crowdworkers.

Analysts *(A)*: In the case of direct quality feedback, i.e., traffic signals, analysts report an increased performance and find the task easier–*"... it's easier to directly choose based on quality... and it takes care of typos too, the typo samples are marked down so the work goes pretty fast"*(A_3). When analysts are shown the visualization interfaces, they are explicitly taught to differentiate the traffic signal colors in the visualizations as being indicative of how the sample affects the overall dataset quality, i.e., the colors in different component views represent individual terms of the components calculated over the whole dataset (analysts can toggle between the states of original dataset and new sample addition). We find that users initially find this more difficult to do– *"It takes a little time to figure out how to go through the views. I learned that in the samples I looked at, components three and seven seemed to be linked. So I'd look at those first the next time I used the system"* (A_6) and *"... it takes me some time to figure out how to read*

Premise	Orig. Hypothesis	DQI	New Hypothesis	New DQI	Label
A woman and a man sweeping the sidewalk.	The couple is sitting down for dinner.	2.416	The couple is meeting for dinner.	3.479	Contradiction
A woman enjoying the breeze of a primitive fan.	The woman has a fan.	2.127	The woman owns a fan.	2.733	Entailment
There is a man in tan lounging outside in a chair.	A man is preparing for vacation.	2.801	A man is arranging to take a vacation.	3.502	Neutral

Table 143: Examples for TextFooler, with DQI's Intra-sample STS values for existing SNLI samples.

Premise	Orig. Hypothesis	DQI	Suggested Words	New Hypothesis based on suggestions	New DQI
A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill	2.4650170	preparing,sleep	A woman is organizing to rest on a treadmill	2.5275722
The dog is catching a treat	The cat is not catching a treat	2.752542	catching	the cat is not getting a treat	3.6909140
Three young men are watching a tennis match on a large screen outdoors	Three young men watching a tennis match on a screen outdoors, because their brother is playing	2.6435402 891414217	young,watching, playing	Three youthful men observing a tennis match on a screen outdoors, because their brother is performing.	2.6787982
A man in a green apron smiles behind a food stand	A man smiles	3.2367785	smiles	A person is grinning.	6.303777

Table 144: A few samples for Autofix with ISSTS in DQI

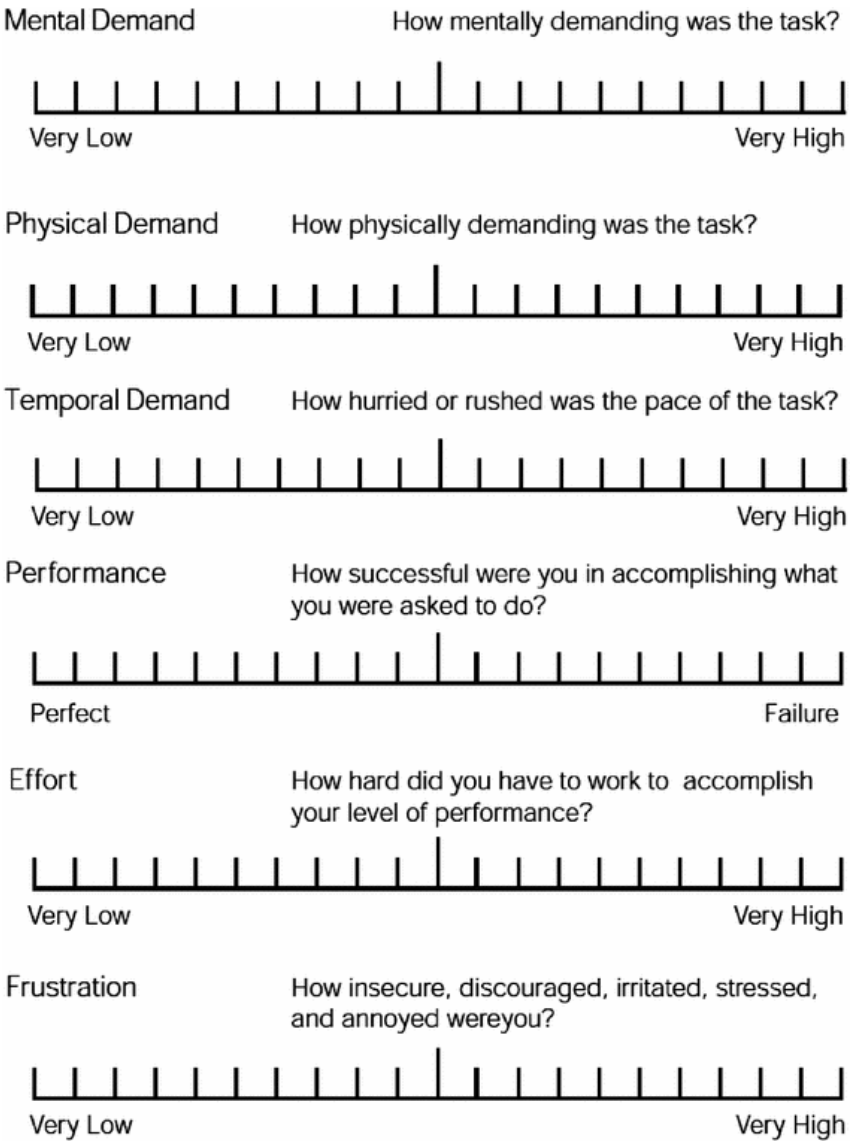


Figure 28: NASA TLX Form

the interfaces effectively, but it does make me more secure in judging sample quality at multiple granularities and that would help if I was doing this for a particular application"(A_1). Analysts averaged

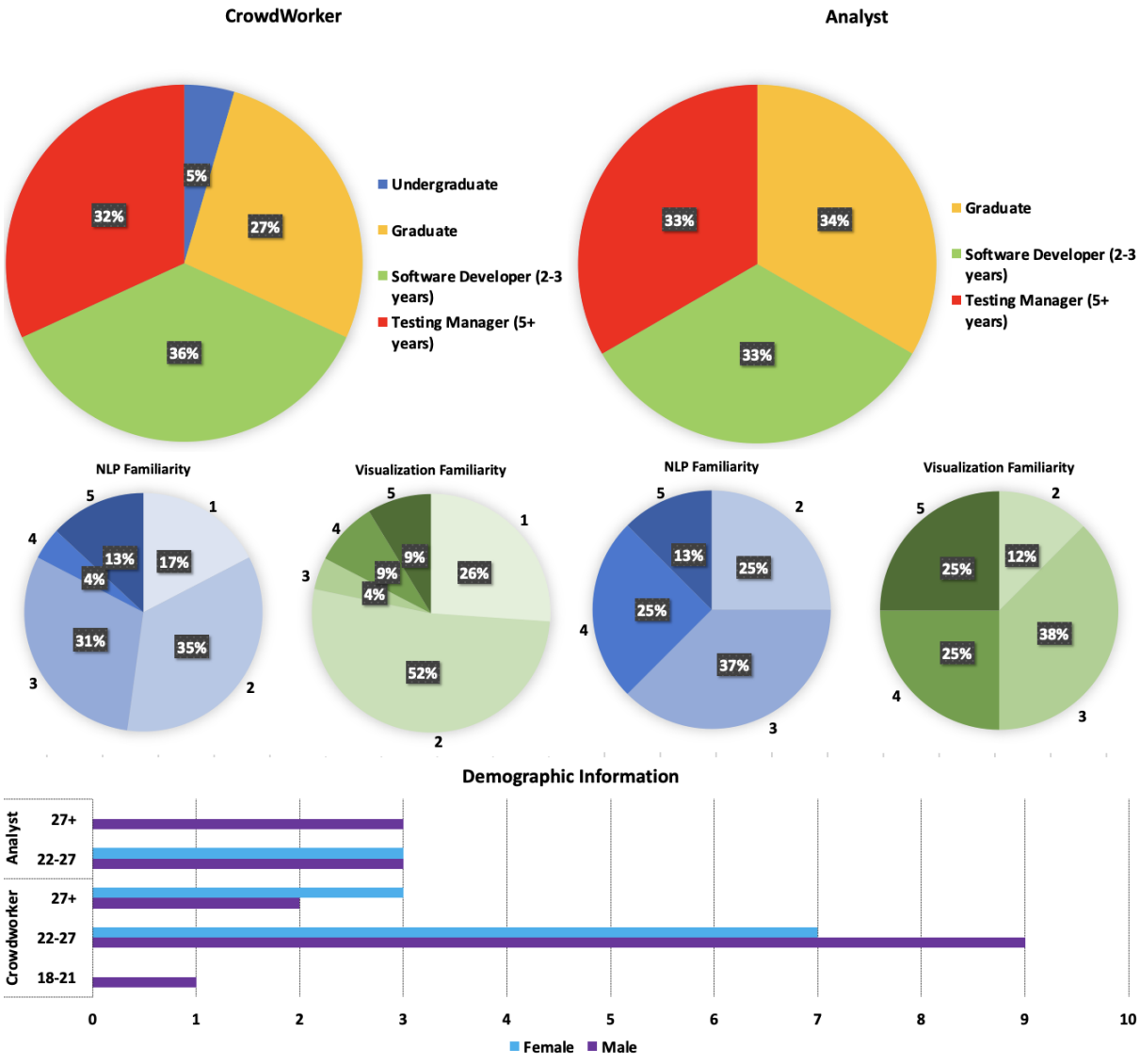


Figure 29: Demographic information for the User Study

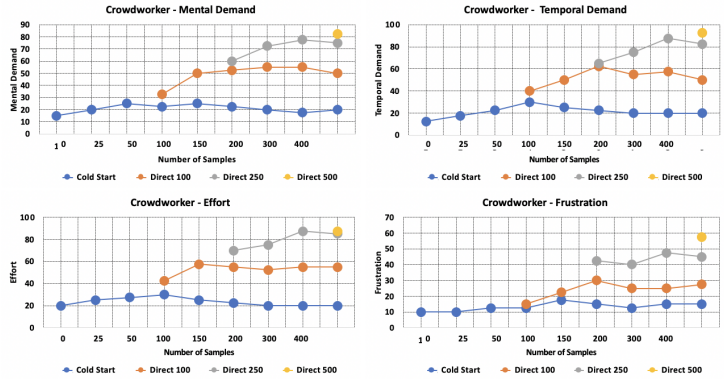


Figure 30: NASA TLX– Crowdworker Subscale Results

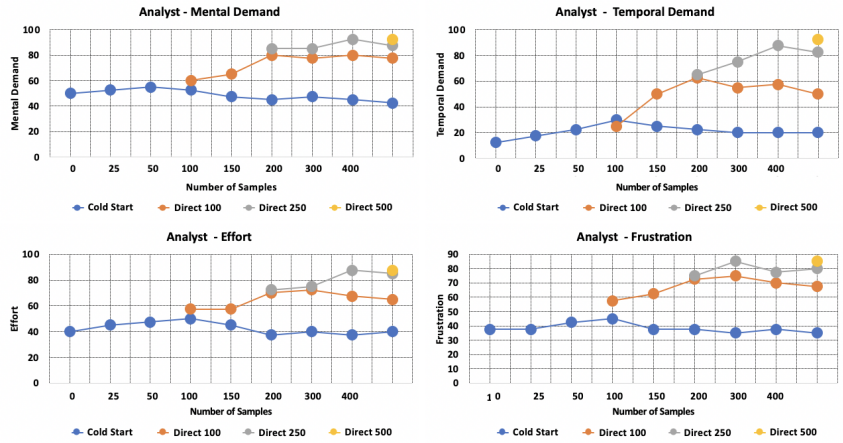


Figure 31: NASA TLX– Analyst Subscale Results

behavior on TextFooler models the conventional approach quite closely, as analysts are seen to have a tendency to either– *"... deciding to reject or repair is difficult when you don't have the sample or dataset feedback... and what if the repaired sample still isn't good enough?"*(A_4), or– *" I like having this option to repair... I don't need to waste time on analyzing something that isn't outright an accept or reject, I can send it to be repaired and come back to it later"*(A_8). When shown the full system, analysts also report improvement in all aspects, particularly mental demand and performance–*"I can be sure of not having to redo things since it's likely that I will be able to get a low hypothesis baseline using this system"*(A_2, A_1). The visualization usage also improves– *"... I went to component three right off the bat this time, I knew that I could look at the linked components..."* (A_6). Altogether, sample evaluation by analysts increases, following this trend, and analysts are more assured of their performance.