UCLA Samueli Computer Science

SEAL Software Evolution & Analysis Laboratory

Plus lab

VT VIRGINIA TECH

Human-in-the-Loop Synthetic Text Data Inspection with Provenance Tracking

Hong Jin* Kang

Fabrice* Harel-Canada

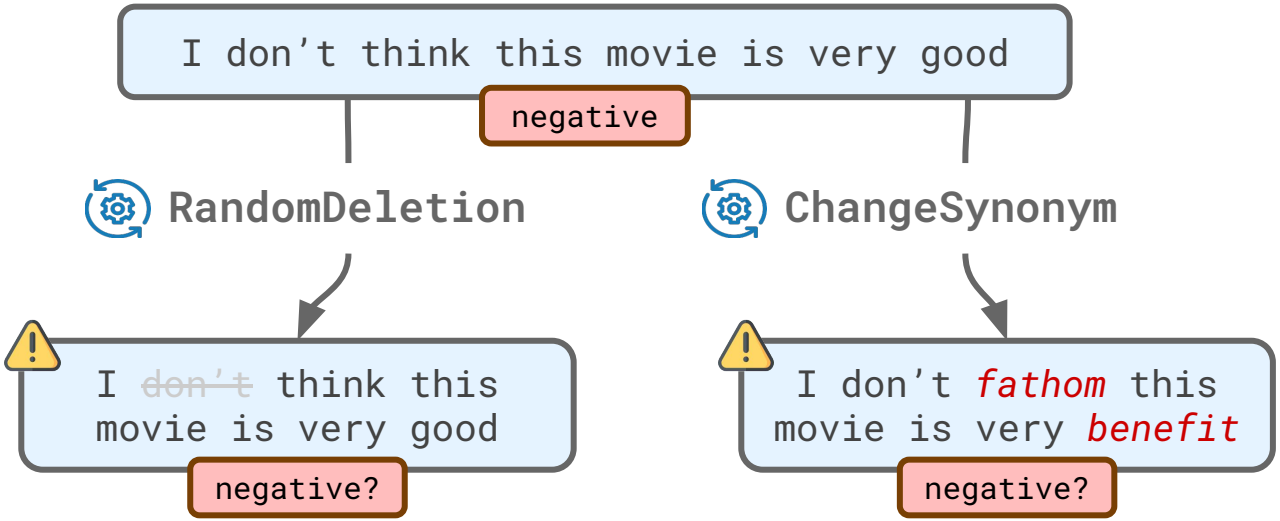
Muhammad Ali Gulzar

Nanyun Peng

Miryung Kim

So what's the problem?

Data augmentation (DA) is commonly used to expand datasets and improve models but often **compromises** quality



QA to the rescue?

Human effort is expensive

Individual inspection is time-consuming

Repetitive tasks undermine focus

A new tool for scaling text quality analysis

INSPECTOR

Provenance Tracking

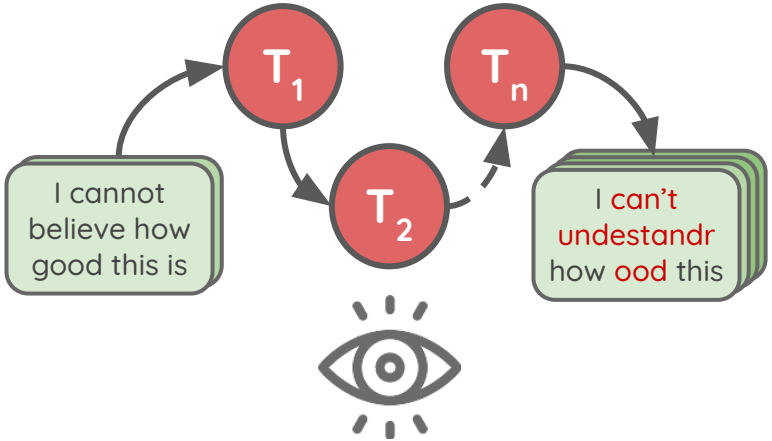
1. **Transforms** applied during augmentation
2. **Text features** that might cause context-specific failures (e.g. negations)

Assistive Labeling

1. **Quality Metrics**
   - Alignment
   - Fluency
   - Grammaticality
2. **LLM Guidance**

Transformation Provenance

- Tracking **sequence of transforms** applied during augmentation
- Helps identify specific transforms that frequently result in low-quality or mislabeled data
- Implemented as a simple wrapper compatible with many augmentation frameworks
  - `NL-Augmenter`, `Sibyl`, `EDA`, `TextAttack` and more

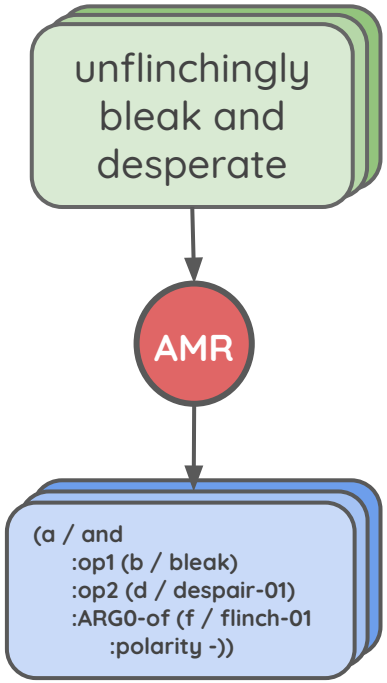


Input	Transform	Args	Output
I cannot …	T_1	**{prob: 0.2}**	I can't be…
I can't be…	T_2	**{idx: 3}**	I can't un…
I can't un…	T_n	**{rand: 0.4}**	I can't un…

Feature Provenance

- Grouping on shared linguistic features derived from **Abstract Meaning Representation** graphs
- Helps identify how specific features might misbehave in the presence of certain transforms

"We should probably focus on every case where ***`WordDeletion`*** *and* ***`:negation`*** *co-occur!"*



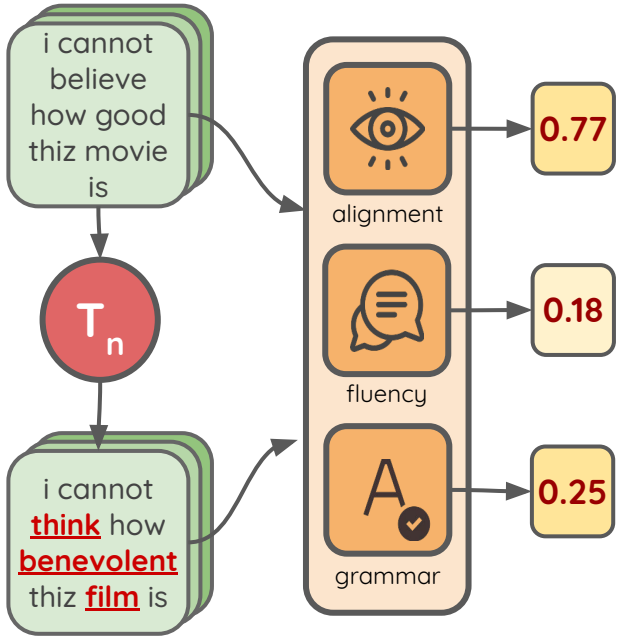
Example Features

```
:negation
:polarity
:purpose
:range
:condition
:expressive
:choice
```

⋮

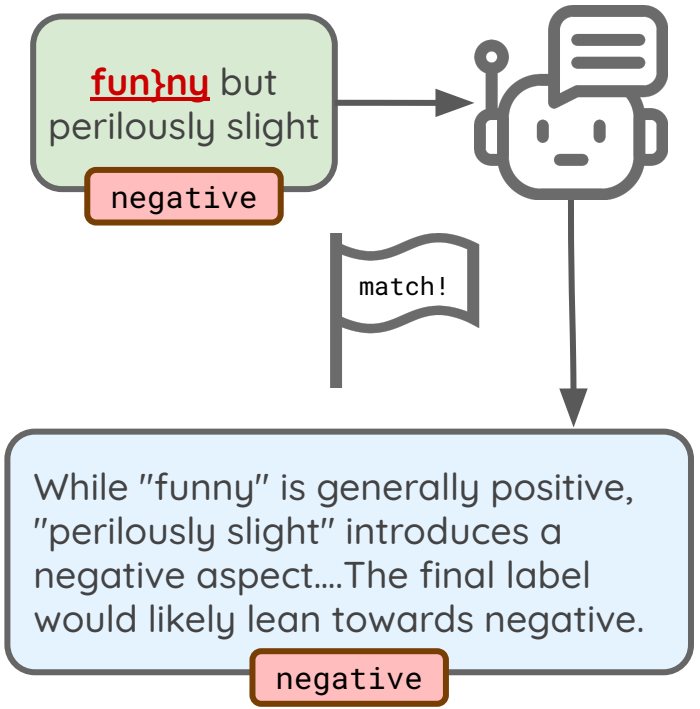
Quality Metrics

- **Alignment**
  - *Does the label match the new text?*
  - Scored via confident learning, which predicts label errors
- **Fluency**
  - *Does the new text sound realistic?*
  - Scored by change in LLM perplexity
- **Grammaticality**
  - *Does the new text violate grammar?*
  - Scored by change in grammar errors



LLM Guidance

- How do humans respond to LLM assistance while inspecting texts?
- Using an LLM (`gpt-3.5-turbo`) to provide pre-loaded:
  - Prediction
  - Justification
- New flag indicating when LLM disagrees with current label




Data Table displaying augmented data
Selected data: Showing both selected and unselected data
2 selected (9, 322);
57 inspected; 35 are high quality
Rank data by similarity to selected
Inspected data: Showing both inspected and uninspected data
Show only inspected
Show only uninspected
Show both inspected and uninspected
Cancel selection of
already inspected data (1 instances)
all selected data (2 instances)
sentence
Align
Fluency
Grammar
LLM predictions
LLM explanation
label-LLM label consistency
label
high quality
Inspected
audience riveted.
it takes a strange kind of laziness to waste the talents of robert forster, anne meara, eugene le y, and reginald veljohnson all in
1
0.99
0.57
neg
The sentiment of the text is negative. The use of the word 'waste' implies that the talents of the actors mentioned were not utilized effectively, which is a
consistent
neg
... the film suffers from ? a lack of , humor ; (something needed to balance out the ! violence)...
1
0.97
0.86
neg
The sentiment of the text is negative. The use of the phrase 'suffers from' indicates that there is a problem with the film. The text goes on to mention a 'lack of
consistent
neg
we root for (clara and paul), still similar them, though perhaps it's an emotion nigher to pity.
0.02
0.96
0.86
neg
The sentiment of the text is negative. The phrase 'though perhaps it's an emotion nigher to pity' suggests that the speaker feels
inconsistent
pos
2
Transform Provenance
Transform
Mark all as high quality
Mark all as inspected
Examples of transformed texts ar labels
RandomCharSubst (replace a randomly selected character in the sentence with another character)
141 total matching instances, select all
1 selected instances (9), 14 already inspected; 11 high quality)
it takes a strange kind of laziness to waste the talents of robert forster, anne meara, eugene levy, and reginald veljohnson all in the same movie.
nicks, seemingly uncertain what's going to make people laugh, runs the gamut from stale parody to raunchy sex gags to formula romantic comedy.
it does n't believe in itself, it has no sense of humor. it's just plain bored.
ChangeSynonym (replace words with other words with the same meaning)
32 total matching instances, select all
1 selected instances (322), 7 already
the film's hackneyweedll-worn message is not heavalped by the thflinmsy characterizations, nonexistent plotgame and pretentious vopti-sucal style.
although laced with hutemeper and a
1
Feature Provenance
Feature
Mark all as high quality
Mark all as inspected
Examples of original text and labels
Contains a conjunction (e.g., "and", "either")
(309 total matching instances, ct all
ected tances (9,322), 30 already inspected; 17 high quality)
it takes a strange kind of laziness to waste the talents of robert forster, anne meara, eugene levy, and reginald veljohnson all in the same movie.
the film's hackneyed message is not helped by the thin characterizations, nonexistent plot and pretentious visual style.
unflinchingly bleak and desperate
Has a description of a location, e.g., "in Los Angeles"
(70 total matching instances, select all
1 selected instances (9), 18 already inspected; 11 high quality)
it takes a strange kind of laziness to waste the talents of robert forster, anne meara, eugene levy, and reginald veljohnson all in the same movie.
audrey tatou has a knack for picking roles that magnify her outrageous charm, and in this literate french comedy, she's as morning-glory exuberant as she was in amélie.
this illuminating documentary transcends our preconceived vision of the holy land and its inhabitants, revealing the
Quality Metrics
C
LLM Predictions
D
Transform Provenance
A
Feature Provenance
B

User Study Design

Baseline

- A simplified version of INSPECTOR
- Lacks provence and assistive labeling

Datasets

- **Sentiment Analysis:** SST-2
- **Hate Speech:** Tweet Eval (Hate)

Participants

- Recruited 15 Computer Science students - 11 PhD, 2 MS, 2 Undergrad
- Users had varied levels of familiarity with data inspection + ML

Protocol

- Identify texts with correct labels as efficiently and accurately as possible
- Inspect one dataset w/ INSPECTOR and the other w/ Annotator

Research Questions

Reduction in Human Effort

Does INSPECTOR increase efficiency in identifying texts with correct labels?

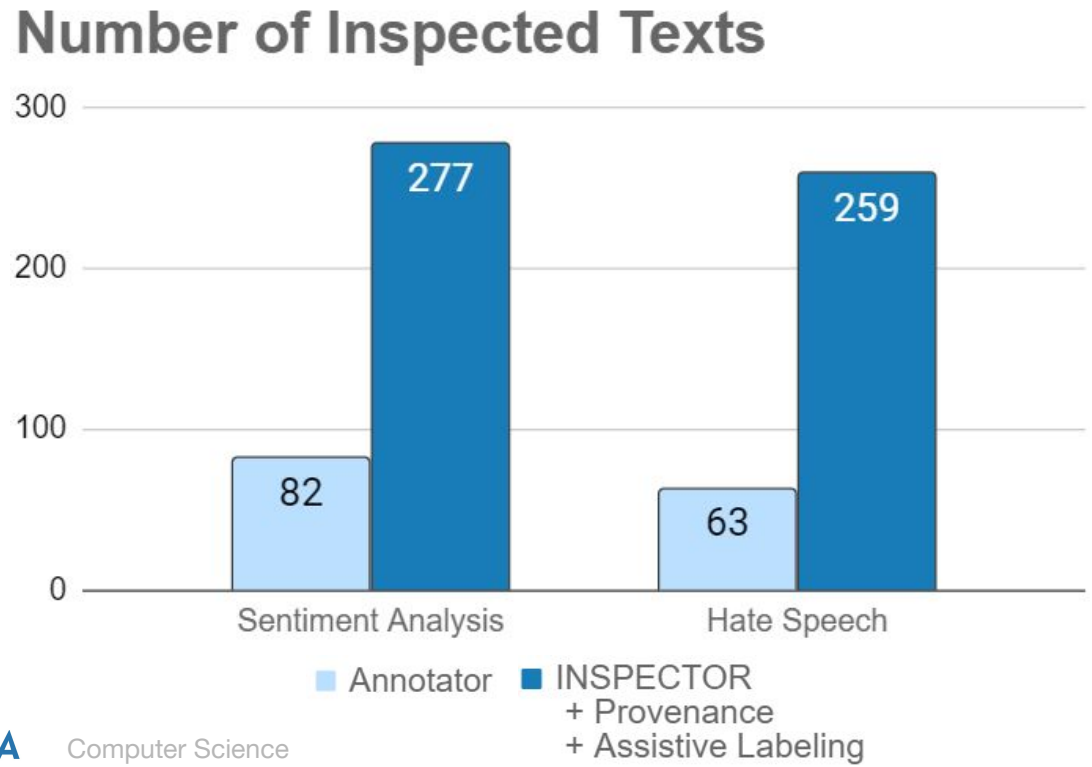
Technique Utility

How useful is each effort reduction technique offered by INSPECTOR?

Model Robustness

Are models more robust when trained using data identified with INSPECTOR?

Reduction in Human Effort



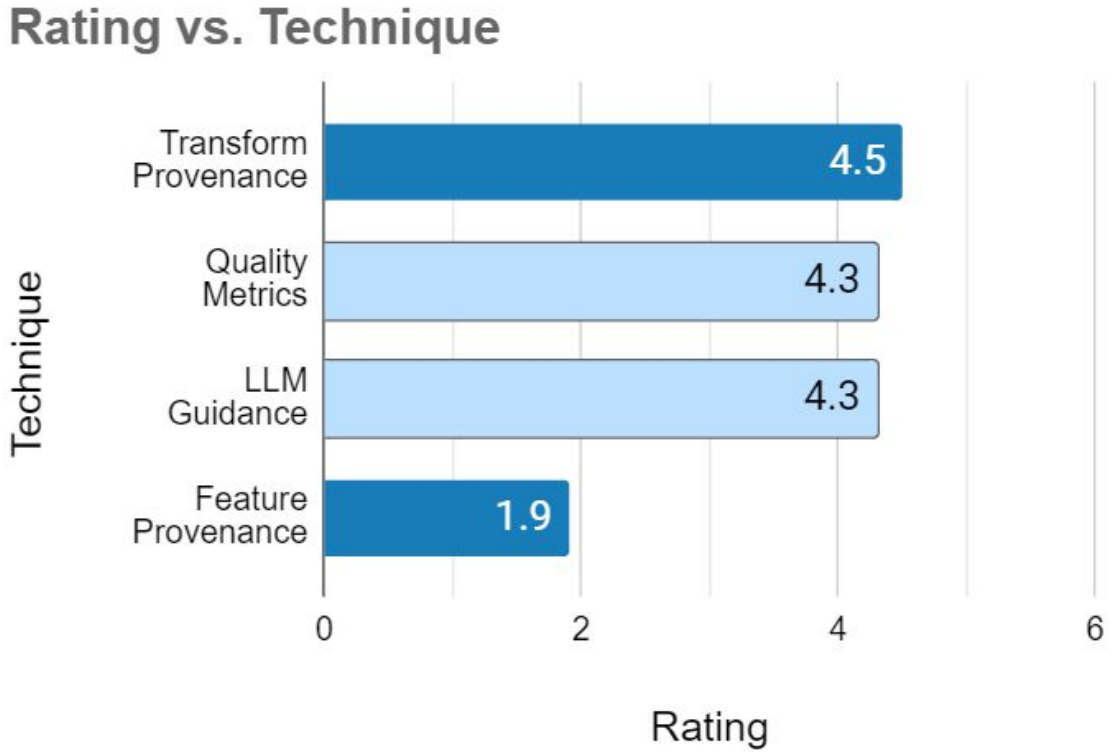
Takeaway

INSPECTOR helped analyze

3✖ - 4✖

more data in the same amount of time

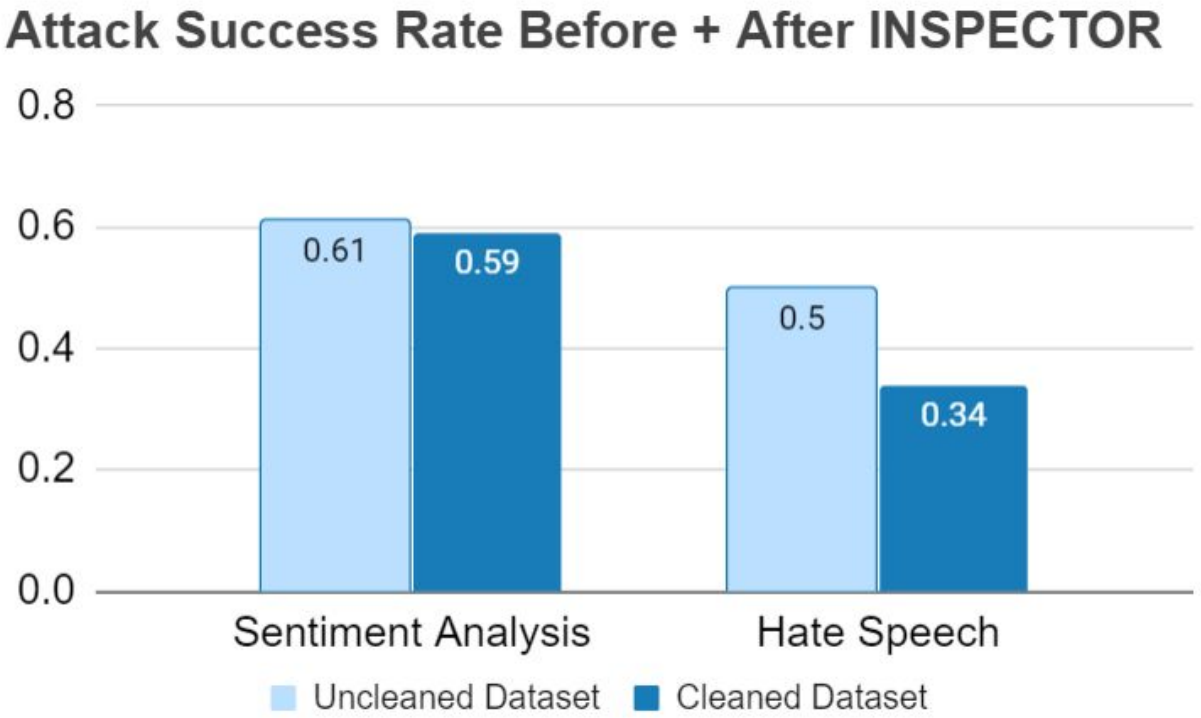
Technique Utility



Takeaway

Transform provenance rated most as useful, but no single technique was universally preferred by all

Model Robustness



Takeaway

Data cleaned with INSPECTOR **improved model robustness** by up to **32%**

Takeaways

INSPECTOR

Effective tool for **scaling** text inspection
Empowers humans to build diverse, **provenance-guided** strategies

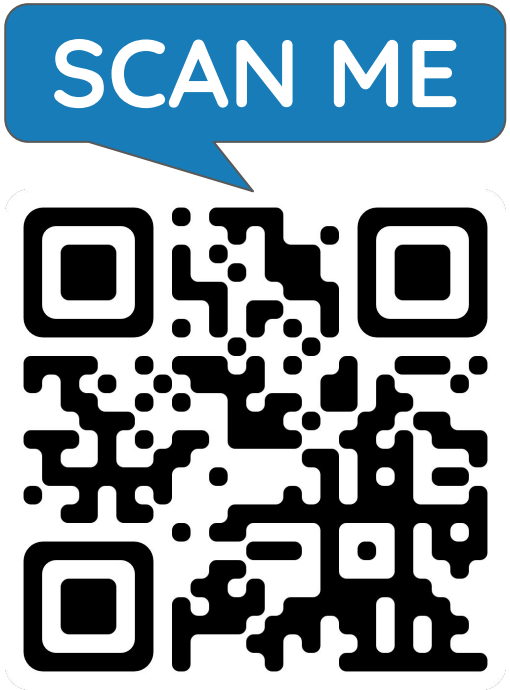
"[I could] reason about whether a specific transform can lead to a reduction of data quality."

"sentences with a grammar score < 0.92 are almost always low-quality."

Use Study

- Inspect **3✖ - 4✖** faster
- Transform provenance preferred over LLMs and quality metrics
- Up to 32% more robust

https://github.com/UCLA-SEAL/ProvenanceInspector



paper link

Thank You!

Questions?

Outline

- Motivation
  - DA expands datasets by applying intuitive transforms to existing text
  - While DA generally improves model performance, the new data is often noisy and of lower quality
  - Inspecting data quality at scale is challenging because of costs and human psychology (boredom with the task).
- We propose Inspector, a tool for scaling text quality analysis which features:
  - Provenance Grouping:
    - Transformations
    - AMR Features
  - Assistive Labeling: alignment, grammaticality, fluency, and LLM best-guess
- Demo interface
  - Show main panels and describe the intended workflow

Outline

- Study Design
  - Participants
  - Datasets
  - Task
- Results
  - RQ1: reduction in human effort? 3-4X
  - RQ2: how helpful is each of inspector’s features?
  - RQ3: models more robust? Up to a 32% improvement
- Takeaways
  - Transform provenance regarded as more useful than LLM best guess (features not at all)
  - Assistive labeling is helpful for trust and users felt more confident in their data quality
- Final slide with links

Grouping by Provenance

Transforms

- **RandomCharSubst**

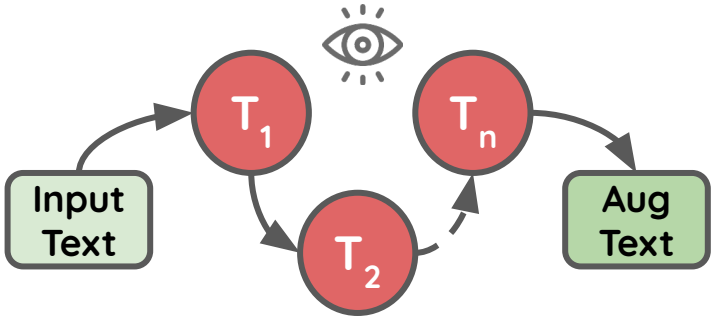
  Replace a randomly selected character in the sentence with another character

- **ChangeHypernym**
- **ChangeHyponym**
- **ChangeLocation**

Provenance Grouping

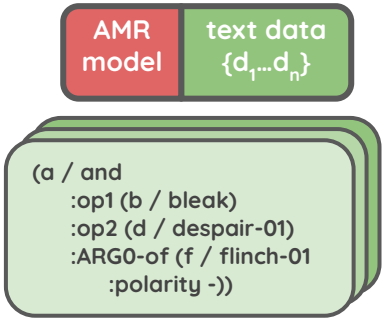
Transformations

Tracking all transforms used to arrive at potentially noisy augmented text



Features

Extract abstract meaning representation (AMR) features from texts

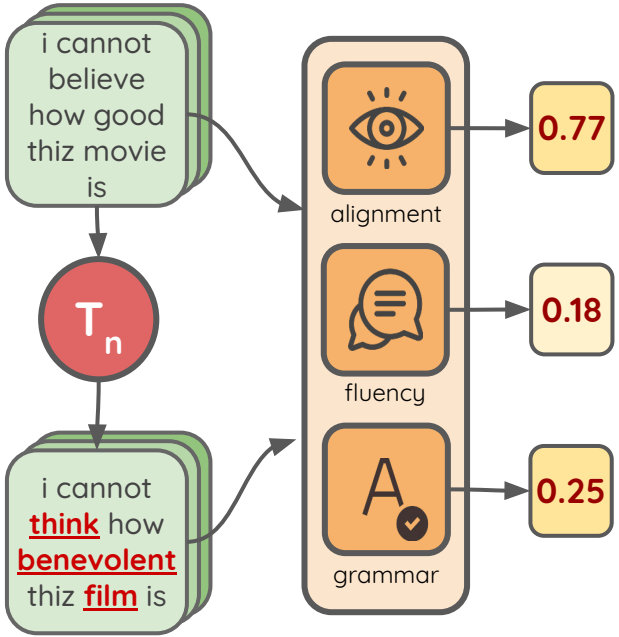


Examples

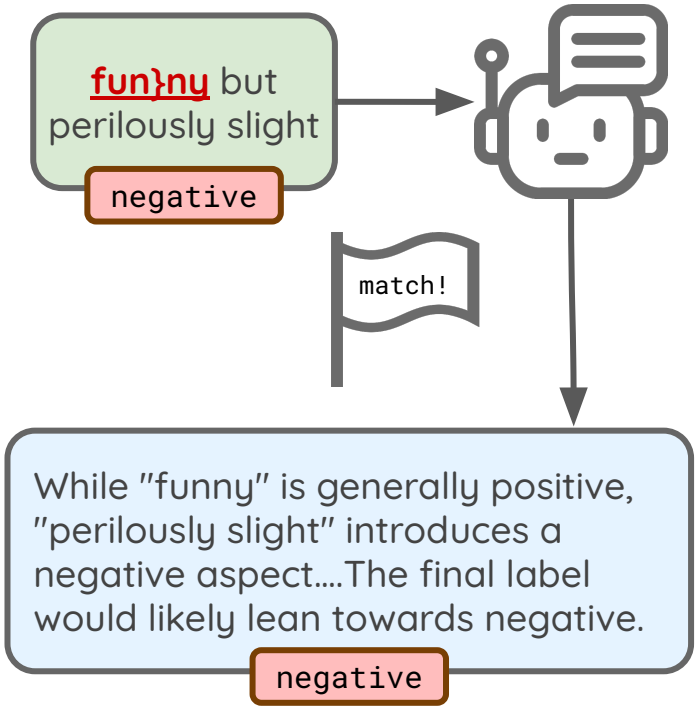
```
:negation
:polarity
:choice
```

⋮

Quality Metrics



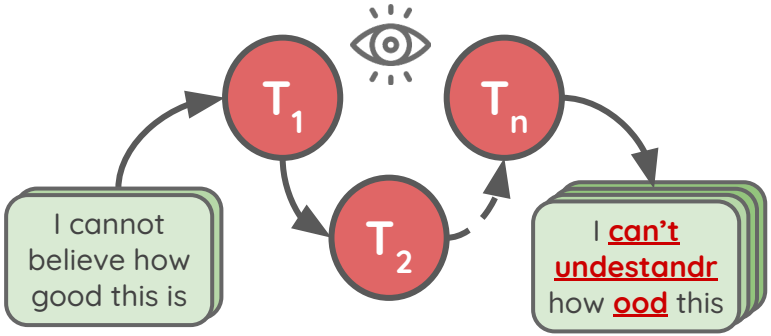
LLM Guidance



Provenance Grouping

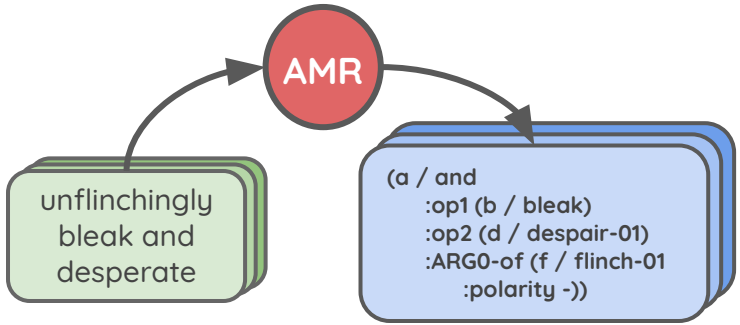
Transformations

Tracking all transforms used to arrive at potentially noisy augmented text



Features

Extract abstract meaning representation (AMR) features from texts



Baseline

- A simplified version of INSPECTOR
- Lacks provence and assistive labeling

Datasets

- **Sentiment Analysis:** SST-2
- **Hate Speech:** Tweet Eval (Hate)

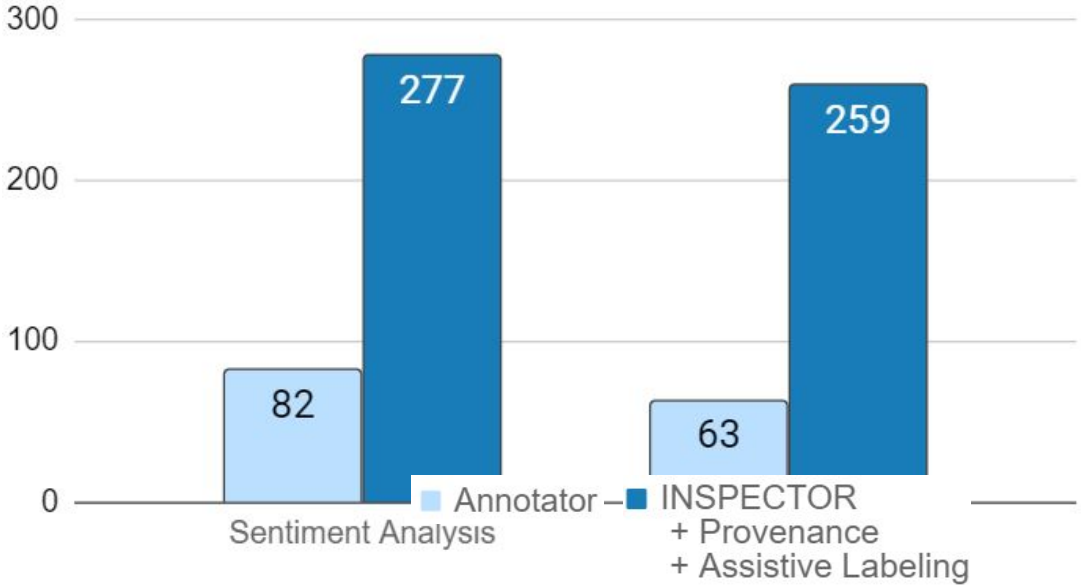
Protocol

- Identify texts with correct labels as efficiently and accurately as possible
- Inspect one dataset w/ INSPECTOR and the other w/ Annotator

Participants

- Recruited 15 Computer Science students - 11 PhD, 2 MS, 2 Undergrad
- Users had varied levels of familiarity with data inspection + ML

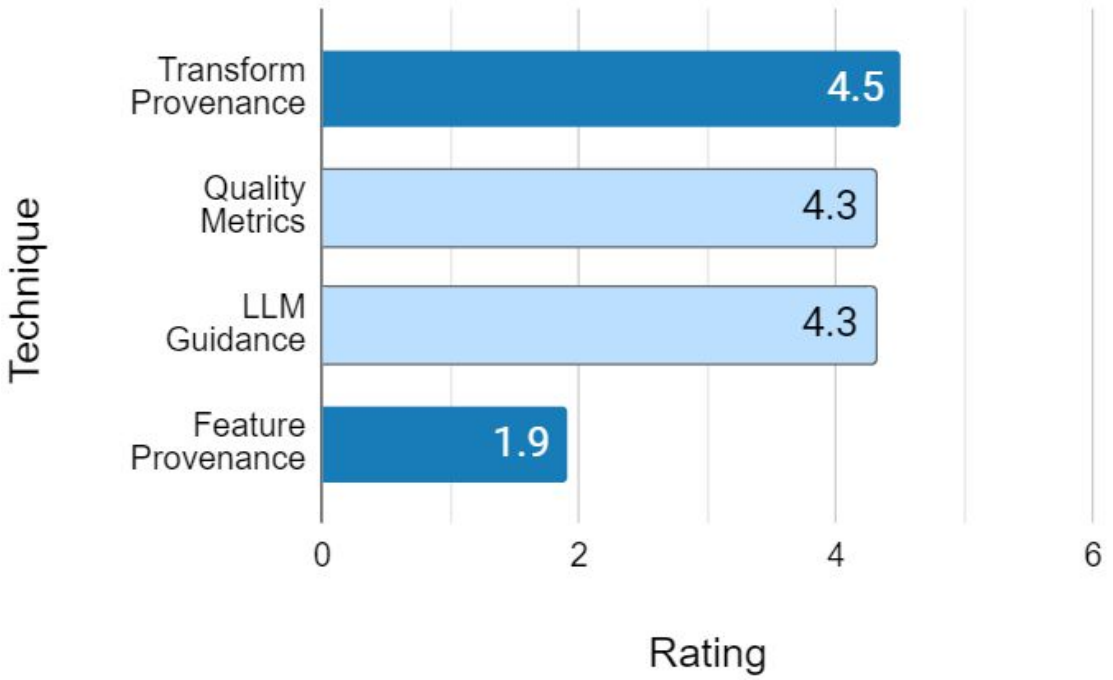
Number of Inspected Texts



Takeaway

INSPECTOR helped analyze
3✖ - 4✖ more data
in the same amount of time

Rating vs. Technique



Takeaway

Transform provenance rated most as useful, but no single technique was universally preferred by all

Attack Success Rate Before + After INSPECTOR

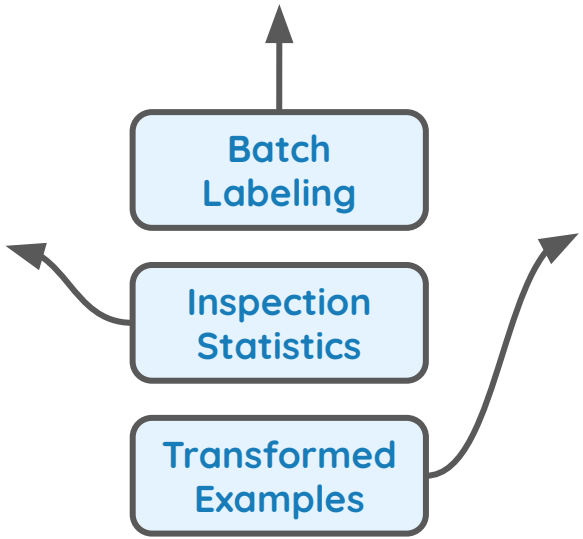
Takeaway

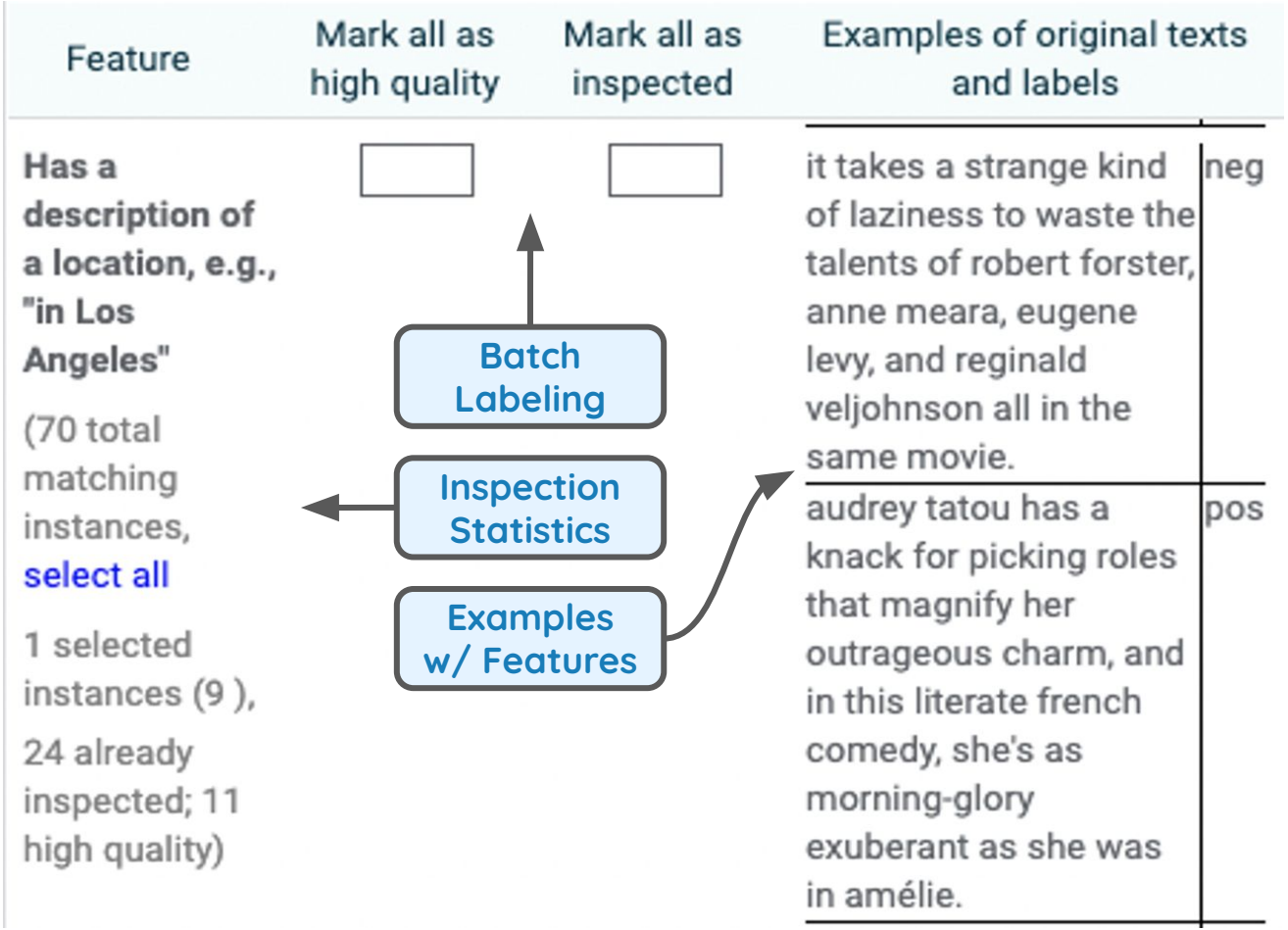
Data cleaned with INSPECTOR **improved model robustness** by up to **32%**

Takeaways


Main Takeaway
Effective tool for **scaling** inspection
Empowers humans to build diverse, **provenance-guided** strategies
"[I could] reason about whether a specific transform can lead to a reduction of data quality."
"sentences with a grammar score < 0.92 are almost always low-quality."

Transform	Mark all as high quality	Mark all as inspected	Examples of transformed texts and labels	
RandomSwapQwerty (substitues random characters with adjacent keys on a keyboard) (35 total matching instances, select all 2 selected instances (28,47), 7 already inspected; 5 high quality)	☐	☐	Niggas be getting on here crying ɘgout not getting no play everyday lol there are women that like weird niggas,funny niggas, athletic niggas,thug niggas etc. Find a bitch in yo arena and stfu	not hate
			In a country scarfred by violence, LGBTI activism is not always welcome. This is Sofia and Daniel's story.	not hate

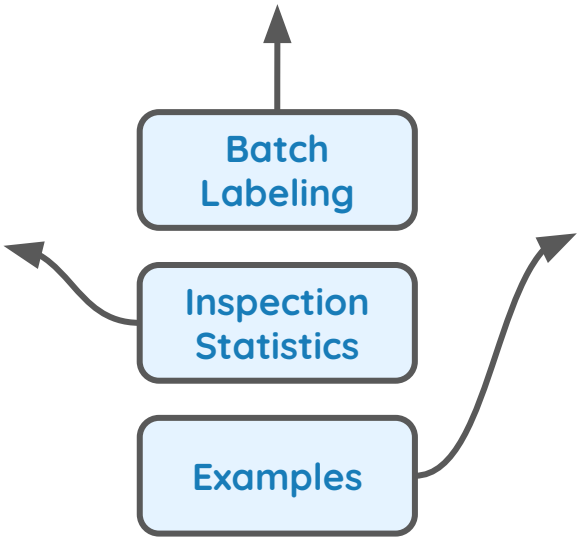



Feature
Mark all as high quality
Mark all as inspected
Examples of original texts and labels
Has a description of a location, e.g., "in Los Angeles"
(70 total matching instances, select all
1 selected instances (9),
24 already inspected; 11 high quality)
Batch Labeling
Inspection Statistics
Examples w/ Features
it takes a strange kind of laziness to waste the talents of robert forster, anne meara, eugene levy, and reginald veljohnson all in the same movie.
neg
audrey tatou has a knack for picking roles that magnify her outrageous charm, and in this literate french comedy, she's as morning-glory exuberant as she was in amélie.
pos

sentence	Align	Fluency	Grammar	LLM predictions	LLM explanation	label-LLM label consistency	label	👍 high quality	✓ Inspected
runs on the pure adrenalin of pacino'c performance.	1	0.5	1	pos	The sentiment of the text is positive. The phrase 'pure adrenalin' suggests excitement and intensity, which is	consistent	pos		
also reachn't embarrassed to make you is for the tissues	1	0.33	1	neutral	The sentiment of the text is neutral. The text is difficult to interpret as it appears to be a jumbled collection of words and does	inconsistent	pos		

Quality Metrics

LLM Guidance


Batch Labeling
Inspection Statistics
Examples