# De-Noising Document Classification Benchmarks via Prompt-based Rank Pruning: A Case Study




**Matti Wiegmann** Martin Potthast Benno Stein

Bauhaus-Universität Weimar Kassel University hessian.AI ScaDS.AI

webis.de

# De-Noising Document Classification Benchmarks

Evaluation of document classification task is based on benchmark datasets.

Those benchmarks are prone to label noise.

- Subjectivity.

  `Is this text a product description or a product advertisement?`

- Many classes.

  `Which of the 188 cognitive biases occur in this text?`

- Need for expert knowledge.

  `Is this LLM-generated essay correct?`

# De-Noising Document Classification Benchmarks

Evaluation of document classification task is based on benchmark datasets.

Those benchmarks are prone to label noise.

- Subjectivity.

  `Is this text a product description or a product advertisement?`

- Many classes.

  `Which of the 188 cognitive biases occur in this text?`

- Need for expert knowledge.

  `Is this LLM-generated essay correct?`

# De-Noising Document Classification Benchmarks

Evaluation of document classification task is based on benchmark datasets.

Those benchmarks are prone to label noise.

- Subjectivity.

  `Is this text a product description or a product advertisement?`

- Many classes.

  `Which of the 188 cognitive biases occur in this text?`

- Need for expert knowledge.

  `Is this LLM-generated essay correct?`

# De-Noising Document Classification Benchmarks

Evaluation of document classification task is based on benchmark datasets.

Those benchmarks are prone to label noise.

- Subjectivity.

  `Is this text a product description or a product advertisement?`

- Many classes.

  `Which of the 188 cognitive biases occur in this text?`

- Need for expert knowledge.

  `Is this LLM-generated essay correct?`

# De-Noising Document Classification Benchmarks

Evaluation of document classification task is based on benchmark datasets.

Those benchmarks are prone to label noise.

- Subjectivity.

  `Is this text a product description or a product advertisement?`

- Many classes.

  `Which of the 188 cognitive biases occur in this text?`

- Need for expert knowledge.

  `Is this LLM-generated essay correct?`

$\rightsquigarrow$ Label noise deteriorates benchmarks and may change *model score*, *score difference*, and *model order*.

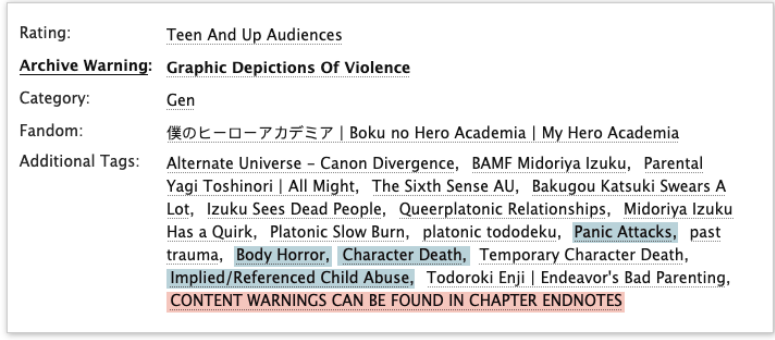
# De-Noising Document Classification Benchmarks

## Dataset

- Fiction documents w/ trigger warnings.[1]
- Labels inferred via weak supervision.

  Via authors' tags, annotations, tag relations, heuristics

Author tags of a document



| | |
|---|---|
| Rating: | Teen And Up Audiences |
| **Archive Warning:** | **Graphic Depictions Of Violence** |
| Category: | Gen |
| Fandom: | 僕のヒーローアカデミア \| Boku no Hero Academia \| My Hero Academia |
| Additional Tags: | Alternate Universe - Canon Divergence, BAMF Midoriya Izuku, Parental Yagi Toshinori \| All Might, The Sixth Sense AU, Bakugou Katsuki Swears A Lot, Izuku Sees Dead People, Queerplatonic Relationships, Midoriya Izuku Has a Quirk, Platonic Slow Burn, platonic tododeku, Panic Attacks, past trauma, Body Horror, Character Death, Temporary Character Death, Implied/Referenced Child Abuse, Todoroki Enji \| Endeavor's Bad Parenting, CONTENT WARNINGS CAN BE FOUND IN CHAPTER ENDNOTES |

[PitViperOfDoom, 2016]

[1] Wiegmann et al. Trigger Warning Assignment as a Multi-Label Document Classification Problem. ACL 2023

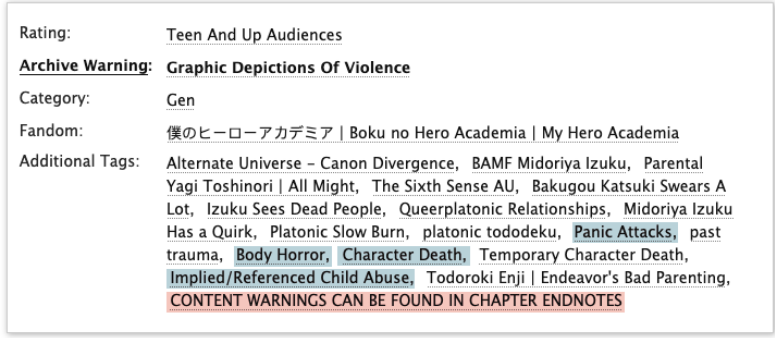
# De-Noising Document Classification Benchmarks

## Dataset

- Fiction documents w/ trigger warnings.[1]
- Labels inferred via weak supervision.

  Via authors' tags, annotations, tag relations, heuristics
- Various sources of label noise:
  - Annotation errors.
  - Author subjectivity.
  - Heuristics fail.

Author tags of a document



| | |
|---|---|
| Rating: | Teen And Up Audiences |
| **Archive Warning:** | **Graphic Depictions Of Violence** |
| Category: | Gen |
| Fandom: | 僕のヒーローアカデミア \| Boku no Hero Academia \| My Hero Academia |
| Additional Tags: | Alternate Universe - Canon Divergence, BAMF Midoriya Izuku, Parental Yagi Toshinori \| All Might, The Sixth Sense AU, Bakugou Katsuki Swears A Lot, Izuku Sees Dead People, Queerplatonic Relationships, Midoriya Izuku Has a Quirk, Platonic Slow Burn, platonic tododeku, Panic Attacks, past trauma, Body Horror, Character Death, Temporary Character Death, Implied/Referenced Child Abuse, Todoroki Enji \| Endeavor's Bad Parenting, CONTENT WARNINGS CAN BE FOUND IN CHAPTER ENDNOTES |

[PitViperOfDoom, 2016]

[1] Wiegmann et al. Trigger Warning Assignment as a Multi-Label Document Classification Problem. ACL 2023

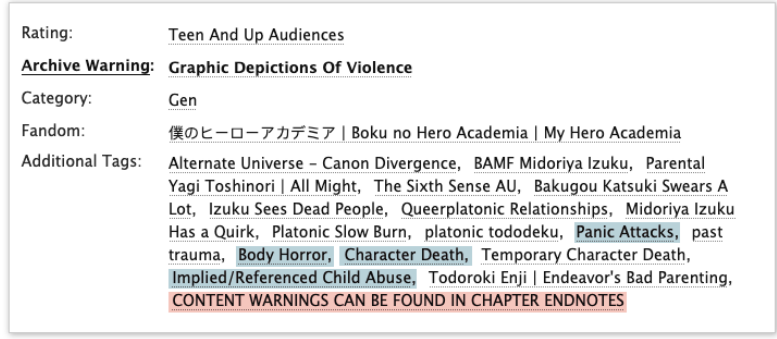
# De-Noising Document Classification Benchmarks

## Dataset

- Fiction documents w/ trigger warnings.[1]
- Labels inferred via weak supervision.

  Via authors' tags, annotations, tag relations, heuristics
- Various sources of label noise:
  - Annotation errors.
  - Author subjectivity.
  - Heuristics fail.
- Author notes may indicate label reliability.

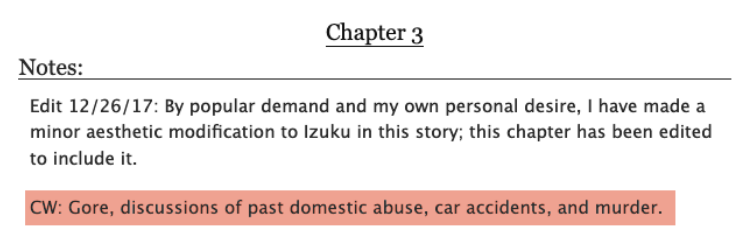
Author tags of a document



| | |
|---|---|
| Rating: | Teen And Up Audiences |
| Archive Warning: | Graphic Depictions Of Violence |
| Category: | Gen |
| Fandom: | 僕のヒーローアカデミア \| Boku no Hero Academia \| My Hero Academia |
| Additional Tags: | Alternate Universe - Canon Divergence, BAMF Midoriya Izuku, Parental Yagi Toshinori \| All Might, The Sixth Sense AU, Bakugou Katsuki Swears A Lot, Izuku Sees Dead People, Queerplatonic Relationships, Midoriya Izuku Has a Quirk, Platonic Slow Burn, platonic tododeku, Panic Attacks, past trauma, Body Horror, Character Death, Temporary Character Death, Implied/Referenced Child Abuse, Todoroki Enji \| Endeavor's Bad Parenting, CONTENT WARNINGS CAN BE FOUND IN CHAPTER ENDNOTES |

[PitViperOfDoom, 2016]

Author notes prepended to a chapter



Chapter 3

Notes:

Edit 12/26/17: By popular demand and my own personal desire, I have made a minor aesthetic modification to Izuku in this story; this chapter has been edited to include it.

CW: Gore, discussions of past domestic abuse, car accidents, and murder.

[PitViperOfDoom, 2016]

[1] * Wiegmann et al. Trigger Warning Assignment as a Multi-Label Document Classification Problem. ACL 2023

# De-Noising Document Classification Benchmarks

## Finding and Pruning Noisy Documents

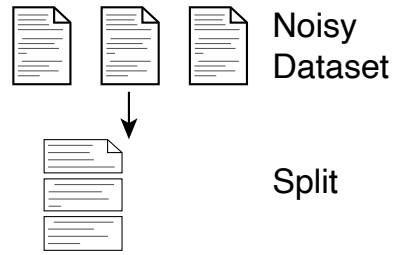
**Idea:** A reliable document contains chunks of text that supports the label *(Signal)*. Remove documents without signal.


1. *Input:* A set of documents w/ finite label set.

# De-Noising Document Classification Benchmarks

## Finding and Pruning Noisy Documents

**Idea:** A reliable document contains chunks of text that supports the label *(Signal)*. Remove documents without signal.



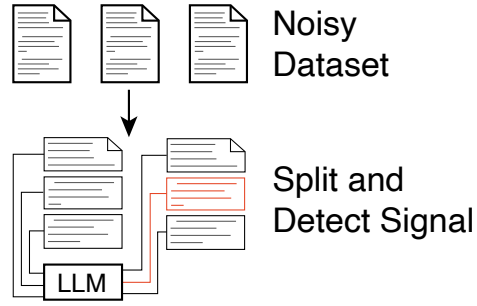

1. *Input:* A set of documents w/ finite label set.
2. Split documents into chunks.

   We use five consecutive sentences as chunks

# De-Noising Document Classification Benchmarks

## Finding and Pruning Noisy Documents

**Idea:** A reliable document contains chunks of text that supports the label *(Signal)*. Remove documents without signal.

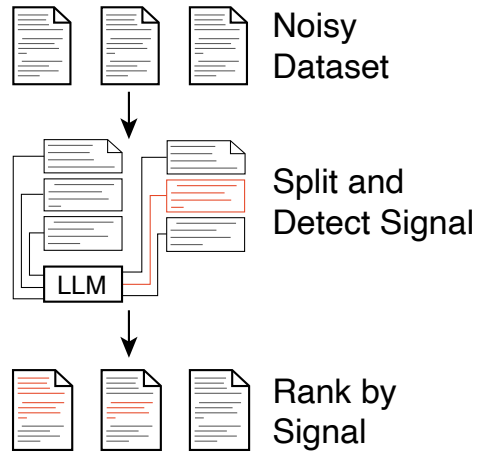



1. *Input:* A set of documents w/ finite label set.
2. Split documents into chunks.
   We use five consecutive sentences as chunks
3. Prompt a LLM to test if a chunk has a signal for its label.

# De-Noising Document Classification Benchmarks

## Finding and Pruning Noisy Documents

**Idea:** A reliable document contains chunks of text that supports the label *(Signal)*. Remove documents without signal.



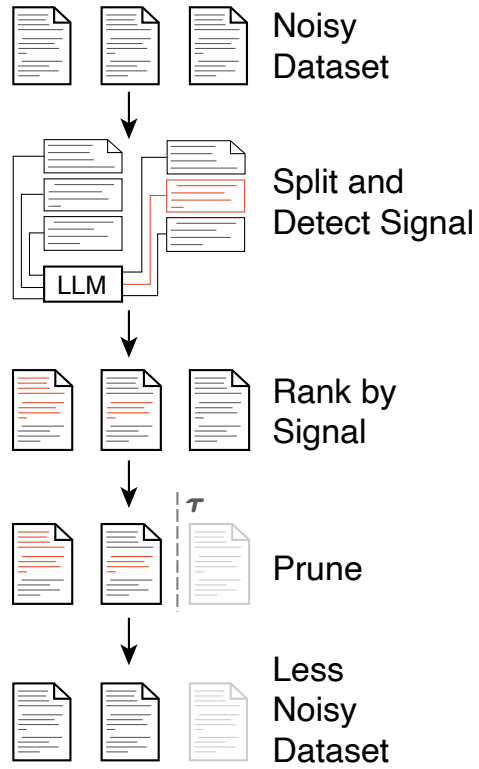

1. *Input:* A set of documents w/ finite label set.
2. Split documents into chunks.
   We use five consecutive sentences as chunks
3. Prompt a LLM to test if a chunk has a signal for its label.
4. Rank the documents descending by signal.
   We use the absolute number of chunks with a signal

# De-Noising Document Classification Benchmarks

## Finding and Pruning Noisy Documents

**Idea:** A reliable document contains chunks of text that supports the label *(Signal)*. Remove documents without signal.

1. *Input:* A set of documents w/ finite label set.
2. Split documents into chunks.
   We use five consecutive sentences as chunks
3. Prompt a LLM to test if a chunk has a signal for its label.
4. Rank the documents descending by signal.
   We use the absolute number of chunks with a signal
5. Prune (noisy) documents with a signal below a threshold $\tau$.



# De-Noising Document Classification Benchmarks

## Experimental Evaluation

1. **Does our de-noising remove noisy labels?**
   - Yes, if the proportion of reliable documents increases
     and/or the proportion of non-reliable documents decreases after pruning.

# De-Noising Document Classification Benchmarks

## Experimental Evaluation

1. **Does our de-noising remove noisy labels?**
   - Yes, if the proportion of reliable documents increases and/or the proportion of non-reliable documents decreases after pruning.

2. **Does our de-noising improve the benchmark?**
   - Yes, if the test scores increase and the relative difference between test scores changes after pruning the test data.

# De-Noising Document Classification Benchmarks

## Experimental Evaluation

### Experimental Dataset

- *Labels:* Death, Violence, Homophobia, Self-harm.
- 1,000 documents per label.

  English documents; 50-10,000 words; no duplicates
- 200 reliable documents.

  Author note has `tw, cw, trigger, content warning` within 20 tokens of a warning term (e.g. `homophobia`)
- 200 documents with synthetic label noise.

  Label was replaced with one of the other three.

Number of documents in corpus

| Warning | All | Reliable |
|---|---|---|
| Death | 124,958 | 1,579 |
| Violence | 119,684 | 1,736 |
| Homophobia | 22,688 | 558 |
| Self-harm | 23,029 | 1,343 |

Lenght of sampled documents

| Warning | Mean | Std |
|---|---|---|
| Death | 3,351 | 2,717 |
| Violence | 4,021 | 2,853 |
| Homophobia | 4,125 | 2,809 |
| Self-harm | 3,478 | 2,688 |

# De-Noising Document Classification Benchmarks

## Finding and Pruning Noisy Documents

### Chunk Classification*

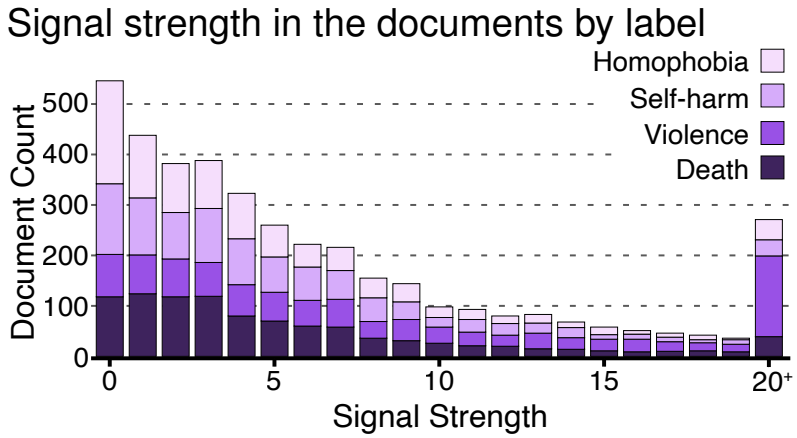
- *Model:* `Mixtral 8x7B`.
- *Prompt:*

  You are a text classification model.

  You determine if a given text contains death, graphic display of death, murder, or dying characters.

  If the text contains intense, explicit, and graphic death, you answer: **Yes.**

  If the text contains mild or implicit death or no death at all, you answer: **No.**




* Wiegmann et al. If there's a Trigger Warning, then where's the Trigger? Investigating Trigger Warnings at the Passage Level. arXiv 2024.

# De-Noising Document Classification Benchmarks

## Experimental Evaluation

### Evaluation

- Models:
  1. RoBERTa.

     Documents truncated to 500 tokens
  2. Longformer.

     Documents truncated to 4,096 tokens
  3. XGBoost.

     Word 1–3-gram; tf $\cdot$ idf vectors; Top 10,000 features selected via $\chi^2$

- Micro-averaged F1 via a 5-fold Monte Carlo cross-validation.

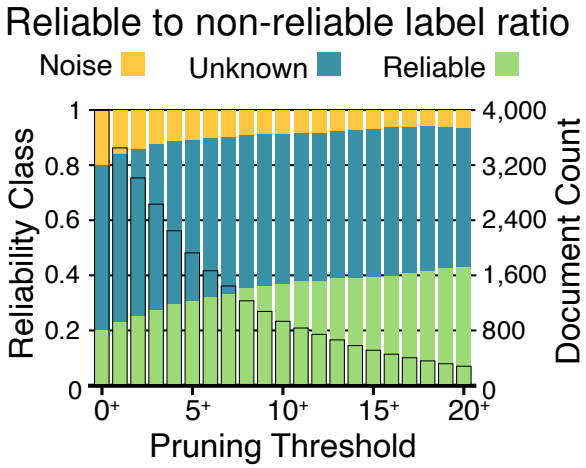
# De-Noising Document Classification Benchmarks

## Results

**Ratio of reliable to non-reliable labels**

- Documents w/ *reliable* labels increase: 0.2 to 0.41.
- Documents w/ *synthetic noise* decrease: 0.2 to 0.05.

⇝ De-noising improves the ratio of reliable-to-noisy labels.



# De-Noising Document Classification Benchmarks

## Results

### Ratio of reliable to non-reliable labels

- Documents w/ *reliable* labels increase: 0.2 to 0.41.
- Documents w/ *synthetic noise* decrease: 0.2 to 0.05.

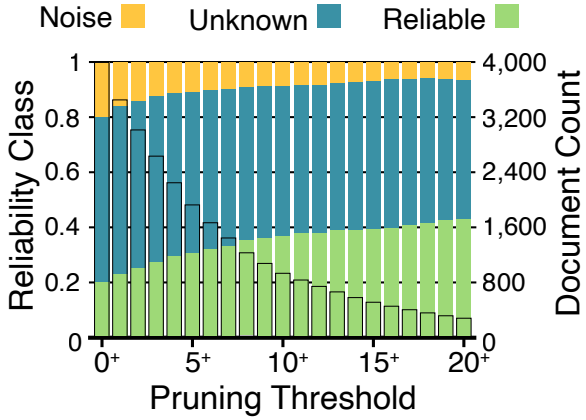
$\rightsquigarrow$ De-noising improves the ratio of reliable-to-noisy labels.

### Model performance and model differences

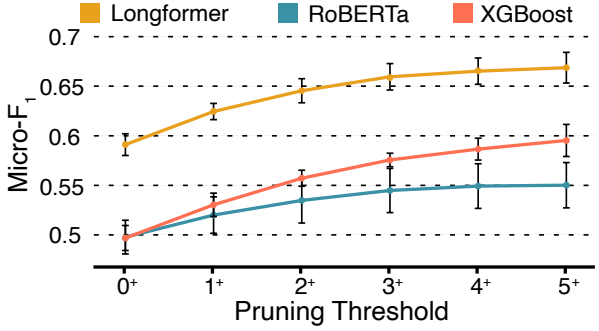
- F1 increases by 0.05–0.1 with $\tau$ = 5+.
  Strongest for XGBoost and weakest for RoBERTa
- XGBoost is significantly better that RoBERTa at $\tau >= 2$.

$\rightsquigarrow$ De-noising can reveal hidden model differences.

Reliable to non-reliable label ratio




$F_1$ for test data pruning



# De-Noising Document Classification Benchmarks

## Summary

- Label noise can deteriorate benchmarks.
- We propose prompt-based rank pruning to remove noisy labels.
- Our method (1) removes noise and (2) reveals hidden model differences.

  One one dataset for three models.

**Data** https://doi.org/10.5281/zenodo.7976807
**Code** https://github.com/webis-de/CLEF-24
**Contact** matti.wiegmann@uni-weimar.de

# Appendix

- Effectiveness when pruning training and test data