# RaFoLa: A Rationale-Annotated Corpus for Detecting Indicators of Forced Labour

**Erick Mendez Guzman, Viktor Schlegel, and Riza-Batista Navarro**

Department of Computer Science, The University of Manchester, Manchester, United Kingdom

## RATIONALES

- **Rationales** (or Language Explanations): Interpretable justifications for a model's prediction.
- **Human Rationales:** Annotations that explain why a human annotator assigned certain classification labels.

## CORPUS & ANNOTATION

**989 news articles (English written) published from January 2019 - September 2021**

**Our goal is to provide richer annotations for training text classification models**, i.e., labels with rationales. When annotating a news article, our annotators also highlight the evidence supporting their annotation, thereby allowing classifiers to learn why the instance belongs to a specific category.

| Forced Labour Indicator | # News Articles | Frequent Words (Articles) | Frequent Words (Rationales) |
|---|---|---|---|
| Abuse of vulnerability | 172 | [*workers, labour, work, rights, forced*] | [*vulnerable, child, children, forced, women*] |
| Abusive working and living conditions | 256 | [*workers, rights, children, human, palm*] | [*conditions, water, living, food, dangerous*] |
| Debt bondage | 72 | [*workers, labour, migrant, trafficking, human*] | [*pay, debt, fees, money, recruitment*] |
| Deception | 51 | [*workers, trafficking, labour, slavery, victims*] | [*promise, job, lured, contracts, recruitment*] |
| Excessive overtime | 117 | [*workers, labour, palm, oil, children*] | [*hours, day, work, week, plantation*] |
| Intimidation and threats | 67 | [*workers, women, labour, forced, rights*] | [*threats, retaliation, refused, reported, bosses*] |
| Isolation | 47 | [*palm, oil, workers, children, plantations*] | [*plantations, remote, phone, guarded, hills*] |
| Physical and sexual violence | 123 | [*workers, labour, children, women, forced*] | [*abuse, sexual, harassment, violence, beaten*] |
| Restriction of movement | 34 | [*workers, labour, forced, trafficking, conditions*] | [*locked, factory, guard, armed, escaping*] |
| Retention of identity documents | 31 | [*workers, labour, trafficking, force, human*] | [*passport, documents, taken, confiscated, migrant*] |
| Withholding of wages | 47 | [*workers, labour, rights, people, human*] | [*wages, pay, unpaid, withheld, money*] |

Table 1. Number of news articles and most frequently occurring words for each forced labour indicator

## BASELINE CLASSIFIERS

We decided to apply a simple random **under-sampling method over the training and validation sets**:

- **Dataset 1:** The whole corpus, including the news articles without any assigned labels (**n=989**).
- **Dataset 2:** We removed half of the news articles without any assigned labels (**n=763** which were randomly selected).
- **Dataset 3:** We kept only news articles with at least one label assigned (**n=538**).

| Model | Dataset | F1(weighted) | LRAP | EMR |
|---|---|---|---|---|
| roberta-base | Data set 1 | 0.47 | 0.86 | 0.49 |
| | Data set 2 | 0.45 | 0.87 | 0.42 |
| | Data set 3 | 0.40 | 0.88 | 0.05 |
| distilroberta-base | Data set 1 | 0.49 | 0.85 | 0.50 |
| | Data set 2 | 0.50 | 0.88 | 0.43 |
| | Data set 3 | 0.43 | 0.89 | 0.09 |
| distilbert-base | Data set 1 | 0.49 | 0.86 | 0.48 |
| | Data set 2 | 0.44 | 0.88 | 0.25 |
| | Data set 3 | 0.36 | 0.88 | 0.06 |
| xlnet-base | Data set 1 | 0.51 | 0.87 | 0.51 |
| | Data set 2 | 0.44 | 0.86 | 0.43 |
| | Data set 3 | 0.38 | 0.87 | 0.06 |
| albert-base | Data set 1 | 0.47 | 0.86 | 0.44 |
| | Data set 2 | 0.47 | 0.88 | 0.34 |
| | Data set 3 | 0.35 | 0.87 | 0.04 |
| roberta-large | Data set 1 | 0.47 | 0.86 | 0.55 |
| | Data set 2 | 0.46 | 0.87 | 0.49 |
| | Data set 3 | 0.39 | 0.88 | 0.13 |

Table 2. Results on the test subsets of our three data set

- **XLNet records the highest micro, macro, and weighted F1 scores** with 0.52, 0.47, and 0.51, respectively (Dataset 1).
- Almost all models, except for the distilroberta-base, **worsened their F1 scores compared to their results on Data set 1**.
- There is a clear trend of **decreasing EMR scores when removing examples without labels**.
- **roberta-large does not outperform significantly smaller versions of the same architecture**, namely roberta-base and distilroberta-base.

## CONTRIBUTIONS

- We design a **rationale-oriented annotation scheme** for capturing indicators of forced labour.
- To the best of our knowledge, we present the **first resource consisting of news articles annotated for indicators of forced labour**, and their respective human-generated rationales.
- We provide **results of multi-class and multi-label baseline models** to predict such indicators.

## FUTURE WORK

**Evaluate the impact of including human rationales as extra supervision information on model performance and explainability** for a multi-class/multi-label text classifier to detect indicators of forced labour.

**SIGNATURE**