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.

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).

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 Dataset 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.