Towards Explainable Multi-Label Text Classification: A Multi-Task Rationalisation Framework for Identifying Indicators of Forced Labour

Erick Mendez Guzman[1] and **Viktor Schlegel**[1,2] and **Riza Batista-Navarro**[1]
[1]The University of Manchester, United Kingdom
[2]Imperial College London, Imperial Global Singapore
erick.mendezguzman@manchester.ac.uk, v.schlegel@imperial.ac.uk
riza.batista@manchester.ac.uk

Abstract

The importance of *rationales*, or natural language explanations, lies in their capacity to bridge the gap between machine predictions and human understanding, by providing human-readable insights into why a text classifier makes specific decisions. This paper presents a novel multi-task rationalisation approach tailored to enhancing the explainability of multi-label text classifiers to identify indicators of forced labour. Our framework integrates a rationale extraction task with the classification objective and allows the inclusion of human explanations during training. We conduct extensive experiments using transformer-based models on a dataset consisting of 2,800 news articles, each annotated with labels and human-generated explanations. Our findings reveal a statistically significant difference between the best-performing architecture leveraging human rationales during training and variants using only labels. Specifically, the supervised model demonstrates a 10% improvement in predictive performance measured by the weighted F1 score, a 15% increase in the agreement between human and machine-generated rationales, and a 4% improvement in the generated rationales' comprehensiveness. These results hold promising implications for addressing complex human rights issues with greater transparency and accountability using advanced NLP techniques.

1 Introduction

Multi-label text classification is a fundamental task in Natural Language Processing (NLP) with wide-ranging applications, including document categorisation, sentiment analysis and content recommendation (Kowsari et al., 2019). Even though deep learning models have achieved state-of-the-art performance for text classification in the last two decades, their black-box nature and the lack of understanding of why they assign specific labels to a text limits their application scope in high-stake domains (Liu et al., 2017; Zini and Awad, 2022).

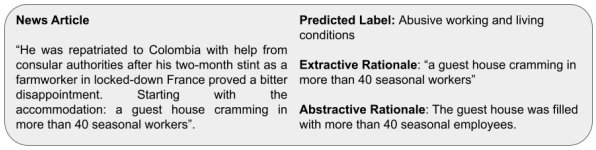

Figure 1: Example of an extractive and abstractive rationale supporting the identification of a forced labour indicator for a news article.

Rationalisation models attempt to explain the outcome of a text classification model by providing a natural language explanation (*rationale*) (Lei et al., 2016). It has been observed that rationales are more understandable and easier to use than other explainability methods since they are verbalised in human-comprehensible natural language (DeYoung et al., 2019; Wang and Dou, 2022). Recent evidence suggests that generating human-readable justifications for a model's predictions could empower users to grasp the reasoning behind a classifier's decisions, facilitating trust, accountability and the development of user-centric applications (Kandul et al., 2023; Zhao et al., 2023).

Rationales for explainable text classification can be categorised into *extractive* and *abstractive* rationales (Figure 1). Extractive rationales are a subset of the input text that supports a model's predictions, while abstractive rationales are explanations that are not constrained to be grounded in the input text (DeYoung et al., 2019; Liu et al., 2018).

Previous research has established that a multi-task learning approach in training a rationalisation model can enhance the model's accuracy and generate more coherent and relevant explanations (Lei et al., 2016, 2017). When a rationalisation model is trained to classify items and explain its predictions, it learns to perform both tasks simultaneously, leveraging shared information to improve its predictions and rationales (Yu et al., 2019). Recent evidence suggests that domain experts can play a pivotal role in this process by providing concise textual snippets (*human rationales*) that encapsu-

Proceedings of the Third Workshop on NLP for Positive Impact, pages 98–112
November 15, 2024 ©2024 Association for Computational Linguistics

late the reasons behind each classification decision (Wang and Dou, 2022; Kandul et al., 2023).

We propose a novel rationalisation framework to explain the outcome of a multi-label text classifier through extractive rationalisation. Our framework uses multi-task learning to independently produce rationales at a label level and allows the alternative of including human rationales during training as an extra supervision signal. We employ our framework to identify indicators of forced labour, as defined by the International Labour Organization (ILO, 2012), for a rationale-annotated corpus of news articles (Mendez Guzman et al., 2022). We hope our framework can help researchers and practitioners (e.g., social scientists or policymakers) in using supervised learning models to detect modern slavery with a more systematic approach. In summary, the main contributions of this paper are: (*i*) We present a novel rationalisation framework to generate extractive rationales at a label level in a multi-label setting using a multi-task learning approach and including human explanations during training; (*ii*) We provide results demonstrating that including human explanations during training can boost predictive performance and explainability of our rationalisation model for identifying indicators of forced labour.

2 Related Work

Most research on extractive rationalisation has been carried out using an *encoder-decoder*[1] architecture (Lei et al., 2016; Arous et al., 2021). The encoder $enc(x)$ serves as a tagging model, where each word in the input sequence x receives a binary tag indicating whether it is included in the rationale z. The decoder $dec(x, z)$ then uses only the rationales and maps them to the target vector (Paranjape et al., 2020).

Lei et al. (2016) pioneered the idea of using a multi-task learning approach modelling rationales as binary latent variables. They proposed jointly training the encoder and decoder to minimise a cost function composed of the classification loss and sparsity-inducing regularisation to keep the rationales short and coherent. Considering that minimising the expected cost is challenging since it involves summing up all possible choices of rationales in the input sequence, they suggested training the architecture using REINFORCE-based optimisation (Williams, 1992). REINFORCE works by sampling rationales from the encoder and training the model to generate explanations using reinforcement learning. As a result, the model is rewarded for producing rationales that align with the desiderata defined in the cost function (Zhang et al., 2021b).

Using this multi-task learning approach, researchers have studied extractive rationalisation methods for binary and multi-class text classification (Wang and Dou, 2022). While some authors have kept using the latent binary variables and sparsity-inducing regularisation to encourage the model to select a limited set of words as rationales while suppressing irrelevant information (Zhao and Vydiswaran, 2020; Paranjape et al., 2020), others have decided to transform the binary latent variables into continuous and differentiable variables. Reparametrisation enables smoother optimisation without using the REINFORCE algorithm and allows for fine-tuning the length of rationales (Bastings et al., 2019; Madani and Minervini, 2023).

Even though research on *learning with rationales* has established over the last fifteen years that incorporating human rationales during training can improve classification performance, it is only in the past four years that studies have started looking into using human rationales to enhance the quality of the generated explanations (Hartmann and Sontag, 2022). Researchers have adapted the original implementation by Lei et al. (2016), incorporating human rationales during training by modifying the model's cost function by adding components to force the generated rationales to be similar to the human explanations (DeYoung et al., 2019; Strout et al., 2019; Arous et al., 2021).

Our rationalisation approach draws inspiration from the work of Lei et al. (2016) and Bastings et al. (2019) around using multi-task learning to enhance predictive performance and explainability when training the encoder-decoder architecture. Following work by DeYoung et al. (2019) and Arous et al. (2021), we also explore using human explanations during training as an extra supervision signal and check whether it significantly impacts the results of our framework. However, our work extends theirs by focussing on independently producing rationales tailored to each predicted label using pre-trained language models.

[1]Not to be confused with encoder-decoder transformer architectures, such as the Text-to-Text Transfer Transformer (T5).

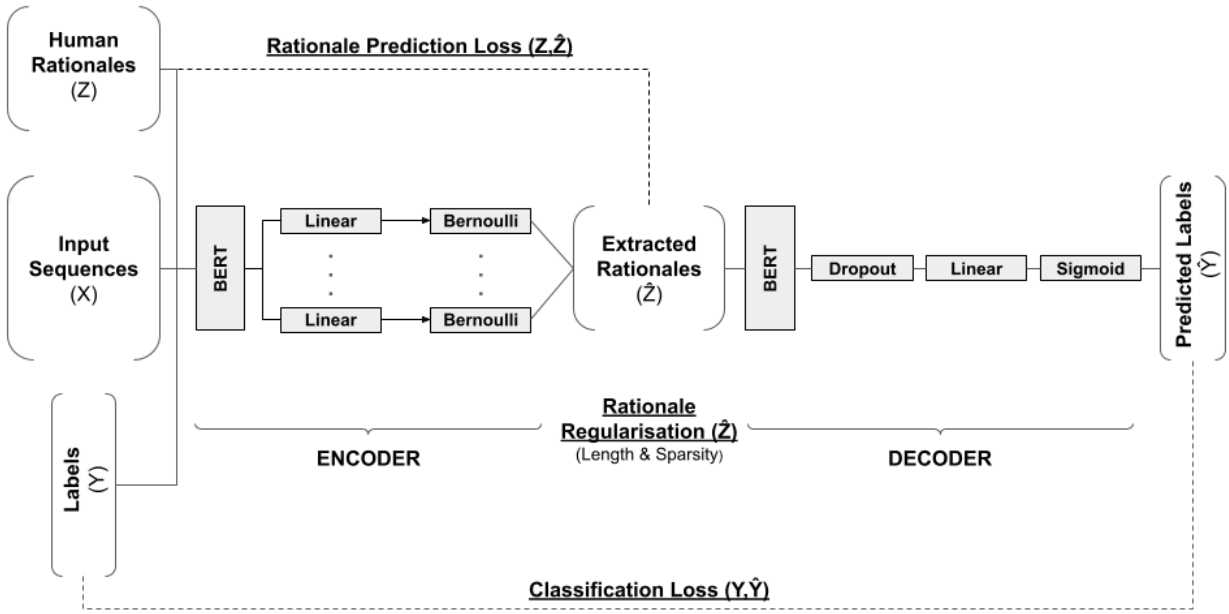

Figure 2: Framework for Explainable Multi-Label Text Classification through Multi-Task Extractive Rationalisation. The encoder processes the input sequences (X) to extract the rationales ($\hat{Z}$) at a label level. The rationales are then input to the decoder to predict the target labels ($\hat{Y}$). The encoder and decoder are trained jointly via REINFORCE-based optimisation using a loss function composed by the Classification Loss ($Y, \hat{Y}$), a Rationale Regularisation ($\hat{Z}$), and an additional Rationale Prediction Loss ($Z, \hat{Z}$) in the case of the supervised rationale extraction variant.

3 Explainable Text Classification Framework

In this section, we detail our framework for explainable text classification based on a multi-task learning implementation of the *encoder-decoder* architecture (Lei et al., 2016) to produce rationales at a label level for a multi-label setting. The encoder is the module responsible for identifying the rationales within the input sequence at a label level, and the decoder is tasked with predicting labels based on the generated rationales (Bastings et al., 2019; Madani and Minervini, 2023).

It is important to note that our framework allows human rationales to be included as an extra supervision signal during training. Throughout our paper, we refer to the architecture using target labels and human rationales during training as 'supervised rationale extraction' and refer to the implementation using only target labels as 'unsupervised rationale extraction'. Figure 2 describes our framework in terms of its input data, encoder, decoder and loss function.

Input Data The input data for our framework is composed of input sequences (X), target labels (Y) and, optionally, human rationales (Z). Target labels are encoded as C-dimensional vectors using one-hot-encoding (Zhang and Zhou, 2013), where C is the number of classes. As mentioned before, the human rationales are snippets of the input sequence that support labelling decisions at the label level. The human rationales for each input sequence are subsequently post-processed and represented in a $C \times L$ matrix format, where L is the maximum sequence length associated with the tokenisation applied over the input sequence (Arous et al., 2021). Each row corresponds to a rationale for a specific label, and it is filled with binary tags that indicate whether the token was selected to be part of the human explanation or not. We refer to Appendix A for a detailed input data example.

Encoder Drawing inspiration from the encoder-decoder architecture proposed by Lei et al. (2016) and DeYoung et al. (2019), we employ a pre-trained language model such as BERT (Devlin et al., 2018), to induce contextualised representation of tokens. The encoder generates a scalar, denoting the probability of selecting that token as part of the rationale, for each BERT hidden state using a set of C linear and Bernoulli layers (Shapiro and Zahedi, 1990). It is important to note that each linear and Bernoulli layer works independently to produce rationales at a label level for our multi-label classifier.

Decoder As a decoder, we use a second pre-trained language model followed by a classifica-

tion layer. The classification layer comprises a dropout (Srivastava et al., 2014), a linear (Svozil et al., 1997) and a sigmoid layers (Menon et al., 1996). In our implementation, the decoder accepts the tokens in the input sequence tagged as rationales by the encoder and independently predicts each label.

Optimisation The encoder and decoder are trained jointly via REINFORCE-based optimisation (Williams, 1992), for which we assume it is possible to efficiently sample rationales from the encoder (Lei et al., 2016; Arous et al., 2021). REINFORCE allows us to extract rationales using reinforcement learning, where our model is rewarded for producing explanations that align with the desiderata defined in our loss function (Zhang et al., 2021b). One of the advantages of this approach is that it is flexible enough to allow us to experiment with variants of the architecture in which we train it with and without rationale-level annotations.

Loss Function For the unsupervised rationale extraction variant, the loss function is a composite of the classification loss and a regulariser over the rationale selection. Following Lei et al. (2016), we guide the encoder to extract short and coherent explanations by penalising the number of words in rationales and discouraging transitions. In this way, the encoder should select only a few words, and those rationales should form phrases rather than isolated and disconnected words (Bastings et al., 2019; Arous et al., 2021). In the variant using human rationales during training, we incorporate an additional component made by the cross-entropy loss over rationale predictions (Strout et al., 2019; DeYoung et al., 2019). We refer to Appendix B for the mathematical formulation of the loss function.

Evaluation The goal of our rationalisation framework is to simultaneously enhance predictive performance and explainability by identifying concise and relevant rationales. We evaluate the outcome of our architecture from various perspectives, attempting to assess the extent to which it meets the expectations of different stakeholders, especially end-users and developers (Doshi-Velez and Kim, 2017; Carton et al., 2020).

We utilise a set of widely used metrics for multi-label classification to evaluate the predictive performance. Even though our primary metric will be the weighted F1 score as it considers the class imbalance in our corpus (Feldman et al., 2007), we also calculate the Label Ranking Average Precision (LRAP) (Ghamrawi and McCallum, 2005) and the Exact Match Ratio (EMR) (Feldman et al., 2007). While LRAP assesses the classifier's ranking performance by quantifying how well it orders the labels in terms of relevance, EMR evaluates the classifier's precision in predicting all labels correctly for a given instance. These metrics offer a robust evaluation framework addressing precision, ranking and overall label prediction accuracy.

To assess the quality of the machine-generated rationales, we measure their plausibility and faithfulness. Plausibility reflects whether the rationales make sense to domain experts or end users, while faithfulness assesses the alignment between the rationales and the model's actual decision-making process (Mohseni et al., 2018; Lertvittayakumjorn and Toni, 2019; Carton et al., 2020). This dual evaluation ensures that the explanations are human-understandable and faithfully represent the model's reasoning, enhancing their overall utility and trustworthiness (Doshi-Velez and Kim, 2017; Hase and Bansal, 2020).

Since measuring exact matches between human rationales (z_{ij}) and machine-generated explanations ($\hat{z}_{ij}$) for the same input sequence i and class j is likely too harsh, we evaluate plausibility using the Intersection-over-Union (IoU) at the token level as it is a more relaxed measure to compare two text sequences (DeYoung et al., 2019):

$$\text{IoU}(z_{ij}, \hat{z}_{ij}) = \frac{|z_{ij} \cap \hat{z}_{ij}|}{|z_{ij} \cup \hat{z}_{ij}|} \quad (1)$$

We count an extracted rationale as a match if it overlaps with the human rationale by more than some threshold (0.5 in our case):

$$\text{match} = \begin{cases} 1 & \text{if IoU}(z_{ij}, \hat{z}_{ij}) > \text{threshold} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Finally, we use these matches to derive an F1 score at the label level and weight them according to the number of items on each class to calculate a weighted average (DeYoung et al., 2019; Paranjape et al., 2020; Chan et al., 2021).

For measuring faithfulness, we calculate sufficiency and comprehensiveness as defined by DeYoung et al. (2019), using $m(x_i)_j$ as the original prediction for the item i provided by a model m for the predicted class j. Similarly, $m(z_{ij})_j$ and $m(x_i \backslash z_{ij})_j$ are the predicted probability for the same class using only the rationales, and using the

input sequence once the rationales were removed, respectively.

Sufficiency (Suff) assesses whether the snippets in the rationales are adequate to make a prediction (Equation 3).

$$\text{Suff} = 1 - \max(0, m(x_i)_j - m(z_{ij})_j) \quad (3)$$

Comprehensiveness (Comp) captures the degree to which all relevant features to make a prediction were selected as rationales (Equation 4).

$$\text{Comp} = \max(0, m(x_i)_j - m(x_i \backslash z_{ij})_j) \quad (4)$$

Sufficiency and comprehensiveness scores go from zero to one, with one being the best score possible. Following these definitions, a faithful rationale should have high sufficiency and comprehensiveness (Zhang et al., 2021a; Chan et al., 2021). All the metrics cited above are calculated at a label level and then aggregated into a weighted average to account for the class imbalance in our dataset.

4 Experimental Design

We conduct experiments using various pre-trained language models to compare and contrast the results of the unsupervised and supervised rationale extraction architectures on our dataset.[2]

4.1 Dataset

Forced labour refers to situations in which individuals are coerced to work against their will through the use of violence, intimidation, or other forms of exploitation (ILO, 1930). According to figures from the International Labour Organization (ILO) and Walk Free, an estimated 27.6 million people worldwide were victims of forced labour in 2022 across various industries, including agriculture, construction, and domestic work (Free et al., 2022).

The dataset utilised in this study is the RaFoLa dataset v.2.0 curated by Mendez et al. (2022) to promote research on explainability and released under the Creative Commons Attribution-NonCommercial 4.0 International License (CC-BY-NC-4.0)[3]. The second release of the RaFoLa dataset comprises a collection of 2,800 news articles retrieved from specialised data sources, such as the Traffik Analysis Hub (TAH, 2012), and annotated by researchers and domain experts to identify indicators of forced labour. Each news article is annotated in a multi-label text classification manner based on the eleven indicators of forced labour defined by ILO (2012). Additionally, the annotators have selected phrases and sentences to support their labelling decisions at a label level. These snippets extracted from the original text are the human rationales used for training our supervised rationalisation architecture and evaluating the plausibility of the generated rationales. For detailed information about the dataset's label distribution, we refer the reader to Appendix C.

4.2 Training

Since there is a relatively small body of literature on using state-of-the-art NLP methods in the humanitarian domain, we decided to explore a set of BERT variations for our explainable framework, considering the trade-off between performance and computational cost (Bliss et al., 2021).

Based on work of Mendez et al. (2022) on text classification to identify forced labour, we utilised the following transformer-based models available on Hugging Face (Wolf et al., 2019):

- **DistilBERT** (Sanh et al., 2019): A compressed and smaller version of BERT leveraging knowledge distillation during the training phase.
- **ALBERT** (Lan et al., 2019): A light version of BERT that introduces parameter-sharing strategies to reduce the model's size.
- **RoBERTa** (Liu et al., 2019): An optimised variant of BERT, achieved by fine-tuning training techniques and leveraging a larger corpus. We use the 'base', 'distil-roberta' and 'large' versions for this model.
- **XLNet** (Yang et al., 2019): A generalised autoregressive pretraining method incorporating a permutation-based training approach, enabling it to capture bidirectional context.
- **DeBERTa** (He et al., 2020): A variant of the BERT model that introduces disentangled attention mechanisms and performs dynamic weight adaptation.

DistilBERT's efficiency is advantageous for rapid experimentation, while ALBERT's

[2]The code will be made publicly available upon paper acceptance.

[3]https://creativecommons.org/licenses/by-nc/4.0/legalcode

	Model	F1 (B)	F1	P	S	C
Unsupervised	distilroberta-base	0.48 ± 0.03	0.55 ± 0.02	0.17 ± 0.01	0.96 ± 0.03	0.30 ± 0.02
	roberta-base	0.48 ± 0.02	0.54 ± 0.01	0.13 ± 0.01	0.90 ± 0.03	0.29 ± 0.02
	distilbert-base	0.50 ± 0.03	0.53 ± 0.04	0.12 ± 0.01	0.94 ± 0.02	0.27 ± 0.01
	xlnet-base	**0.53 ± 0.02**	**0.58 ± 0.02**	**0.19 ± 0.00**	**0.92 ± 0.02**	**0.32 ± 0.01**
	albert-base	0.48 ± 0.01	0.51 ± 0.02	0.15 ± 0.01	0.94 ± 0.02	0.26 ± 0.01
	roberta-large	0.47 ± 0.04	0.55 ± 0.04	0.11 ± 0.00	0.91 ± 0.02	0.26 ± 0.01
	deberta-base	0.52 ± 0.03	0.57 ± 0.03	0.18 ± 0.01	0.91 ± 0.03	0.31 ± 0.02
Supervised	distilroberta-base	0.48 ± 0.03	0.57 ± 0.03	0.19 ± 0.01	0.94 ± 0.02	0.34 ± 0.02
	roberta-base	0.48 ± 0.02	0.56 ± 0.03	0.14 ± 0.01	0.91 ± 0.02	0.29 ± 0.01
	distilbert-base	0.50 ± 0.03	0.56 ± 0.03	0.13 ± 0.01	0.96 ± 0.03	0.28 ± 0.02
	xlnet-base	**0.53 ± 0.02**	**0.64 ± 0.04**	**0.22 ± 0.02**	**0.92 ± 0.03**	**0.36 ± 0.02**
	albert-base	0.48 ± 0.01	0.57 ± 0.05	0.16 ± 0.01	0.95 ± 0.02	0.27 ± 0.02
	roberta-large	0.47 ± 0.04	0.57 ± 0.04	0.11 ± 0.00	0.89 ± 0.03	0.28 ± 0.01
	deberta-base	0.52 ± 0.03	0.62 ± 0.04	0.20 ± 0.01	0.90 ± 0.02	0.31 ± 0.01

Table 1: Cross-validation results for the unsupervised and supervised architectures **F1 (B)**: Baseline weighted F1 Score using the whole input sequence **F1**: Weighted F1 Score **P**: Plausibility. **S**: Sufficiency **C**: Comprehensiveness

parameter-reduction techniques allow us to reduce model size without sacrificing its predictive performance (Sanh et al., 2019; Lan et al., 2019). RoBERTa's robustness, XLNet's bidirectional context capture, and DeBERTa's attention mechanisms all provide a versatile toolkit for improving our rationalisation framework's performance and explainability capabilities (Liu et al., 2019; Yang et al., 2019; He et al., 2020).

We leverage the power of the EGG toolkit to implement our multi-task learning rationalisation approach for explainable text classification (Kharitonov et al., 2021). EGG is a Pytorch-based (Paszke et al., 2019) toolkit that allows researchers to implement multi-agent games, where agents are trained to communicate and jointly solve a task. EGG's flexible and user-friendly APIs allowed us to train our architecture using the transformer-based models listed above with REINFORCE-based optimisation. Moreover, it is essential to note that EGG includes an easy-to-adapt boilerplate code to include human rationales during training with minimal changes in the implementation logic.

We split the RaFoLa dataset (v.2.0) into training, validation and test sets according to a 70:10:20 ratio using stratified sampling (Neyman, 1992) and search for the hyperparameter values that minimise the corresponding loss function over the validation set for the unsupervised and supervised variants of our rationalisation architecture. To optimise the training process, we tuned the architecture hyperparameters using a random search method (Bergstra and Bengio, 2012) and ran ten training runs, one for each combination of hyperparameters. Each trial was fine-tuned for twenty-five epochs on the training set. For a detailed description of the hyperparameter tuning process and its results, we refer the reader to Appendix D.

Finally, we merged the training and validation sets in preparation for evaluating the architectures. We utilised k-fold validation (k=5) (Anguita et al., 2012), where each fold was trained for a hundred epochs using the hyperparameters selected by the search method described above. Finally, we used t-test (Student, 1908) and ANOVA (Girden, 1992) analysis to determine if there are statistically significant differences among the different variants of our architecture.

To ensure consistency and comparability of our results, all our models were trained and evaluated on a Google Colab (Bisong and Bisong, 2019) runtime equipped with an NVIDIA A100 GPU with 40 GB of memory.

5 Results and Discussion

Table 1 shows the results obtained for each unsupervised and supervised rationalisation architecture in the cross-validation test sets in terms of each metric's mean and standard deviation.

Results from the ANOVAs, performed separately for each performance and explainability indicator using a significance level of 0.05, suggest a statistically significant difference in the architectures' performance in all metrics.

We replicate the experiments described by

Label	F1	P	S	C
Abuse of vulnerability	0.41 ± 0.04	0.06 ± 0.02	0.92 ± 0.03	0.09 ± 0.02
Abusive working and living conditions	0.79 ± 0.04	0.30 ± 0.03	0.94 ± 0.02	0.25 ± 0.04
Debt bondage	0.61 ± 0.03	0.48 ± 0.02	0.92 ± 0.02	0.55 ± 0.07
Deception	0.56 ± 0.04	0.13 ± 0.08	0.85 ± 0.05	0.31 ± 0.03
Excessive overtime	0.67 ± 0.05	0.16 ± 0.07	0.90 ± 0.04	0.63 ± 0.02
Intimidation and threats	0.26 ± 0.07	0.09 ± 0.03	0.89 ± 0.04	0.53 ± 0.08
Isolation	0.64 ± 0.06	0.06 ± 0.02	0.92 ± 0.03	0.24 ± 0.02
Physical and sexual violence	0.42 ± 0.10	0.35 ± 0.02	0.95 ± 0.01	0.57 ± 0.03
Restriction of movement	0.90 ± 0.03	0.14 ± 0.03	0.73 ± 0.09	0.50 ± 0.03
Retention of identity documents	0.29 ± 0.11	0.04 ± 0.01	0.68 ± 0.08	0.53 ± 0.02
Withholding of wages	0.58 ± 0.07	0.08 ± 0.02	0.77 ± 0.09	0.43 ± 0.03

Table 2: Cross-validation results for the best-performing supervised architecture (XLNet) **F1**: Weighted F1 Score **P**: Plausibility. **S**: Sufficiency **C**: Comprehensiveness

Mendez Guzman et al. (2022) using the RaFoLa dataset (v.2.0) and use the weighted F1 score as a baseline for the predictive performance of our framework. From this data, it can be seen that there is an increase in the average predictive performance when comparing our unsupervised rationalisation architecture with transformer-based classifiers using the whole input sequence as an input. For LRAP and EMR scores for each architecture, we refer the reader to Appendix E.

What stands out in Table 1 are the high sufficiency scores for all architectures, regardless of whether they exploit human rationales during training and the transformer model they are based on. The sufficiency scores of 0.9 or above on average, indicate that the generated rationales provide enough information to justify the classification outcome (DeYoung et al., 2019; Bastings et al., 2019; Paranjape et al., 2020). However, there is room for improvement regarding the plausibility and comprehensiveness of the machine-generated explanations. Low plausibility signifies that the extracted rationales differ from the snippets the domain experts picked, potentially undermining the model's trust in a real-world setting (Strout et al., 2019; Arous et al., 2021). Additionally, low comprehensiveness suggests that the rationales fail to encompass the essential information, including more information than necessary, potentially reducing the explanation's effectiveness (Doshi-Velez and Kim, 2017; Carton et al., 2020).

The implementation based on the XLNet model performed the best among the unsupervised architectures in all metrics except for the sufficiency. Data from previous research suggests that permutation-based training of this model, which captures bidirectional context efficiently, might enhance our architecture's predictive performance and rationales' quality (Mendez Guzman et al., 2022; Kashapov et al., 2022).

Regarding the supervised rationalisation models, data in Table 1 shows that incorporating human rationales during training enhances the model's performance and explainability. The ANOVA analysis revealed a significant difference between the unsupervised and supervised rationalisation architectures in the F1 score, plausibility and comprehensiveness scores. Even though the difference in the sufficiency scores between the two variants was not statistically significant, results of the supervised architectures are still around 0.9 on average. We refer the reader to Appendix E for detailed results regarding LRAP and EMR scores for the supervised rationalisation architectures.

Similar to the results for the unsupervised rationalisation models, the architecture based on the XLNet model performed best among the supervised variants. Results show a 10%, 15%, and 4% improvement in F1 score, plausibility and comprehensiveness, compared to the unsupervised architecture based on the same model. It is worth noting that these results are significant at a $p = 0.05$ level.

Table 2 presents results at the forced labour indicator level for the supervised rationalisation architecture based on the XLNet model. A closer inspection of the table shows the disparity in the results among indicators of forced labour. While there are labels such as 'Debt bondage' and 'Physical and sexual violence' where the model performs significantly better than the overall results, there are also indicators, namely 'Retention of identity documents' and 'Withholding of wages', where

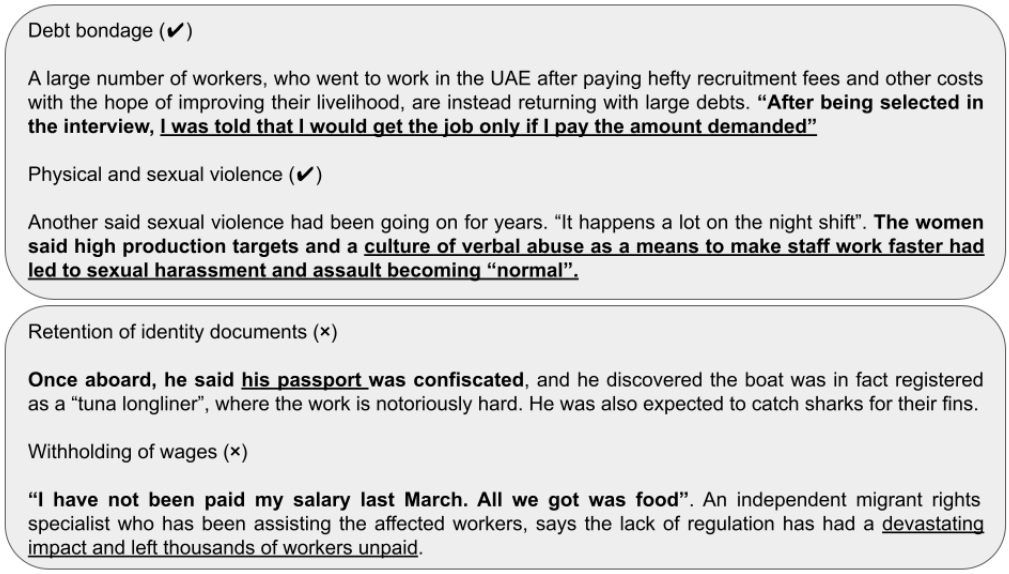

Figure 3: Examples of "good" (✓) and "bad" (×) rationales extracted using the supervised rationalisation architecture based on the XLNet model. Human rationales are depicted in bold for each example, while the machine-generated rationales are underlined.

the model is not able to identify nor explain them adequately.

Figure 3 presents examples of some "good" and "bad" rationales extracted using the supervised rationalisation architecture based on the XLNet model. On average, machine-generated rationales are 37% shorter than human rationales. "Good" rationales are often a subset of the human rationales containing the most relevant information regarding the predicted forced labour indicator. In contrast, "bad" machine-generated rationales are either too short or have no intersection with the human-provided explanations. In both cases, the IoU scores are very low, directly affecting the generated rationales' plausibility (DeYoung et al., 2019; Carton et al., 2020).

The results of our rationalisation approach may vary among labels due to differences in the number of news articles per indicator and the intrinsic complexity associated with each label (Ghamrawi and McCallum, 2005; Lertvittayakumjorn and Toni, 2019; Carton et al., 2020). Labels with fewer examples (e.g., "Retention of identity documents") or inherently complex criteria (e.g., "Intimidation and threats") may exhibit more significant variability in rationalisation performance compared to labels with readily available training data (e.g., "Abusive working and living conditions") and more distinctive language characteristics (e.g., "Debt bondage") (ILO, 2012; Mendez et al., 2022).

We observe that our rationalisation approach exploiting rationale-level supervision often improves the predictive performance and rationales' quality, as in prior work (Zhang et al., 2016; Strout et al., 2019; Arous et al., 2021). Nevertheless, there is a disparity in the results among labels where rationales for less-represented forced labour indicators tend to have low predictive performance and shallow agreement with human-provided rationales.

6 Conclusions

Forced labour is the most common type of modern slavery, affecting an estimated 27.6 million people worldwide. Explainable text classification can aid stakeholders, such as NGOs, police forces, and policy-makers, in understanding, addressing, and preventing the spread of forced labour by empowering them with actionable insights (Tambe and Tambay, 2020; Weinberg et al., 2020). In this work, we presented a novel multi-task rationalisation framework to extract rationales at a label level in a multi-label setting that allows the inclusion of human explanations during training. Our experiments showed that using human rationales as an extra supervision signal can improve the classification performance of our model while enhancing the quality of the generated explanations. Whilst the small number of cases limits the results of our framework in some of the forced labour indicators, it offers valuable insights into cases of "Abusive working and living conditions" and "Restriction of movement", among other indicators. In the future, we will focus on improving the framework's performance for less-represented labels.

7 Limitations

We attempted to develop a novel framework for explainable multi-label text classification in a multi-task setting using human explanations as additional supervision signals during training. However, our approach is subject to certain limitations, as enumerated below: (*i*) Since our rationalisation approach uses human rationales during training, there is the potential for limited or biased annotations which may not cover the full range of possible rationales. One could consider employing data augmentation techniques to create additional diverse rationales or incorporating more expert feedback to enhance the diversity and representativeness of the training dataset. (*ii*) A limitation of evaluating the machine-generated rationales using only plausibility and faithfulness metrics is that these may not fully capture the utility of the explanations for end-users. One could incorporate additional user–centric evaluation metrics obtained through user studies or feedback to provide a more comprehensive assessment of rationale quality from the user's perspective. (*iii*) The proposed methodology has been validated on an English-based dataset. Further research would be required to scale up to other languages prevalent in regions and countries where forced labour is more widespread.

8 Ethics Statement

One potential harm of our rationalisation approach to identify indicators of forced labour is the risk of inadvertently revealing sensitive information through the generated rationales, which could jeopardise the safety of victims. Additionally, if not carefully trained and implemented, the methodology may be exploited to produce misleading explanations, potentially hindering the accurate identification of forced labour indicators. To address concerns around potential harms, we believe that our framework should be used by data professionals and domain experts trained to handle and analyse sensitive information and interpret the rationalisation results appropriately.

References

Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, Sandro Ridella, et al. 2012. The'k'in k-fold cross validation. In *ESANN*, pages 441–446.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 5868–5876.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*.

James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-Parameter Optimization. *Journal of machine learning research*, 13(2).

Ekaba Bisong and Ekaba Bisong. 2019. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64.

Nadya Bliss, Mark Briers, Alice Eckstein, James Goulding, Daniel Lopresti, Anjali Mazumder, and Gavin Smith. 2021. CCC/Code 8.7: Applying AI in the Fight Against Modern Slavery. *arXiv preprint arXiv:2106.13186*.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. *arXiv preprint arXiv:2010.04736*.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2021. Unirex: A unified learning framework for language model rationale extraction. *arXiv preprint arXiv:2112.08802*.

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*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ronen Feldman, James Sanger, et al. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge university press.

Walk Free et al. 2022. Global estimates of modern slavery: Forced labour and forced marriage.

Nadia Ghamrawi and Andrew McCallum. 2005. Collective Multi-label Classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200.

Ellen R Girden. 1992. *ANOVA: Repeated measures*. 84. sage.

Mareike Hartmann and Daniel Sonntag. 2022. A survey on improving nlp models with human explanations. *arXiv preprint arXiv:2204.08892*.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

ILO. 2012. ILO Indicators of Forced Labour. In: Special Action Programme to Combat Forced Labour (SAP-FL). *Special Action Programme to Combat Forced Labour*.

ILO ILO. 1930. Forced labor convention, 1930 (no. 29).

Serhiy Kandul, Vincent Micheli, Juliane Beck, Markus Kneer, Thomas Burri, François Fleuret, and Markus Christen. 2023. Explainable ai: A review of the empirical literature. *Available at SSRN 4325219*.

Amir Kashapov, Tingmin Wu, Sharif Abuadbba, and Carsten Rudolph. 2022. Email summarization to assist users in phishing identification. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, pages 1234–1236.

Eugene Kharitonov, Roberto Dessì, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2021. EGG: a toolkit for research on Emergence of lanGuage in Games. https://github.com/facebookresearch/EGG.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Tao Lei et al. 2017. *Interpretable Neural Models for Natural Language Processing*. Ph.D. thesis, Massachusetts Institute of Technology.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. *arXiv preprint arXiv:1908.11355*.

Hui Liu, Qingyu Yin, and William Yang Wang. 2018. Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.

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*.

Mohammad Reza Ghasemi Madani and Pasquale Minervini. 2023. Refer: An end-to-end rationale extraction framework for explanation regularization. *arXiv preprint arXiv:2310.14418*.

Erick Mendez, Viktor Schlegel, and Riza Batista-Navarro. 2022. RaFoLa: A rationale-annotated corpus for detecting indicators of forced labour. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3610–3625, Marseille, France. European Language Resources Association.

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2022. RaFoLa: A rationale-annotated corpus for detecting indicators of forced labour. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3610–3625, Marseille, France. European Language Resources Association.

Anil Menon, Kishan Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. 1996. Characterization of a class of sigmoid functions with applications to neural networks. *Neural networks*, 9(5):819–835.

Sina Mohseni, Jeremy E Block, and Eric D Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*.

Jerzy Neyman. 1992. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 123–150. Springer.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. *arXiv preprint arXiv:2005.00652*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Samuel S Shapiro and Hassan Zahedi. 1990. Bernoulli trials and discrete distributions. *Journal of Quality Technology*, 22(3):193–205.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Julia Strout, Ye Zhang, and Raymond J Mooney. 2019. Do human rationales improve machine explanations? *arXiv preprint arXiv:1905.13714*.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. 1997. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62.

TAH. 2012. Traffikanalysis.org. https://www.traffikanalysis.org/.

Pratap Tambe and Prerna Tambay. 2020. Reducing modern slavery using ai and blockchain. In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*, pages 22–27. IEEE.

Hao Wang and Yong Dou. 2022. Recent development on extractive rationale for model interpretability: A survey. In *2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT)*, pages 354–358. IEEE.

Nyasha Weinberg, Adriana Bora, Francisca Sassetti, Katharine Bryant, Edgar Rootalu, Karyna Bikziantieieva, Laureen van Breen, Patricia Carrier, Yolanda Lannquist, and Nicolas Miailhe10. 2020. Ai against modern slavery: Digital insights into modern slavery reporting—challenges and opportunities. In *AI for Social Good, Association for the Advancement of Artificial Intelligence Fall Symposium*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in neural information processing systems*, 32.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*.

Dongyu Zhang, Cansu Sen, Jidapa Thadajarassiri, Thomas Hartvigsen, Xiangnan Kong, and Elke Rundensteiner. 2021a. Human-like explanation for text classification with limited attention supervision. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 957–967. IEEE.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. 2021b. Sample efficient reinforcement learning with reinforce. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10887–10895.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 795. NIH Public Access.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*.

Xinyan Zhao and VG Vydiswaran. 2020. Lirex: Augmenting language inference with relevant explanation. *arXiv preprint arXiv:2012.09157*.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31.

A Input Data

Here is a description of the input data of our multi-task rationalisation framework

- **News article**: Original text.
- **Human Rationales**: Human explanations at a label level, including an index of the first character, an index of the last index, the snippet justifying the label, and the indicator of forced labour.
- **Rationale Mask**: The human rationales' matrix format is filled with binary tags indicating the tokens selected as rationales per label.
- **Labels**: Target labels encoded as C-dimensional vectors using one-hot-encoding.

Below, we provide an example of each of the above-mentioned elements.
news article': "But the job wasn't what he expected. He found himself held captive in a call centre, forced to scam foreigners, including Australians, out of thousands of dollars. After arriving in Cambodia at the start of May last year, John Doe was taken to a compound with about 70 other people, he says. His passport was confiscated and he was taught how to scam victims online, tricking Australians, Europeans and Chinese into handing over their money"

'**human rationales**':[[37, 83, 'He found himself held captive in a call centre', 'Restriction of movement'] [222, 278, 'John Doe was taken to a compound with about 70 other people', 'Restriction of movement'], [289, 317, 'His passport was confiscated', 'Retention of identity documents']

'**rationale mask**':
[[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0]
[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]]

Labels: [0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0]

B Loss Function Details

Before describing the loss function, we would like to remind you of the inputs and outputs of our framework and their dimensions:

- **Inputs**: Input Sequence ($B \times L$), Labels ($B \times C$), and Human Rationales ($B \times C \times L$).
- **Outputs**: Predicted Labels ($B \times C$) and Extracted Rationales ($B \times C \times L$).

B corresponds to the batch size, C is the number of classes, and L is the maximum sequence length.

The loss function is a composite of the classification loss and a regularisation over rationale selection. Additionally, we incorporate a loss over rationale prediction in the variant using human rationales. Compiling all the components, the total loss averaged over the batch size is (Equation 5):

$$\text{Loss} = \text{Classification Loss} + \text{Length Regularisation} + \text{Sparsity Regularisation} + \text{Rationale Loss} \quad (5)$$

B.1 Classification Loss

The binary cross-entropy loss with logits for the whole batch is averaged across all instances B and all classes C (Equation 6):

$$\text{Classification Loss} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C}\frac{1}{C}\left[y_{ij}\cdot\log(\sigma(\hat{y}_{ij})) + (1-y_{ij})\cdot\log(1-\sigma(\hat{y}_{ij}))\right] \quad (6)$$

Here, y_{ij} and $\hat{y}_{ij}$ denote the true labels and predicted logits for the j-th class of the i-th instance in the batch, respectively. This formula captures the binary classification loss for each class within each example in the batch.

B.2 Length Regularisation

This component of the loss function (Equation 7) penalises the total length of the rationale to encourage compact rationales:

$$\text{Length Regularisation} = \frac{\lambda}{B}\sum_{i=1}^{B}\sum_{j=1}^{C}\sum_{k=1}^{L}\hat{z}_{ijk} \quad (7)$$

$\hat{z}_{ijk}$ are the elements of the extracted rationale, with λ serving as the regularisation coefficient.

B.3 Sparsity Regularisation

To encourage minimal changes between adjacent rationale elements, the sparsity regularisation is defined as (Equation 8):

$$\text{Sparsity Regularisation} = \frac{\gamma}{B}\sum_{i=1}^{B}\sum_{j=1}^{C}\sum_{k=1}^{L-1}|\hat{z}_{i,j,k+1} - \hat{z}_{ijk}| \quad (8)$$

Here, γ is the coherence factor, emphasising minimal variation between adjacent elements in the extracted rationale, enhancing the coherence of selected rationales.

B.4 Rationale Classification

The loss for rationale classification against the human rationales can be expressed with binary cross-entropy as follows (Equation 9):

$$\text{Rationale Prediction} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C}\sum_{k=1}^{L}\left[z_{ijk}\log(\sigma(\hat{z}_{ijk})) + (1-z_{ijk})\log(1-\sigma(\hat{z}_{ijk}))\right] \quad (9)$$

$\hat{z}_{ijk}$ represents elements of the human rationale used for rationale comparison.

C Label Distribution in the Dataset

Table 3 illustrates the number of news articles assigned to each forced labour indicator and the percentage with respect to the total number of articles in the corpus.

Label	# News Articles	% of the Total
Abuse of vulnerability	731	26.09
Abusive working and living conditions	594	21.20
Debt bondage	107	3.81
Deception	107	3.81
Excessive overtime	160	5.71
Intimidation and threats	30	1.09
Isolation	15	0.54
Physical and sexual violence	289	10.33
Restriction of movement	46	1.63
Retention of identity documents	30	1.09
Withholding of wages	46	1.63

Table 3: Distribution of the number of labels

D Hyperparameter Tuning

Here are the details of the hyperparameter tuning process used in our experiments, including a brief description of each hyperparameter.

- **Regularisation - Length** (λ): The length rationale regularisation term aims to control the length of generated rationales by penalising models for producing excessively long or verbose explanations during training.
- **Regularisation - Sparsity** (γ): The sparsity regularisation term encourages continuity of selections in the generated rationales, discouraging transitions or isolated words as explanations during training.
- **Entropy Coefficient**: The entropy coefficient modulates the exploration-exploitation trade-off of the REINFORCE algorithm by adding a penalty term based on the entropy of the rationale distribution.
- **Rationale Threshold**: Threshold value is used to determine which tokens are included in the generated rationales, allowing the model to select only features surpassing the predefined threshold during inference.

Table 4 describes the search space for each hyperparameter in terms of their sampling distribution and possible values. As mentioned in Section 4, these values were tuned for each classifier using a random search method.

Hyperparameter	Distribution	Value ranges
R - Length (λ)	random	[0.03, 0.06, 0.09, 0.12, 0.15]
R - Sparsity (γ)	random	[0.06, 0.12, 0.18, 0.24, 0.30]
Entropy Coefficient	random	[0.05, 0.10, 0.15, 0.20, 0.25]
Threshold	random	[0.40, 0.45, 0.50, 0.55, 0.60]

Table 4: Hyperparameter search space

Table 5 and table 6 shows the hyperparameter values used for the unsupervised and supervised architectures, respectively.

Model	R - Length (λ)	R - Sparcity (γ)	Entropy Coefficient	RS - Threshold
distilbert-base	0.03	0.12	0.05	0.45
albert-base	0.03	0.18	0.05	0.50
roberta-base	0.09	0.12	0.10	0.40
distilroberta-base	0.06	0.06	0.05	0.50
roberta-large	0.09	0.18	0.10	0.55
xlnet-base	0.03	0.06	0.05	0.45
deberta-base	0.06	0.12	0.15	0.50

Table 5: Hyperparameters used in the unsupervised architectures.

Model	R - Length (λ)	R - Sparcity (γ)	Entropy Coefficient	RS - Threshold
distilbert-base	0.06	0.12	0.10	0.50
albert-base	0.09	0.12	0.10	0.55
roberta-base	0.03	0.06	0.05	0.55
distilroberta-base	0.12	0.18	0.15	0.55
roberta-large	0.06	0.12	0.15	0.50
xlnet-base	0.06	0.12	0.10	0.55
deberta-base	0.09	0.24	0.05	0.55

Table 6: Hyperparameters used in the supervised architectures.

E Detailed Predictive Performance Results

This section details the predictive performance, Label ranking average precision (LRAP) and exact match ratio (EMR), for the unsupervised and supervised rationalisation architectures.

	Unsupervised Architecture		Supervised Architecture	
Model	**LRAP**	**EMR**	**LRAP**	**EMR**
distilroberta-base	0.91 ± 0.03	0.06 ± 0.03	0.90 ± 0.02	0.08 ± 0.02
roberta-base	0.93 ± 0.02	0.08 ± 0.02	0.94 ± 0.03	0.09 ± 0.03
distilbert-base	0.85 ± 0.04	0.10 ± 0.02	0.90 ± 0.02	0.09 ± 0.01
xlnet-base	0.95 ± 0.07	0.11 ± 0.03	0.94 ± 0.04	0.09 ± 0.04
albert-base	0.87 ± 0.03	0.09 ± 0.03	0.91 ± 0.02	0.10 ± 0.02
roberta-large	0.87 ± 0.06	0.06 ± 0.02	0.88 ± 0.03	0.09 ± 0.03
deberta-base	0.93 ± 0.03	0.10 ± 0.03	0.92 ± 0.02	0.12 ± 0.04

Table 7: Cross-validation results for the unsupervised and supervised rationalisation architectures **LRAP**: Label ranking average precision **EMR**: Exact match ratio.