Rationalized Co-Training

Anonymous ACL submission

Abstract

Co-training is a semi-supervised learning technique that leverages two views of the data. It trains a classifier for each view using a small labelled dataset and uses the classifiers to label training data for each other. Intuitively, co-training works by encouraging agreement between the classifiers; an idea exploited in co-regularization. In this work, we propose rationalized co-training: a variant of co-training that encourages agreement between the *rationales* of the classifiers' predictions. Experiments on two datasets showed that rationalized co-training reduces the error rates of the partially and fully supervised models by 32.3%. This error rate reduction outperformed that of vanilla co-training by 8.51%.

1 Introduction

Co-training (Blum and Mitchell, 1998) is a semi-supervised learning technique that uses a large unlabelled dataset to improve a model's performance when only a small labelled dataset is available. To do so, co-training requires two views of the data. Co-training first trains a classifier on each view of the labelled dataset. The most confident predictions of each classifier on the unlabelled dataset are then selected as pseudo-labelled training data for both. Co-training has been successfully applied to many natural language processing (NLP) tasks such as machine translation (Callison-Burch, 2002), sentiment classification (Wan, 2009), and named entity recognition (Li et al., 2013).

Separately, prior works have improved a model's performance by exploiting human-labelled *rationales*: subsets of the inputs that justify the classification (Du et al., 2019; Zaidan et al., 2007; Zhang et al., 2016). Using the movie review task as example, the phrase *"the acting is great!"* in a long movie review is the rationale that justifies the *positive* review. These works argue that rationales can aid learning by 1) directing the learning algorithm's to the important features and 2) reducing overfitting on dataset biases. However, human-labelled rationales may not be readily available. To circumvent this, Bhat et al. (2021) proposed a self-training framework wherein a teacher model generates both task and rationale labels for a student model to learn from. In this paper, we investigate the following idea: can we improve co-training by encouraging the two classifiers to share the rationales behind their predictions? This sharing ideally models how humans discuss: a two-way exchange of thought processes. This question is important because it proposes a possible improvement to co-training, which has been successfully applied to many NLP tasks (Callison-Burch, 2002; Wan, 2009; Li et al., 2013). In this work, we use machine rationales and do not assume access to human-labelled rationales which may not be readily available in practice.

To study this possibility, we propose rationalized co-training: a variant of co-training that encourages agreement between the *rationales* of the classifiers' predictions. Concretely, we use the classifier's most confident predictions and the corresponding rationales as pseudo-labels. Our proposed approach requires a mapping between the rationales of the two views. For example, if we use languages as views, the word alignments between the translated texts are valid mappings. Using a model's rationale as pseudo-labels constraints the choice of models to those that (1) expose their rationales, and (2) can leverage rationales to improve performance. To satisfy these constraints, we used Hierarchical Attention Networks (HAN) (Yang et al., 2016) with an additional rationale agreement loss. Concretely, HAN's attention weights naturally proxies as the model's rationale (first constraint) and can be trained using the loss between the attention weights and the pseudo-labelled rationales (second constraint). While we experimented with HAN, we believe that the same modifications can be extended to most attention neural networks, which

dominate the state-of-the-art neural architectures in NLP (Devlin et al., 2019; Vaswani et al., 2017). Experiments on two datasets from the ERASER benchmark (DeYoung et al., 2020) showed that rationalized co-training reduced the error rates between the partially and fully supervised models by 32.3%. This error rate reduction outperformed that of vanilla co-training by 8.51%.

2 Problem Formulation

In this paper, we consider sentence pairs in two languages (obtained by machine translation) as the two views for co-training. Languages as views have been used in prior works (Callison-Burch, 2002; Wan, 2009; Li et al., 2013). We denote the datasets as D_s for the source language and D_t for the target language. These datasets are partitioned into a large unlabelled dataset U_s, U_t, and a limited labelled dataset, L_s, L_t. These labels consist of only task labels, and not human-labelled rationales.

To apply rationalized co-training, we need a mapping between the rationales in the two views. With languages as views, we use a word aligner $A_{s,t}$, whose objective is to find the correspondence between words of a sentence pair in two languages (Dou and Neubig, 2021). To allow the use of rationales as pseudo-labels, we employ the Hierarchical Attention Networks (Yang et al., 2016) as the base classification model. Finally, the goal is to maximise the task performance of classifiers in both languages C_s and C_t.

3 Hierarchical Attention Networks

We address the problem of document classification (e.g. movie reviews). We assume that a document has N sentences, $\{s_i\}_{i\in[1,N]}$, and each sentence s_i contains T words, $\{w_{it}\}_{t\in[1,T]}$. The classifier C represents each document as a vector, and outputs a single label (e.g., positive/negative).

Word Encoder. We embed words with multilingual BERT (Devlin et al., 2019), $x_{ij} = W_e(w_{ij})$ where W_e denotes the embedding function. To obtain each word's representation, we used a bidirectional GRU (Bahdanau et al., 2016) containing the forward GRU $\overrightarrow{f}$ which reads the sentence s_i from x_{i1} to x_{iT} and a backward GRU $\overleftarrow{f}$ which reads from x_{iT} to x_{i1}:

$$\overrightarrow{h_{it}} = \overrightarrow{GRU}(x_{it}), t \in [1, T],$$
$$\overleftarrow{h_{it}} = \overleftarrow{GRU}(x_{it}), t \in [T, 1].$$

We then obtain each word's representation by concatenating the two hidden states, $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$.

Word Attention. The word attention weighs the importance of each word in the sentence:

$$u_{it} = \tanh(W_w h_{it} + b_w), \quad (1)$$
$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}, \quad (2)$$
$$s_i = \sum_t \alpha_{it} h_{it}. \quad (3)$$

Equation 1 represents a linear layer. Equation 2 measures a word's importance as the normalized similarity of u_{it} with a word level context vector u_w. Equation 3 computes the sentence vector s_i as the weighted sum of the word representations based on the attention.

Sentence Encoder. The sentence encoder is similar to the word encoder. We used a bidirectional GRU to encode the sentences:

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i), i \in [1, N],$$
$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i), i \in [N, 1].$$

We then concatenate the hidden states, $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$. The sentence attention mechanism is similar to the word attention mechanism:

$$u_i = \tanh(W_s h_i + b_s), \quad (4)$$
$$\alpha_i = \frac{\exp(u_i^\top u_s)}{\sum_i \exp(u_i^\top u_s)}, \quad (5)$$
$$v = \sum_i \alpha_i h_i, \quad (6)$$

where v is the document vector summarizing information from all sentences. Lastly, we pass the document vector as input to a linear layer and softmax for classification: $p = softmax(W_c v + b_c)$, where p represents the probability vector.

4 Rationalized Co-Training

Rationale agreement loss. Rationalized co-training uses one classifier's most confident predictions *and corresponding rationales* as pseudo-labels for the other. To do so, we first need a notion of what a classifier's rationale is. Observe that the attention weights α of HAN naturally *proxies* as rationales (equations 2, 5). For example, if the attention weight for the i^{th} sentence is high, the document vector v will then place more weight to

the i^{th} sentence's representation h_i. Similarly for the word attention α_{it}. The classification is based on the document vector v, which is the weighted sum of the sentence representations. Each sentence representation is in turn the weighted sum of the word representations. Thus, the attention weights on both the word and sentence level directly influences the classification, and can be a *proxy* for the classifier's rationale.

We now describe how one classifier can use the rationales of the other to improve task performance. To this end, we introduce $L1$ losses over the attention weights of both classifiers. Formally,

$$L_{sent} = L1(\alpha_{it}, \alpha'_{it}), i \in [1, N]\, t \in [1, T_i],$$
$$L_{doc} = L1(\alpha_i, \alpha'_i), i \in [1, N],$$
$$L = L_{task} + \gamma L_{sent} + \beta L_{doc},$$

where α'_{it}, α'_i are the other classifier's word and sentence rationales, γ, β are hyperparameters to balance the weighted sum. The $L1$ loss is suitable because the attention weights are softmaxed, thus the classifier should not assign weights that are too high or low. Intuitively, the loss encourages the classifier to pay attention to the same words/sentences as the other classifier.

Rationalized co-training algorithm. The rationalized co-training algorithm is similar to the vanilla co-training. The key difference is that we used the classifier's rationales in addition to their most confident task labels as pseudo-labels.

We start by training our classifiers C_s, C_t on the labelled datasets L_s, L_t respectively. The algorithm then iterates the following procedure. First, it computes the number of examples to label: $s = p \times |U_s| \times k$, where p denotes the proportion of the unlabelled set to label and k denotes the iteration number. Second, C_s, C_t label s most confident examples and corresponding rationales from L_s, L_t to form the pseudo-labelled set L'_{ss}, L'_{tt} respectively. L'_{ss} denotes the examples labelled by C_s, whose pseudo-labelled rationales are in language s. However, C_t cannot leverage L'_{ss} directly as C_t is trained in language t. Thus, we used $A_{s,t}$ to align the pseudo-labelled rationales of L'_{ss} in language s to L'_{st} in language t. We do the same with L'_{tt} to create L'_{ts}. Finally, we augment the training datasets L_s, L_t with L'_{ts}, L'_{st} respectively, i.e. $L_s \cup L'_{ts}, L_t \cup L'_{st}$. Note that we did not include L'_{ss} in L_s to avoid C_s from learning from it's own pseudo-labelled examples (without loss of generality to C_t). Lastly, we finetune C_s, C_t on the augmented sets L_s, L_t respectively. While the original co-training paper retrained the classifiers, we opted to fine-tune to save training time.

5 Experiments

We used the following datasets from the ERASER benchmark (DeYoung et al., 2020): 1) *Movie Reviews* is the task of labelling a movie review as positive/negative (size: 2K). 2) *e-SNLI* is the task of determining the inference relation between two short texts: entailment, contradiction, or neutral (size: 570K). We used 20% of the training set as our labelled dataset L, and the remaining 80% as the unlabelled dataset U. The validation set is used for early stopping and model selection. We repeated all experiments thrice and took the average results.

We used the AWESOME word aligner for it's state-of-the-art performance (Dou and Neubig, 2021). As the aligner performs well between English and French, we used these languages as views. As the ERASER benchmark is in English, we Google Translated it to French and mapped the rationales with the aligner.

To measure task performance, we used macro F1 score. To measure co-training performance, we used the error rate reduction between partially and fully supervised models: $1 - \frac{F1_{sup} - F1_{cotrain}}{F1_{sup} - F1_{partial}}$, where $F1_{sup}$, $F1_{cotrain}$, $F1_{partial}$ denote the best F1 score of the fully supervised, partially supervised, and co-training experiments. We also leveraged ERASER's human-labelled rationales to evaluate the correctness of the model's proxy rationales. To this end, we used the L1 loss between the model and human-labelled attention, denoted by L'_S, L'_W for the sentence and word level respectively. To compute the human-labelled word attention, we normalised the human-labelled rationales. To compute the human-labelled doc attention, we compute the proportion of rationale tokens in each sentence and normalized this vector of proportions.

Results and Analysis Our results are summarised in Table 1. Rationalized co-training reduced the error rates between the partially and fully supervised models by an average of 32.3%. This error rate reduction outperformed that of vanilla co-training by an average of 8.51%. Furthermore, the test F1 score is still increasing after 2 co-training epochs whereas vanilla co-training generally tapers after the first epoch (Figure 1). We did not con-

Method	Movies								eSNLI							
	En				Fr				En				Fr			
	ER	F1	L'_S	L'_W	ER	F1	L'_S	L'_W	ER	F1	L'_S	L'_W	ER	F1	L'_S	L'_W
Vanilla	37	73.9	4.1	8.8	45	74.0	4.0	9.5	0.0	69.6	26	13	13	66.7	26	13
Ours	**44**	75.5	**3.4**	8.7	**61**	75.9	**3.5**	9.2	**4.3**	70.0	**17**	**12**	**20**	67.2	**18**	**12**
Full	-	87.7	3.6	8.8	-	80.9	4.0	10	-	75.9	26	14	-	73.7	26	13
Partial	-	65.9	3.8	**8.6**	-	68.3	3.7	**9.1**	-	69.7	26	14	-	65.7	26	13

Table 1: Test set results comparing rationalized co-training to vanilla co-training, fully supervised and partially supervised models. ER denotes error rate reduction. L'_S, L'_W denote the L1 Loss between the best model's and human-labelled sentence/word attention weight (1e-2 units). The highest ER and lowest L'_S, L'_W is bolded. Results shown are the average of three repeats.

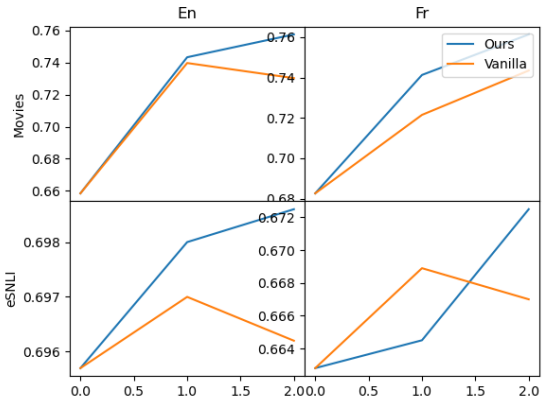

Figure 1: Test F1 scores against co-training epochs.

tinue training due to time constraints. The error rate reduction is higher for movies than eSNLI because the average number of tokens in the former is higher than the latter (774 > 16), thus benefitting more from rationale supervision.

Rationalized co-training also has an average L'_S, L'_W of 0.106, the lowest amongst vanilla (0.131), fully (0.133) and partially supervised baselines (0.131). This result suggests that rationalized co-training aids the learning of human rationales, even without supervision on them. We also note that the fully supervised model has the highest loss, which suggests overfitting to dataset biases.

6 Related Work

Co-training. Co-training was initially proposed with the assumption of conditionally independent views given the class labels (Blum and Mitchell, 1998). However, it has been shown to succeed in many applications that do not satisfy such conditions (Callison-Burch, 2002). Works have since proposed assumptions in weaker forms (Abney, 2002; Wang and Zhou, 2010). Other variants of co-training addressed the assumption of having two conditionally independent views. Zhou and Li (2005) proposed to use three classifiers trained on the same dataset to teach each other. Qiao et al. (2018) trained multiple neural models to be of different views by exploiting adversarial examples to encourage view difference. Sindhwani and Niyogi (2005) extended the idea of co-training by proposing a co-regularisation approach that encourages agreeement between views. Our work builds on this line of work by encouraging agreement between the rationales of the classifier's predictions.

Exploiting Rationales Prior works have shown that training with human-labelled rationales can improve performance. Zaidan et al. (2007) first demonstrated the usefulness of rationales in support vector machines. Zhang et al. (2016) augmented neural networks with rationale supervision for text classification. Melamud et al. (2019) used rationales to augment unsupervised pre-training for text classification. However, human labelled rationales may not be readily available. To circumvent this, Bhat et al. (2021) proposed a self-training framework wherein a teacher model generates both task and rationale labels for a student model to learn from. Our work builds on this line of work by encouraging classifiers to share the rationales behind their predictions.

7 Conclusion

We proposed rationalized co-training: a variant of co-training that encourages agreement between the *rationales* of the classifiers' predictions. Our method requires a mapping between rationales in the two views. As we used languages as views here, the rationales are transferable using the word alignments between sentence pairs. Our approach also requires classifiers which expose their rationales and can leverage rationales to improve performance. We hope to generalise to other classifiers and to cases where the mapping between rationales is non-trivial (e.g. between latent representations).

References

Steven Abney. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata (Subho) Mukherjee. 2021. Self-training with few-shot rationalization: Teacher explanations aid student in few-shot nlu. In *EMNLP 2021*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA. Association for Computing Machinery.

Chris Callison-Burch. 2002. Co-training for statistical machine translation. Master's thesis, University of Edinburgh.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora.

Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2019. Learning credible deep neural networks with rationale regularization.

Yegang Li, Heyan Huang, Xingjian Zhao, and Shumin Shi. 2013. Named entity recognition based on bilingual co-training. In *Chinese Lexical Semantics*, pages 480–489, Berlin, Heidelberg. Springer Berlin Heidelberg.

Oren Melamud, Mihaela Bornea, and Ken Barker. 2019. Combining unsupervised pre-training and annotator rationales to improve low-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3884–3893, Hong Kong, China. Association for Computational Linguistics.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition.

Vikas Sindhwani and Partha Niyogi. 2005. A co-regularized approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Wei Wang and Zhi-Hua Zhou. 2010. A new analysis of co-training. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1135–1142, Madison, WI, USA. Omnipress.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541.

A Rationalized Co-Training Pseudocode

We show the rationalized co-training algorithm in Algorithm 1.

B Experiment Details

B.1 Data Processing

We used the publicly available ERASER benchmark: link. We used the ERASER dataset as it contains human-labelled rationales, which allows us to evaluate the correctness of our model proxy

Algorithm 1 Rationalized Co-Training algorithm.

Require:

- Large unlabelled datasets U_s, U_t in source language s and target language t
- Limited labelled dataset L_s, L_t
- Word aligner $A_{s,t}$

Train classifiers C_s, C_t on L_s, L_t respectively
for $k \leftarrow 1$ to K **do**
  $s \leftarrow p \times |U_s| \times k$
  C_s, C_t labels s most confident examples and corresponding rationales from L_s, L_t to form L'_{ss}, L'_{tt} respectively
  Use $A_{s,t}$ to align L'_{ss} in language s to L'_{st} in language t and L'_{tt} in language t to L'_{ts} in language s.
  $L_s \leftarrow L_s \cup L'_{ts}$
  $L_t \leftarrow L_t \cup L'_{st}$
  Finetune C_s, C_t on L_s, L_t respectively
end for

rationales. To translate the ERASER benchmark from English to French, we google translated the documents and used our word alignment tool to map the rationales. For words that are in English but not in French, we take the next closest word in the alignment. We intend to release the translated datasets and codebase. For our movies dataset, we excluded the query from the document since it is constant across all examples (e.g. "What is the sentiment of this review?).

B.2 Hyperparameter Configurations

We trained the models with the Adam optimiser, setting epochs=20, patience=5, learning rate=1e-3, batch size=64, proportion p=0.1, number of co-training iterations K=2. We also set the attention loss weights γ, β to (5, 25) and (10, 2.5) for the movie reviews and eSNLI dataset respectively. These weights were not tuned. They were chosen with the prior knowledge that sentence attention loss matters more in shorter documents (e.g. eSNLI), while the doc attention loss matters more in longer documents (e.g. Movie Reviews). The values were also not too high such that they overwhelm the task loss and not too low such that it does not aid learning.

B.3 Definition of Human-Labelled Attention

Human-labelled rationales are provided as token-level binary rationale labels, indicating if the token is a rationale or not (e.g. *"The movie is really great"* $\rightarrow$ (0, 0, 1, 1)). To compute the human-labelled word attention, we divided the vector by it's sum (e.g. (0, 0, 1, 1) $\rightarrow$ (0, 0, 0.5, 0.5)). This normalised vector represents the gold attention weights, thus allowing us to evaluate the model's word attention weights via the L1 loss between them. To compute the human-labelled doc attention, we first compute the proportion of rationale tokens in each sentence (e.g. (0, 0, 1, 1) $\rightarrow$ (0.5)). A document is then represented as a vector of proportions (e.g. (0.2, 0.5, 0.4) for a document of 3 sentences). We then divided the vector of proportions by it's sum (e.g. (0.2, 0.5, 0.4) $\rightarrow$ (0.18, 0.45, 0.36)), which allows us to evaluate the model's doc attention weights via the L1 loss. Since the L1 loss is differentiable, we can also use this approach to train HAN with human rationale labels if given.

Dataset	Size (train/val/test)	#Tokens
Movies	1600 / 200 / 200	774
eSNLI	549309 / 9823 / 9807	16

Table 2: Statistics of the datasets used. Tokens is the average number of tokens in each document.

B.4 Reproducibility Checklist

1. **A clear description of the mathematical setting, algorithm, and/or model.** We show the rationalized co-training algorithm in Algorithm 1. The mathematical setting and model is described in Problem Formulation 2 of the main text.

2. **A link to a downloadable source code, with specification of all dependencies, including external libraries (recommended for camera ready, though welcome for initial submission)**. We intend to release the source code (and translated datasets) soon.

3. **A description of computing infrastructure used.** We ran our experiments on a GeForce RTX 2080 Ti.

4. **The average runtime for each model or algorithm, or estimated energy cost**. The estimated runtime for co-training on the Movies and eSNLI dataset is 1 hour and 3 hours respectively.

5. **The number of parameters in each model.** Our HAN model has 1003402 parameters.

6. **Corresponding validation performance for each reported test result.** We intend to release the validation performance in the next iteration of the paper.

7. **A clear definition of the specific evaluation measure or statistics used to report results.** The definition of error rate reduction is detailed in the experiments 5. The definition of L'_S, L'_W is detailed in the appendix B.2.

8. **The exact number of training and evaluation runs.** We trained for 20 epochs and co-trained for 2 epochs.

9. **The bounds for each hyperparameter. The method of choosing hyperparameter values (e.g. manual tuning, uniform sampling, etc.) and the criterion used to select among them (e.g. accuracy)** Hyperparameter tuning was not done. We used the default (PyTorch) values.

10. **Summary statistics of the results (e.g. mean, variance, error bars, etc.).** All results are reported in Table 1. The results are averaged across three runs.

11. **Relevant statistics such as number of examples and label distributions.** The dataset statistics are summarised in Table 2. More details can be found in the ERASER benchmark (DeYoung et al., 2020).

12. **Details of train/validation/test splits.** Specified in Table 2.

13. **An explanation of any data that were excluded, and all pre-processing steps**. Detailed in the appendix section: Experimental Details B.

14. **For natural language data, the name of the language(s)**. English and French.

15. **A link to a downloadable version of the dataset or simulation environment.** The ERASER dataset is available at: link. We intend to release the translated Movies and eSNLI dataset in French soon.

16. **For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.** Detailed in the appendix section: Experimental Details B.