Are Multilingual Sentiment Models Equally Right for the Right Reasons?

Rasmus Kær Jørgensen[1,2], **Fiammetta Caccavale**[3], **Christian Igel**[1] and **Anders Søgaard**[1]
[1]University of Copenhagen, Denmark
[2]PricewaterhouseCoopers (PwC), Denmark
[3]Technical University of Denmark
rasmuskj,igel,soegaard@di.ku.dk
fiacac@kt.dtu.dk

Abstract

Multilingual NLP models provide potential solutions to *the digital language divide*, i.e., cross-language performance disparities. Early analyses of such models have indicated good performance across training languages and good generalization to unseen, related languages. This work examines whether, between related languages, multilingual models are equally *right for the right reasons*, i.e., if interpretability methods reveal that the models put emphasis on the same words as humans. To this end, we provide a new trilingual, parallel corpus of rationale annotations for English, Danish, and Italian sentiment analysis models and use it to benchmark models and interpretability methods. We propose rank-biased overlap as a better metric for comparing input token attributions to human rationale annotations. Our results show: (i) models generally perform well on the languages they are trained on, and align best with human rationales in these languages; (ii) performance is higher on English, even when not a source language, but this performance is not accompanied by higher alignment with human rationales, which suggests that language models favor English, but do not facilitate successful transfer of rationales.

1 Introduction

NLP models are sometimes right for the wrong reasons, e.g., when sentiment analysis models correctly predict a movie review to be positive because it contains the word *Shrek* (Sindhwani and Melville, 2008). Human rationale annotations can be used to evaluate the extent to which models are right for the right reasons, i.e., whether model rationales align with human rationales. Datasets with rationale annotations exist for sentiment analysis (Zaidan and Eisner, 2008), fact-checking (Thorne et al., 2018), natural language inference (Camburu et al., 2018a), and hate speech detection (Mathew et al., 2020),[1]

[1]Several of these datasets can also be found in the ERASER Benchmark (DeYoung et al., 2020).

EN	**A**	**deep**	**and**	**meaningful**	**film**
	2.34	1.69	2.70	1.92	0.09
DA	**En**	**dyb**	**og**	**meningsfuld**	**film**
	0.20	0.79	0.67	2.32	0.11
IT	**Un**	**film**	**profondo**	**e**	**significativo**
	0.44	0.28	1.72	1.79	1.43

Table 1: Tokens with machine generated importance scores for direct translations of the same sentence into English, Danish, and Italian. We see machine rationales are nevertheless quite different; e.g., consider the importance scores for the connectives *and*, *og* and *e*.

but so far only for the English language. While multilingual language models often fail to generalize across distant languages (Singh et al., 2019a; Pires et al., 2019; Rust et al., 2020), they do bridge between related languages and have become a standard solution to data sparsity (Zheng et al., 2021), as well as a way to reduce the overall energy consumption of training language-specific language models (Sahlgren et al., 2021). Benchmark performance does not tell us whether multilingual models are more prone to spurious correlations in some languages rather than others, i.e., whether models are *equally right for the right reasons* or to different degrees, see Table 1.

This paper presents a trilingual parallel corpus of human rationale annotations in Danish, Italian, and English, for the task of sentiment analysis. To this end, we translate an existing sentiment analysis dataset into different languages following a similar procedure as Hu et al. (2020), with human post-correction. We then collect rationales from native speakers of these languages. We evaluate the quality of our human rationale annotations in two ways: using inter-annotator agreement metrics and using human forward prediction experiments (Nguyen, 2018). We then use the corpus to evaluate the extent to which multilingual language models are equally right for the right reasons across languages, and whether agreement with human rationales aligns

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 131 - 141

with downstream performance.

Contributions Our contributions are as follows: (a) We present a trilingual corpus of human rationales, based on post-corrected translations of the Stanford Sentiment Treebank (Socher et al., 2013) and annotated by native speakers. The corpus is made publicly available at https://github.com/RasmusKaer/BlackBox2022. (b) We propose better metrics for comparing ranked rationales than has previously been used, as well as a sequence-wise normalization of LIME's token scores to make scores comparable across sequences. (c) We evaluate MBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019), in conjunction with two interpretability methods, LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), across three languages, quantifying the extent to which these models are *equally right for the right reasons*.

2 Multilingual Rationale Annotation

Our multilingual corpus of human rationales is based on post-corrected translations of the Stanford Sentiment Treebank. We obtain Danish and Italian translations of a sample of validation data, correct the translations manually, and have native speakers annotate the original English sentences, as well as their post-corrected translations. We then validate the annotations by quantifying human inter-annotator agreement and by performing human forward prediction experiments (Doshi-Velez and Kim, 2017; Nguyen, 2018; Hase and Bansal, 2020; Gonzalez and Søgaard, 2020; González et al., 2021). We describe each step in detail in this section.

Stanford Sentiment Treebank (SST) Our dataset builds on a sample of the Stanford Sentiment Treebank, which originally consists of 11,855 sentences from movie reviews, annotated with sentiment labels, and split in training, validation and evaluation sections of 8,544, 1,101, and 2,210 sentences. The sample selected for annotation of the rationales consists of 250 sentences from the validation section.

Translation We translate the English dataset into the target languages using Google Cloud API[2]. We carefully correct the translations of the rationales set manually and assess the quality of corpus through a language analysis. The post-correction process is presented in 6. We are aware that it would have been beneficial to have a set of languages that was more representative of linguistic diversity, but for this work we only had access to professional annotators in the three languages.

Annotation We ask native speakers of English, Danish and Italian to annotate the sample with rationales. Our aim is to identify two types of information for each sentence: the rationales span, snippets of text that support the outcome; and the rank, the most meaningful words to justify the sentiment of the sentence. Inspired by previous explainability work in NLP using human rationale annotations (Mathew et al., 2020; DeYoung et al., 2019; Zhang et al., 2016), we follow the annotation guidelines in Zaidan et al. (2007). For the rank, we are interested in single words that carry a semantic meaning for the output (positive or negative sentiment). Annotators are asked to rank up to five words from most (1) to least (5) meaningful. See Table 2 for an example. The four annotators used in this study had linguistic training and participated on a voluntary basis.

S	John and Adam are such likeable actors.
R	John and Adam are such [2] likeable [1] actors.
S	A warm , funny , engaging film.
R	A warm [3], funny [1], engaging [2] film.

Table 2: Text annotation showing span (S) annotation and rank (R) annotation.

Annotator agreement The inter-annotator agreement is measured as Cohen's κ (Cohen, 1960) and accuracy; see Table 3. The κ coefficients suggest that the two annotators for each language have substantial agreement across all languages.

Lang.	κ	Acc.	Span	Rank	Tokens
DA	0.705	0.882	1,114	722	4157
EN	0.731	0.890	1,250	770	4232
IT	0.642	0.857	1,067	736	4411

Table 3: Annotator agreement and rationales by token. The minimum sentence length is 3 tokens for all three languages. The average length for both EN and DA is 17 and the maximum is 42 tokens per sentence, while in IT it is, respectively, 18 and 44 tokens per sentence.

Forward prediction Besides calculating the inter-annotator agreement, we also validate the

[2]Advanced version (v3), September 2021

quality of our annotations through human forward prediction (Doshi-Velez and Kim, 2017; Nguyen, 2018; Hase and Bansal, 2020; Gonzalez and Søgaard, 2020; González et al., 2021). We recruited 9 annotators from our professional network, and everyone had degrees in computer science or linguistics. In a small-scale side experiment, we show participants 28 examples in which rationales identified by the annotators are highlighted. Participants are then asked to guess the ground truth (positive or negative sentiment) from these highlighted spans. We compare this to a baseline setting in which our participants have to guess the ground truth from raw text. We explicitly mentioned in the task that the results will be used for scientific research. If the rationales help participants predict the ground truth, they have been shown to be good rationales. Humans predicted the ground-truth for 82% of the examples with rationales, compared to 70% of the examples *without* rationales. For example, without rationales provided, 22.2% of annotators struggled in identifying the correct sentiment of a review such as *"Turns a potentially forgettable formula into something strangely diverting"*, while having less difficulties with equally challenging reviews when the rationales are provided. The high inter-annotator agreement and the usefulness of our rationales together indicate that our annotations are of high quality.

3 Comparing Ranked Rationale Lists

To evaluate the agreement between human rationales and rationales identified by interpretability methods applied to automatic sentiment analyses, we need a similarity measure for comparing ranked rationale lists. Common correlation tests are not sufficient, because the measure must be applicable to non-conjoint, uneven lists and should put a higher weight on higher-ranked words.

The human annotator selects the most relevant words in a sentence until exhausted. The ranking is ordered, but may only contain a few words. On the other hand, the interpretability methods provide by design a rank for each word in a sentence. Thus, the annotator's ranking is typically *incomplete* (not all items are ranked), while the automatically computed ranking is *complete*. That is, the two rankings are mutually *non-conjoint*. Furthermore, we need to deal with *indefiniteness* (Webber et al., 2010) in the sense that the annotator may truncate the complete list at an arbitrary depth. The measure we propose for evaluating rationale rankings is the extrapolated version of the *rank-biased overlap* (Webber et al., 2010), RBO_{EXT}, which is a generalization of average based overlap for indefinite rankings. It ranges from 0 (disjoint) to 1 (identical). The RBO_{EXT} measure satisfy the criteria needed for evaluating the agreement of list rationale rankings of both sentences and documents by being able to handle tied ranks, rankings of different lengths and top-weighted rankings.

The degree of top-weightedness is determined by a parameter $p \in [0, 1]$. Consider a person comparing two rankings by sequentially going through the lists starting with the highest rank. In each step, one additional rank is considered. That is, in the beginning only the highest ranked elements are compared, then additionally the top two elements are compared, and so on. At each step, the person stops the comparison with a probability $1 - p$. Roughly speaking, RBO_{EXT} measures the expected similarity computed by this randomized comparison. The parameter p induces a weighting of the ranks that decreases with decreasing rank (i.e., decreasing importance). Following Webber et al. (2010), we choose p such that 86% of the weight is concentrated on the first d ranks. They show that the concentration of weights on the first d ranks given p can be computed as

$$1 - p^{d-1} + \frac{1-p}{p} d \left(\ln \frac{1}{1-p} - \sum_{i=1}^{d-1} \frac{p^i}{i} \right).$$

Table 3 shows that annotators on average rank 3 words per sentence. Hence, we set $p = 0.68$, because this leads to a concentration of roughly 86% for $d = 3$. The annotators were asked to rank up to 5 words. Therefore, we also considered only the top-5 elements in the rankings produced by the interpretability methods (still, we apply RBO_{EXT} as derived for indefinite rankings).

4 Experiments

Our experiments below rely on two pretrained multilingual language models, which we briefly introduce, three different experimental protocols, and two different interpretability methods.

Pretrained language models The experimental protocol is based on two pretrained multilingual transformer language models (Vaswani et al., 2017), namely MBERT (Devlin et al., 2019)[3] and

[3]https://huggingface.co/bert-base-multilingual-cased

XLM-R (Conneau et al., 2019)[4]. We used the base, cased version from the Hugging Face transformers library[5]. Following (Devlin et al., 2019), we added a classification layer on top of the [CLS] token. We fine-tuned these models for 3 epochs on a single Tesla K80 GPU, with a training batch size of 16 and a learning rate of $3 \cdot 10^{-5}$. The parameters were found using manual hyperparameter tuning based on the authors' recommendations of batch-sizes $\{16, 32\}$, epochs $\{2, 3, 4\}$. The learning rate was fine-tuned over $\{2 \cdot 10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$ with 3 trials each.

Experimental protocols In our experiments, we fine-tune MBERT and XLM-R on the SST training data and/or translations thereof (into Danish or Italian). We rely on three standard protocols, which we call the BASE-SETTING, the CROSS-SETTING, and the MULTI-SETTING. In the BASE-SETTING, we fine-tune MBERT and XLM-R on a single language, e.g., English, and evaluate them on the evaluation data in the *same* language. This corresponds to the situation in which you use a multilingual language model to learn a monolingual model in the presence of training data. This scenario is common for *medium-resourced* languages. In the CROSS-SETTING, we evaluate such models, e.g., trained on English, on another language. This scenario is common for *low-resourced* languages. Finally, in the MULTI-SETTING, we train and evaluate on all three languages, inducing a *multilingual* sentiment analysis model for three languages. In all three settings, we evaluate the extent to which the fine-tuned MBERT and XLM-R models align with human rationales, relying on interpretability methods.

Interpretability methods A variety of methods for deriving explanations are currently being used by the NLP community. Examples of such methods are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), LRP (Bach et al., 2015), and DTD (Montavon et al., 2017). For this study, we consider SHAP and LIME, since they are two of the most widely used post-hoc model interpretability methods, also used in similar studies such as ERASER (DeYoung et al., 2020) and HateXplain (Mathew et al., 2020). LIME is a model-agnostic approach that returns an explanation for a prediction on an input example (a text) by virtue of a local linear approximation of the model's behavior around that example. The linear approximation is a sparse linear model induced from hundreds of perturbations of the example. In the case of text examples, perturbations are obtained by randomly removing tokens or words. SHAP is also model-agnostic and based on Shapley values (Shapley, 1953), a concept from cooperative game theory, which refers to the average of the marginal contributions to all possible coalitions. When applied to text, the method, like LIME, produces explanations in terms of tokens or words. We kept the hyperparameters of the two methods to their default-setting, except for the size of neighbourhood used to learn linear models for LIME, which we set to 500 for computational reasons.

[4]https://huggingface.co/xlm-roberta-base

[5]https://huggingface.co/docs/transformers, V4.15.0

5 Results

Table 4 presents the results of the experimental protocol on our trilingual corpus. We compare the effectiveness of LIME and SHAP on human rationales. The agreements is evaluated using ROC AUC for rationale span and RBO_{EXT} for rank similarity based on all 250 samples. The protocol sets two properties for fine-tuning: a single language, denoted by DA, EN and IT, or multiple languages, denoted MULTI. The fine-tuned models are tested across DA, EN and IT with 3 runs per setting.

Performance of MBERT and XLM-R The accuracy of the multilingual models across languages and settings is presented in Table 4. The results confirm the findings of the original works (Conneau et al., 2019), that XLM-R is consistently better than MBERT.

While MBERT-based models consistently obtain their highest accuracy in the BASE-SETTING, XLM-R-based models always perform best on English as the target language, independently from the source language. MBERT-based models exhibit a high variation in the CROSS-SETTING (5.11 p.p. difference between the average accuracy of the BASE compared to the CROSS settings), e.g., EN-MBERT achieves 81.48% accuracy when tested on the English test set, but has only 70.42% accuracy on Danish. In contrast, XLM-R shows less variation between BASE and CROSS settings (0.52 p.p. difference).

But does a higher performance correspond to higher agreement with human rationales? Table 4 presents the results for agreement, evaluated using ROC AUC for rationale span and RBO_{EXT} for rank similarity of the two list rankings. The results sug-

Protocol settings				SHAP		LIME	
Source	Model	Target	Acc.	ROC AUC	RBO_{EXT}	ROC AUC	RBO_{EXT}
English	EN-MBERT	EN	81.48 ± 0.3	68.69 ± 0.7	51.63 ± 0.0	67.08 ± 0.0	53.76 ± 0.0
		IT	74.28 ± 0.6	70.11 ± 1.0	49.92 ± 0.0	66.18 ± 0.0	47.77 ± 0.0
		DA	70.42 ± 0.9	67.41 ± 1.0	44.38 ± 0.0	62.05 ± 0.0	42.35 ± 0.0
	EN-XLM-R	EN	85.37 ± 0.2	69.95 ± 1.4	52.78 ± 0.0	66.83 ± 0.0	56.87 ± 0.0
		IT	82.16 ± 0.2	69.80 ± 0.4	48.52 ± 0.0	68.05 ± 0.0	54.48 ± 0.0
		DA	82.50 ± 0.3	68.85 ± 0.7	50.68 ± 0.0	66.19 ± 0.0	53.33 ± 0.0
Italian	IT-MBERT	IT	80.66 ± 1.2	69.24 ± 1.1	53.24 ± 0.0	68.23 ± 0.0	55.37 ± 0.0
		EN	76.08 ± 1.7	68.79 ± 1.0	50.46 ± 0.0	66.04 ± 0.0	48.62 ± 0.0
		DA	68.94 ± 0.5	65.13 ± 0.6	43.11 ± 0.0	62.66 ± 0.0	43.95 ± 0.0
	IT-XLM-R	IT	82.56 ± 0.0	71.79 ± 1.2	52.79 ± 0.0	69.94 ± 0.0	56.72 ± 0.0
		EN	84.15 ± 0.7	70.62 ± 0.8	55.48 ± 0.0	66.79 ± 0.0	55.22 ± 0.0
		DA	81.24 ± 1.0	69.59 ± 0.4	53.03 ± 0.0	66.16 ± 0.0	52.98 ± 0.0
Danish	DA-MBERT	DA	79.17 ± 0.5	67.40 ± 2.0	49.07 ± 0.0	66.37 ± 0.0	51.33 ± 0.0
		IT	72.10 ± 0.3	68.36 ± 0.8	45.84 ± 0.0	64.74 ± 0.0	45.39 ± 0.0
		EN	75.60 ± 0.7	69.95 ± 0.5	49.50 ± 0.0	66.17 ± 0.0	48.37 ± 0.0
	DA-XLM-R	DA	83.41 ± 0.5	69.74 ± 1.6	55.88 ± 0.0	65.99 ± 0.0	53.27 ± 0.0
		IT	82.07 ± 0.6	69.16 ± 0.6	49.75 ± 0.0	67.57 ± 0.0	52.12 ± 0.0
		EN	84.80 ± 0.2	70.39 ± 1.1	53.63 ± 0.0	66.34 ± 0.0	52.59 ± 0.0
Multi	MULTI-MBERT	EN	81.51 ± 0.1	65.02 ± 2.1	43.49 ± 0.0	65.97 ± 0.0	51.68 ± 0.0
		IT	80.62 ± 0.2	66.16 ± 1.6	45.57 ± 0.0	66.21 ± 0.0	49.60 ± 0.0
		DA	78.34 ± 0.9	63.99 ± 0.4	42.65 ± 0.0	63.89 ± 0.0	49.71 ± 0.0
	MULTI-XLM-R	EN	85.83 ± 0.4	67.79 ± 0.8	50.45 ± 0.0	64.48 ± 0.0	48.66 ± 0.0
		IT	83.67 ± 0.3	69.10 ± 0.7	46.41 ± 0.0	66.52 ± 0.0	51.88 ± 0.0
		DA	82.88 ± 0.7	66.99 ± 1.3	48.89 ± 0.0	64.61 ± 0.0	49.59 ± 0.0

Table 4: Evaluation results on the multilingual corpus of rationales. All results are averaged over three trials. We report the results in percentages. We observe that generally models perform well on the languages they are trained on (source languages), and align best with human rationales in these languages. Generally, MBERT aligns better with human rationales, but XLM-R performs better. We also observe, however, that performance is high on English, even when not a source language, but that this performance is not accompanied by higher alignment with human rationales. This suggests that language models favor English, but do not facilitate successful transfer of rationales.

gest that the accuracy of the models does not generally seem to influence ROC AUC and RBO_{EXT} scores, since a much higher accuracy does not imply better span prediction.

Interpretability methods Our evaluation of the span agreement shows an average across all models and languages of 68.50% for SHAP and 66.04% for LIME, indicating that SHAP has a higher (2.46 p.p.) agreement with human span rationales than LIME. The average rank agreement across all models and languages measured using RBO_{EXT} is 49.46% for SHAP and 51.07% for LIME, the latter being 1.61 p.p. higher in agreement than SHAP. These experiments show that we do not have a single best method across rank and span. Our results suggest a trend of SHAP being a more successful method for capturing good weights for span agreement and LIME being slightly more in accordance with human ranking.

Languages The best rank agreement is achieved when English is used as target language, with the overall highest for both LIME (51.97%) and SHAP (50.93%), as presented in Table 5.

Metric	Method	Target-EN	Target-IT	Target-DA
RBO_{EXT}	SHAP	50.93	49.01	48.46
	LIME	51.97	51.67	49.56
ROC AUC	SHAP	68.90	69.22	67.39
	LIME	66.21	67.18	64.74
Overall		59.50	59.27	57.54

Table 5: To investigate whether explanations are in equal agreement across languages, we group target languages together across the BASE, CROSS and MULTI settings.

The second best rank agreement is obtained in Italian, while the worst is in Danish for both LIME and SHAP. The highest average span score is achieved on Italian, while English follows close and Danish again remain the lowest in agreement. While English is slightly higher in rank agreement, Italian obtains a better span agreement. The lowest span and rank agreement is generally seen with Danish as target language. As we are interested in how languages compare across models, settings and metrics, we can derive the total from the target languages column in Table 5. Altogether, these results indicate that we have better explanations for English (59.50%) than we have for Italian (59.27%)

and Danish (57.54%). The explanations for English are 1.96 p.p. higher in agreement with human rationales than the explanations derived from Danish, while Italian is 1.73 p.p. higher than Danish.

Evaluation metrics An interpretation of the evaluation metrics across settings and languages shows a span agreement that ranges from 62.05% to 71.79%, with an average of 67.27%. What we can interpret from the score is a satisfactory span agreement, suggesting that there is a $\frac{2}{3}$ chance that the model is able to distinguish a token inside a span and a token outside a span. That is, the machine rationale agrees with a human rationale. Regarding the rank agreement across all settings and languages, we see it ranges from 42.35% to 56.87% with an overall average of 50.27%. The score can be interpreted as neither disjoint nor identical, thus implying a fair agreement.

6 Analysis

In this section, we present our analysis of our results and findings. First, we address whether models are *equally right for the right reasons* and how performance compares to agreement. Next, we analyze the translations and the post-corrections. Lastly, we examine whether token scores predict human rationales.

Are models equally right for the right reasons across languages? The idea of being right for the right reasons refers to learning from reliable signals in your data, which are causally related to the ground truth classification. While some models can be used to illuminate complex causal dynamics, others adapt Clever Hans strategies of relying on pervasive, yet spurious correlations in the training data. In this paper, we ask if multilingual language models such as MBERT and XLM-R are equally prone to spurious correlations across languages? Or could it be that these models adopt Clever Hans strategies for some languages, but not for others?

Our results show, very consistently, that MBERT and XLM-R are *less* right for the right reasons for Danish: When the training language is English or Italian, or when multilingual training language is used, Danish never aligns best with human rationales. For English and Italian, it comes in worst in 18/20 cases, and in the multilingual setting, Danish is least right for the right reasons in 6/10 cases. For English and Italian, things are more or less *on par*. While English is slightly higher in rank agreement, then Italian obtains a better span agreement, but the lowest span and rank agreement is generally seen with Danish as the target language. We conclude that multilingual language models are *not* equally right for the right reasons across languages.

How indicative is accuracy for agreement? It seems intuitive that a good model with high performance will also align better with human rationales, but theoretically, models may adopt radically different strategies, if multiple strategies are possible. Even if we expect a positive correlation between performance and alignment, how strong is this correlation in practice? To answer this question, we compute the correlation between the accuracy of the language models and the agreement of span and rank. We use Spearman's rank-order correlation test and Pearson's correlation test, across both explanation methods and all datasets. Both tests show that performance is only weakly (positively) correlated with alignment with human rationales; see Table 6 for details. That is, we see better alignment if models are better, but performance explains only a little of the variance, suggesting multiple possible strategies for prediction exist. This aligns well with our results, also, where a larger difference in accuracy between models does not transfer into a significant difference in agreement.

Lang.	Spearman's ρ	Pearson's ρ
Acc/AUC	0.059^{**}	0.092^{**}
Acc/RBO	0.076^{**}	0.153^{**}

Table 6: Correlation scores for performance (Acc) and alignment with human rationales (AUC/RBO).

Humans may base their rationales on different parts than machine-based rationales. While humans consider *and* necessary for the snippet of *deep and meaningful* (see example in Table 1), a model may not find it a useful predictor of sentiment. Humans and models may agree on the sentiment, but for slightly different reasons.

Language analysis The translated corpus is post-corrected to obtain a high overall quality, ensuring that the corpus can be used to evaluate the interpretability methods in our experiments. To quantify the translations quality, we report the number or sentences that needed corrections and the average number of corrected words in Table 7. The percentage of sentences that needed to have corrections in Italian and Danish are, respectively, 17.20% and

Lang.	% corrected sentences	Avg. corrected words
DA	15.60	1.46
IT	17.20	1.74

Table 7: Percentage of corrected sentences and average number of corrected words per sentence in Italian and Danish.

15.60%. Among these corrected sentences, 1.74 words were corrected on average in Italian, 1.46 in Danish. The results indicate that overall the quality of the translations is high. This is also supported by the performance of the fine-tuned models in Table 4. A selection of original translation and the post-corrected equivalent is presented in Table 8. We can highlight some limitations found during post-correction. The original sentences sometimes present an informal register, sprinkled with colloquial and slang words, which may result in suboptimal and literal translations. Some of the original sentences present idiomatic expressions that might result in a literal translation, as in A-DA, not corresponding to actual terms in the target language. Moreover, some translations may contain

A-IT ORG.	..., sbalorditivo, assurdamente *cattivo*.
A-IT COR.	..., sbalorditivo, assurdamente **brutto**
B-IT ORG.	Questo film *fa impazzire*.
B-IT COR.	Questo film **è esasperante**.
A-DA ORG.	Der er *parcelhuller*, der er store nok til, ...
A-DA COR.	Der er **plothuller**, der er store nok til, ...
B-DA ORG.	Det er en *greb taske* med genrer, ...
B-DA COR.	Det er en **rodekasse** med genrer, ...

Table 8: Examples of corrected translations (COR.) and the original translations (ORG.).

subpar syntactic structure or lexicon, e.g., in A-IT *brutto* is more suiting to refer to *films*, although it presents the same polarity and magnitude of the original adjective. In B-IT the sentiment of the expression could be misinterpreted, since *fa impazzire* is sometimes used in a positive connotation. Lastly, sometimes the original English sentences contain typos and other errors, which the model is understandably not able to correct or process, therefore transferred into the translations.

Do token scores predict human rationales Meaningful token scores produced by an interpretability method should be predictive of human rationales (Doshi-Velez and Kim, 2017; Nguyen, 2018; DeYoung et al., 2019). To verify this, we map the token score $s(w)$ of a word w to an estimate of the probability that the word is in the rationales span. We assume a logistic model

$$P(w \text{ in rationales span} \mid s(w)) = \sigma_{a,b}(|s(w)|) \ ,$$

where $\sigma_{a,b}(x) = (1 + \exp(ax + b))^{-1}$ with scalar parameters a and b. These parameters are determined by maximum likelihood estimation on a training set pairing token scores and corresponding human annotations. We consider the absolute value of the score because we are interested in the importance of a word regardless of whether it contributes to a positive or negative sentiment. This approach corresponds to calibrating the (absolute) scores to posterior probabilities as suggested by Platt (Platt, 1999; Niculescu-Mizil and Caruana, 2005). It can also be viewed as logistic regression from the absolute score to the dependent variable indicating whether a word is in the rationale span or not.

The logistic model gives us the probability of a word being a rationale, which allows for an interpretation of token scores and a comparison of scores across different interpretability methods. In particular, the model suggests a criterion for deciding whether a word should be considered part of the rationales span or not by applying the natural 50% threshold on the probabilities (we pay for this additional information by using training data to fit the models). To fit the model and to compare the different interpretability methods, we split our data into a training and a validation set. We used 25 positive and 25 negative samples for validation and trained on the remaining 200 data points.

Let $\boldsymbol{s} = (s(w_1), s(w_2), \dots)^{\mathrm{T}}$ denote the vector of scores for a word sequence $w_1, w_2, \dots$ and $\min(\boldsymbol{s})$ and $\max(\boldsymbol{s})$ the minimum and maximum element of $\boldsymbol{s}$, respectively. To compare token scores across sequences, their scaling should not differ across the sequences. That is, because we can assume that each sequence contains at least one word within and one outside the span, for two sequence $\boldsymbol{s}$ and $\boldsymbol{s}'$ we should have $\min(\boldsymbol{s}) = \min(\boldsymbol{s}')$ and $\max(\boldsymbol{s}) = \max(\boldsymbol{s}')$. We found this property to be violated, in particular for LIME. Thus, we normalized the scores at the sequence level using

$$s(w) \leftarrow \frac{s(w) - \min(\boldsymbol{s})}{\max(\boldsymbol{s}) - \min(\boldsymbol{s})}$$

for each score $s(w)$ in a sequence with scores $\boldsymbol{s}$.

Table 9 shows the accuracies on the held-out sets in BASE-SETTING. Both methods performed better

		LIME MBERT	LIME XLM-R	SHAP MBERT	SHAP XLM-R	BASE LINE
(A)	EN	70.03	71.51	71.68	72.76	67.74
	DA	69.50	70.23	70.83	72.75	67.49
	IT	70.94	72.73	72.73	73.78	67.80
(B)	EN	73.75	73.03	70.97	71.68	67.74
	DA	72.34	72.75	71.47	72.70	67.49
	IT	73.47	75.44	73.30	73.13	67.80

Table 9: The accuracies on the hold-out sets in BASE-SETTING. The BASELINE is a majority classifier that naively predicts all tokens as not a rationale. (A) refers to the original token scores and (B) to the normalized token scores.

than simply predicting the majority class. Without normalization, SHAP outperformed LIME on our (rather small) validation data set. LIME was only slightly better than the baseline, but after normalization LIME surpassed SHAP, which did not profit from the normalization. When evaluating explanations on how well the token scores generalize to human rationales, we see a similar pattern of Italian and English sharing the highest agreement where Danish consistently shows the lowest agreement.

Human annotated rationales include connectives, determiners, and similar, which are irrelevant for our binary task and are therefore not used by the logistic models. This suggests that methods for adding the relevance of these could be a promising direction for improving our approach and the evaluation between human and machine rationales.

7 Related work

Transformer-based multilingual models have been analyzed in many ways: Researchers have, for example, looked at performance differences across languages (Singh et al., 2019b), looked at their organization of language types (Rama et al., 2020), used similarity analysis to probe their representations (Kudugunta et al., 2019), and investigated how learned self-attention in the Transformer blocks affects different languages (Ravishankar et al., 2021). Human rationales have been used to supervise attention for various text classification tasks, such as sentiment analysis (Zhong et al., 2019) and machine translation (Yin et al., 2021). Feature attribution methods such as LIME and SHAP have also been applied to multilingual models: LIME has been applied to MBERT for analysis of hate speech models (Aluru et al., 2020), and SHAP has been applied to MBERT in biomedical NLP (Zaragoza, 2021). LIME has also been applied to XLM-R in the context of hate speech (Socha, 2020), as well as in a biomedical context (Koloski et al., 2021). Shapley values have also been used to estimate the influence of source languages on the final predictions of models based on MBERT (Parvez and Chang, 2021). None of these applications have been evaluated, however. Feature attributions have been applied to monolingual models, especially for English, more often than multilingual models. For English, we have a set of datasets with human rationales that we can use to evaluate feature attribution methods. These include BeerAdvocate (Bastings et al., 2019) and e-SNLI (Camburu et al., 2018b), as well as other datasets, several of which were collected in the ERASER benchmark (DeYoung et al., 2020). The reason feature attribution methods have not been properly evaluated in a multilingual context, is simple: There was, until now, no gold standard with which to evaluate the rationales produced by multilingual models.

8 Conclusions

We introduced a new trilingual, parallel corpus of human rank and span rationales in three related languages, English, Danish and Italian. We proposed rank-biased overlap as a better metric for rank evaluation when common correlation tests are not sufficient. We found that a sequence-wise normalization of LIME's token scores is required to make scores comparable across sequences. Evaluations on the corpus showed that generally, models perform well on the languages they are trained on, and align best with human rationales in these languages. Models can be right for different reasons. The main results suggest that multilingual models are *not* equally right for the right reasons in the sense that interpretability methods indicate that the models not necessarily put emphasis on the same words as humans. We also observed that performance is high on English, even when it is not a source language, but that this superior performance is not accompanied by higher alignment with human rationales. In other words, this zero-shot advantage of English as a target language seems to come at the cost of being more prone to spurious correlations. With this work, we hope to inspire further progress on multilingual interpretation and collection of rationales in different languages.

9 Limitations

All the languages chosen for the presented work belong to the Indo-European language family, since we only had access to professional annotators in the three languages. A clear limitation of this study is the lack of linguistic diversity in the set of languages used. It would be beneficial in the future to build larger rationale datasets for less related languages, including languages from different language families. Another limitation to be highlighted is the limited size of the multilingual parallel corpus of rationales, consisting on 250 annotations per language. Finally, although the parallel corpus was post-corrected, the language models are fine-tuned on the translations.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018a. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018b. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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

Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online. Association for Computational Linguistics.

Ana Valeria Gonzalez and Anders Søgaard. 2020. The reverse turing test for evaluating interpretability methods on unknown tasks. In *NeurIPS 2020 Workshop on Human And Model in the Loop Evaluation and Training Strategies*.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Boshko Koloski, Timen Stepišnik-Perdih, Senja Pollak, and Blaž Škrlj. 2021. Identification of covid-19 related fake news via neural stacking. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 177–188, Cham. Springer International Publishing.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alex J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? On the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*.

Magnus Sahlgren, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson. 2021. It's basically the same language anyway: the case for a nordic language model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 367–372, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Lloyd S. Shapley. 1953. *A Value for n-Person Games*, pages 307–318. Princeton University Press.

Vikas Sindhwani and Prem Melville. 2008. Document-word co-regularization for semi-supervised sentiment analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1025–1030.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019a. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019b. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Kasper Socha. 2020. KS@LTH at SemEval-2020 task 12: Fine-tuning multi- and monolingual transformer models for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2045–2053, Barcelona (online). International Committee for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages

1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online. Association for Computational Linguistics.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. 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.

Omar Emilio Contreras Zaragoza. 2021. Explainable antibiotics prescriptions in nlp with transformer models. Master's thesis, Stockholm University.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas. Association for Computational Linguistics.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.