Are Multilingual Sentiment Models Equally Right for the Right Reasons?

Rasmus Kær Jørgensen\textsuperscript{1,2}, Fiammetta Caccavale\textsuperscript{3}, Christian Igel\textsuperscript{1} and Anders Søgaard\textsuperscript{1}

\textsuperscript{1}University of Copenhagen, Denmark \textsuperscript{2}PricewaterhouseCoopers (PwC), Denmark \textsuperscript{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 \textit{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),\textsuperscript{1} 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

\begin{tabular}{|c|c|c|c|c|}
\hline
& EN & deep & and & meaningful & film \\
\hline
DA & En & dyb & og & meningfuld & film \\
\hline
0.20 & 0.79 & 0.67 & 2.32 & 0.11 \\
\hline
IT & Un & film & profondo & e & significativo \\
\hline
0.44 & 0.28 & 1.72 & 1.79 & 1.43 \\
\hline
\end{tabular}

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 \textit{and}, \textit{og} and \textit{e}.

\textsuperscript{1}Several of these datasets can also be found in the ERASER Benchmark (DeYoung et al., 2020).
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 API2. 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>A warm, funny, engaging film.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Lang</th>
<th>κ</th>
<th>Acc.</th>
<th>Span</th>
<th>Rank</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>0.705</td>
<td>0.882</td>
<td>1,114</td>
<td>722</td>
<td>4157</td>
</tr>
<tr>
<td>EN</td>
<td>0.731</td>
<td>0.890</td>
<td>1,230</td>
<td>770</td>
<td>4232</td>
</tr>
<tr>
<td>IT</td>
<td>0.642</td>
<td>0.857</td>
<td>1,067</td>
<td>736</td>
<td>4411</td>
</tr>
</tbody>
</table>

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
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

$$1 - p^{d-1} + \frac{1 - p}{p} \sum_{i=1}^{d-1} \frac{p^i}{i!}$$

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 batches-
izes \(\{16, 32\}\), 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 lan-
guage, e.g., English, and evaluate them on the eval-
uation data in the **same** language. This corresponds
to the situation in which you use a multilingual lan-
guage 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 set-
ings, 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 (Lund-
berg 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 inter-
pretability methods, also used in similar studies
such as ERASER (DeYoung et al., 2020) and Hat-
eXplain (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 behav-
ior 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 ran-
domly removing tokens or words. SHAP is also
model-agnostic and based on Shapley values (Shap-
ley, 1953), a concept from cooperative game the-
ory, which refers to the average of the marginal
contributions to all possible coalitions. When ap-
plied 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 neighbour-
hood used to learn linear models for LIME, which
we set to 500 for computational reasons.

### 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\(_{\text{EXT}}\) for rank simi-
ilarity based on all 250 samples. The protocol sets
two properties for fine-tuning: a single language, de-
noted 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 ac-
curacy of the multilingual models across languages
and settings is presented in Table 4. The results con-
firm the findings of the original works (Conneau
et al., 2019), that XLM-R is consistently better
than MBERT.

While MBERT-based models consistently ob-
tain their highest accuracy in the **BASE-SETTING**, XLM-R-based models always perform best on En-
GLISH as the target language, independently from
the source language. MBERT-based models ex-
hibit 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 varia-
tion 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\(_{\text{EXT}}\) for rank
similarity of the two list rankings. The results sug-

---

\(^4\)https://huggingface.co/xlm-roberta-base

\(^5\)https://huggingface.co/docs/transformers, V4.15.0
suggest that the accuracy of the models does not generally seem to influence ROC AUC and RBO\textsubscript{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\textsubscript{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.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Target-EN</th>
<th>Target-IT</th>
<th>Target-DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBO\textsubscript{EXT}</td>
<td>SHAP</td>
<td>50.93</td>
<td>49.01</td>
<td>48.46</td>
</tr>
<tr>
<td></td>
<td>LIME</td>
<td>51.97</td>
<td>51.67</td>
<td>49.56</td>
</tr>
<tr>
<td>ROC</td>
<td>SHAP</td>
<td>68.90</td>
<td>69.22</td>
<td>67.39</td>
</tr>
<tr>
<td></td>
<td>LIME</td>
<td>66.21</td>
<td>67.18</td>
<td>64.74</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Spearman’s $\rho$</th>
<th>Pearson’s $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc/AUC</td>
<td>0.059**</td>
<td>0.092**</td>
</tr>
<tr>
<td>Acc/RBO</td>
<td>0.076**</td>
<td>0.153**</td>
</tr>
</tbody>
</table>

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
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

subpar syntactic structure or lexicon, e.g., in A-IT brutto is more suited 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.

<table>
<thead>
<tr>
<th>Lang.</th>
<th>% corrected sentences</th>
<th>Avg. corrected words</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>15.60</td>
<td>1.46</td>
</tr>
<tr>
<td>IT</td>
<td>17.20</td>
<td>1.74</td>
</tr>
</tbody>
</table>

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

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

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 $s = (s(w_1), s(w_2), \ldots)^T$ denote the vector of scores for a word sequence $w_1, w_2, \ldots$ and $\min(s)$ and $\max(s)$ the minimum and maximum element of $s$, 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 $s$ and $s'$ we should have $\min(s) = \min(s')$ and $\max(s) = \max(s')$. We found this property to be violated, in particular for LIME. Thus, we normalized the scores at the sequence level using

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

for each score $s(w)$ in a sequence with scores $s$. Table 9 shows the accuracies on the held-out sets in BASE-SETTING. Both methods performed better
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


