Investigating the Saliency of Sentiment Expressions in Aspect-Based Sentiment Analysis

Joachim Wagner and **Jennifer Foster**
School of Computing, ADAPT Centre
Dublin City University
Dublin, Ireland
firstname.lastname@dcu.ie

Abstract

We examine the behaviour of an aspect-based sentiment classifier built by fine-tuning the $\text{BERT}_{\text{BASE}}$ model on the SemEval 2016 English dataset. In a set of masking experiments, we examine the extent to which the tokens identified as salient by LIME and a gradient-based method are being used by the classifier. We find that both methods are able to produce faithful rationales, with LIME outperforming the gradient-based method. We also identify a set of manually annotated sentiment expressions for this dataset, and carry out more masking experiments with these as human rationales. The enhanced performance of a classifier that only sees the relevant sentiment expressions suggests that they are not being used to their full potential. A comparison of the LIME and gradient rationales with the sentiment expressions reveals only a moderate level of agreement. Some disagreements are related to the fixed length of the rationales and the tendency of the rationales to contain content words related to the aspect itself.

1 Introduction

Saliency approaches to understanding the output of a model attempt to locate the parts of the input that contribute most to the model's decision. These salient words (or rationales) are often evaluated by measuring their faithfulness to the model, i.e. the extent to which the model relies on them to arrive at a prediction. The jury is still out regarding the best saliency method, whether it is a black-box, model-independent method such as LIME (Ribeiro et al., 2016) or a model-dependent method such as those which use the training loss function gradient to obtain the rationale (Baehrens et al., 2010). Atanasova et al. (2020) compare saliency methods on three text classification tasks and find that gradient-based rationales perform the best overall. DeYoung et al. (2020), on a different set of tasks, find that LIME produces more faithful rationales than gradient or attention-based rationales.

Rationales produced by saliency methods are often compared to human rationales. DeYoung et al. (2020) introduce the ERASER benchmark of human rationales for datasets covering the tasks of fact verification, movie review sentiment polarity classification, evidence inference, natural language inference and various forms of QA. We build on this work by sourcing a set of human rationales for the task of *aspect-based sentiment analysis*.

Aspect-based sentiment analysis (ABSA) is a form of fine-grained sentiment analysis that attempts to determine the opinion about some aspect of a topic. For example, given a restaurant review

Example 1. *I love where it is located but the service leaves much to be desired*

a system should return the polarity `positive` for the "location" aspect and `negative` for the "service" aspect. Much work in ABSA has taken place within the context of the SemEval 2014-2016 shared tasks which use online consumer reviews of laptops, restaurants and hotels (Pontiki et al., 2014, 2015, 2016). State-of-the-art approaches (Sun et al., 2019; Truşcă et al., 2020) are underpinned by pre-trained models such as BERT (Devlin et al., 2019).

We carry out a novel interpretability study for aspect-based sentiment analysis, focusing on the SemEval 2016 English dataset and a polarity classifier built by fine-tuning BERT. Our study attempts to answer the following questions:

1. Which approach produces the most faithful rationales for this task: LIME or a gradient-based approach?
2. How faithful are the human rationales if taken as rationales for the BERT-based classifier?
3. To what extent do the LIME and gradient-based rationales agree with human rationales?

In answer to our first research question, we conduct a masking study and find that, while both

Findings of the Association for Computational Linguistics: ACL 2023, pages 12751–12769

methods are useful for selecting rationales, LIME produces more faithful rationales.

In order to answer our second research question, we need human rationales for the dataset we are using. It turns out that such data already exists, in the form of manually annotated sentiment expression spans in the SemEval 2016 data. These annotations were created by Kaljahi and Foster (2018) in the context of an effort to boost classifier performance rather than to produce human rationales for an interpretability study. However, they are well suited as human rationales as the sentiment expressions are the words the annotators judged to be expressing the sentiment towards the aspect. We show that the performance of our BERT model improves substantially when all but the sentiment expression tokens are masked (during training and testing), suggesting that they are not currently being used to their full potential. This experiment also reveals that sentiment expressions that are not referring to the aspect in focus are also being used.

For our third research question, we measure the alignment of the LIME and gradient rationales with these human rationales or sentiment expressions. Measures of the alignment between automatic and human rationales can serve as an evaluation measure of rationale extraction methods, providing a different view than existing automatic evaluation measures. At the same time, we should not expect the two to fully agree since they are not designed to reflect exactly the same thing. Human rationales are the words that human annotators judged to be contributing to the sentiment towards the aspect, and the automatic rationales are designed to reflect what the classifier used regardless of whether they are sentiment-bearing, related to the aspect or neither. We find only moderate agreement between the two types of rationales. The highest overlap with the human rationales (approx 60%) is obtained by an automatic rationale length of 50% of the input.

How much of the difference can be explained? Some of the words in the LIME and gradient rationales that are not in the human rationales are words that are used to identify the focus of the sentiment, i.e. those that are related to the aspect, e.g. *service* in Ex. 1. It is reasonable to expect these to be used by the classifier in order to locate the sentiment expressions. We also note structural differences – the human rationales have been mostly annotated as continuous spans (containing function words)whereas the automatic rationales do not have this property; the automatic rationales are a fixed proportion of the input length whereas the human rationales are not. The fixed length of the automatic rationales is an important factor because agreement with the human rationales jumps to over 80% when an oracle length is employed.

2 Dataset

We use the SemEval 2016 English ABSA dataset, focusing on SubTask B[1] where the label to be predicted is the sentiment polarity (positive, negative or neutral), the text granularity is sentence-level and the aspect category is supplied with the input sentence. The aspect category is of the form ENTITY#ASPECT, e. g. LAPTOP#BATTERY. There are a total of 2000 training sentences and 676 test in the restaurant domain, with 2500 training and 808 test sentences in the laptop domain. The number of training and test instances is higher as approximately 15% of sentences are annotated with multiple aspects as in Ex. 1.

Kaljahi and Foster (2018) add an additional layer of annotation[2] to this dataset by marking the spans of the sentiment expressions (SEs) in each sentence. In Ex. 2, the SE is the phrase *good quality*.

Example 2. *The display is **good quality** .*

For sentences with neutral polarity where no opinion is expressed, the SE is empty – see Ex. 3 from Kaljahi and Foster (2018).

Example 3. *We had lunch in that restaurant last week.*

We use these SEs as the human rationales in our experiments.

3 Sentiment Classifier

Following Sun et al. (2019), we fine-tune English uncased $\text{BERT}_{\text{BASE}}$ using auxiliary questions that are fed into BERT as sequence "A" together with the review sentence as sequence "B".

Example 4. **A** *restaurant: What do you think of the QUALITY of FOOD?* **B** *The food was lousy - too sweet or too salty and the portions tiny*

We reserve 5% of the training data as development data, keeping the same distribution of domains and target labels as in the full training data. We jointly train on the laptop and restaurant domains, concatenating the respective training and development sets,

[1] http://alt.qcri.org/semeval2016/task5/
[2] https://opengogs.adaptcentre.ie/rszk/sea

prefixing the question with a domain label ("laptop:" or "restaurant:") to help the model to adjust to domain-specific patterns.

We train twelve classifiers for ten epochs, selecting the epoch with highest accuracy according to development data. The same set of twelve random seeds is used in all settings. The seed controls the randomisation of the training-development split, initialisation of the classification head and any other randomised parts of training, e. g. training batch creation. Hyperparameters are detailed in the appendix.

4 Saliency Methods and Rationales

We employ a gradient-based saliency method and a black-box explanation method from which we derive a saliency score. We do not include attention-based saliency methods as recent discussion suggests caution with these (Jain and Wallace, 2019; Bastings and Filippova, 2020) and as results of Chrysostomou and Aletras (2022b) show superior faithfulness of gradient-based rationales over attention-based rationales for seven of eight settings tested. We also do not include methods that modify a classifier to produce a rationale as a side-product or as an intermediate step of the classification and that would produce classifiers that are not comparable to our baseline classifier (Lei et al., 2016; Bastings et al., 2019; Glockner et al., 2020; Paranjape et al., 2020).

Once each token of the input sequence "B" with length n has received a saliency score, we select the top $k = \lfloor 0.5 + nL \rfloor$ tokens as the rationale, where L is the relative rationale length. We try $L \in \{0.25, 0.5, 0.75\}$.[3] This means that in each setting the relative rationale length is fixed.

Gradient-based Saliency Methods Using the absolute value of the gradient for measuring instance-level feature importance was proposed by Baehrens et al. (2010) and has been introduced to NLP by Denil et al. (2015) and Li et al. (2016). Gradient-based saliency methods typically use the derivative of the loss function used in model training with respect to the inputs, either with the gold label or the model's predictions as an indicator of feature importance. We use the model's prediction as the reference label in our experiments.

Based on empirical results of Atanasova et al. (2020), we multiply the gradient with the input and apply the L_2 norm to aggregate saliency scores for each input BERT subword unit. The former is a method based on work of Kindermans et al. (2016) who explore saliency measures based on Taylor decomposition. L_2 normalisation is also used by Arras et al. (2016). In the (for English) rare case that an input token is split into two or more BERT subword units, we further aggregate subword scores to token scores taking the score of the highest scoring BERT unit. In the following, we write $R_{x\nabla x}$ for these rationales.[4]

Black-box Saliency Methods Black-box saliency methods obtain saliency scores by observing changes in the prediction when parts of the input are masked. Methods vary in how the input sequences are sampled from all possible 2^n masked sequences for a sequence of length n and how saliency scores are derived from the observations. LIME (Ribeiro et al., 2016) samples from all masks according to a probability distribution that keeps the number of samples with each possible number of mask tokens (zero to sentence length) uniform. We set the number of LIME samples to 10,000.[5] LIME trains a linear model to replicate the observed behaviour when the input is only a binary vector indicating which tokens were masked. The parameters of this model are used as indicators of the importance of each input feature. LIME prefers to work with class probabilities rather than just the predicted class to also gain information from input modifications that do not cause a change in the predicted class. We provide LIME with such class probabilities from our classifiers. In case of multi-class classification, as in our task, LIME produces a score for each class. We derive rationales, R_{LIME}, using the absolute value of the score for the predicted class, thereby including tokens as influential that oppose the overall prediction.

[3] The average SE length would suggest using a small value for L. However, Figure 3 shows the highest agreement of rationales with SEs for $L \geq 50$. Our choice $L \in \{0.25, 0.5, 0.75\}$ balances the trend in Figure 3, the need for keeping the number of values small for ROAR (Section 5.1) and the desire to include short rationales.

[4] We also tried integrated gradients but they did not perform better than point gradients in our experiments.

[5] We detect duplicate samples and query the classifier only once for each unique input, reducing the number of queries from 70.8 million to 32.8 million per run.

5 Experiments

5.1 Setup

In our first experiment, to answer our first research question, we mask all input tokens apart from the rationales identified by the saliency methods, $R_{x\nabla x}$and R_{LIME}, and then perform the complement masking, i. e. masking the rationale tokens and training/testing on the remaining words. Adopting the terminology of DeYoung et al. (2020), the former tests the *sufficiency* of the rationales, and the latter tests their *comprehensiveness*.[6] We compare to a baseline with random tokens as rationale.

In our second experiment, corresponding to our second research question, we carry out similar masking experiments using the SEs identified by Kaljahi and Foster (2018) as the human rationales. In our third experiment, we attempt to answer our third research question by examining how well the rationale tokens align with the SEs.

For the masking experiments, we derive new training, development and test sets by applying the same type of masking to all three sets.[7] We then train dedicated classifiers for the type of masking so that all tests are in-distribution. This approach is known as ROAR (Hooker et al., 2019).[8] We choose this method to avoid testing a model outside of its training distribution. The use of ROAR changes *how* sufficiency and comprehensiveness are measured but not *what* these metrics measure, i. e. whether the selected rationale tokens are enough to make the prediction and cover all supporting evidence for the prediction.

All code of our experiments will be made available on `https://github.com/jowagner/absa-rationale-eval`.

5.2 Baselines: Masking Random Tokens

We train classifiers as follows:

- **Full**: No masking. Sequence B is the full review sentence.
- **None**: All tokens in sequence B are masked. The classifier can still use the number of tokens in sequence B and the domain and aspect information provided in sequence A.
- $\mathbf{R_{RAND}}@L$ and $\neg\mathbf{R_{RAND}}@L$: For each item in the data with length n, we randomly select $k = \lfloor 0.5 + nL \rfloor$ tokens. We try $L \in \{0.25, 0.5, 0.75\}$. The same masks are used in each epoch and the masks are incremental, e. g. all tokens selected in $\mathbf{R_{RAND}}@0.25$ are also included in $\mathbf{R_{RAND}}@0.5$.

5.3 Experiment 1: Masking LIME and Gradient Rationales

We obtain rationales with relative length L, $L \in \{0.25, 0.5, 0.75\}$ by applying a saliency method R to the **Full** classifier, R = $R_{x\nabla x}$, R_{LIME}, producing six data sets, e. g. $R_{x\nabla x}@0.25$ are gradient-based rationales with quarter the length of the input. Furthermore, we produce another six datasets with the complementary masks, e. g. $\neg R_{x\nabla x}@0.25$ are the 75% tokens not selected by $R_{x\nabla x}@0.25$.

5.4 Experiment 2: Masking Human Rationales (Sentiment Expressions)

We train and test classifiers with input masked according to sentiment expressions:

- **SE**: Sequence B is the sentiment expression. Other words of the review are masked.
- **¬SE**: Sequence B is the review sentence with the sentiment expression masked.
- **U-SE** and **¬U-SE**: As SE/¬SE but with SEs extended to the union of all SEs (U-SE) for sentences with multiple aspects (see Ex.1).

Example 5 shows the first training sentence of the laptop domain with the SE (in bold) and the masked sequences for some of the settings introduced above.

Example 5.

Full	*All I can say is **W-O-W** .*
SE	*[MASK] [MASK] [MASK] [MASK] [MASK] W-O-W [MASK]*
¬SE	*All I can say is [MASK] .*
$\mathbf{R_{x\nabla x}}$**@.5**	*All [MASK] can say [MASK] W-O-W [MASK]*
$\neg\mathbf{R_{x\nabla x}}$**@.5**	*[MASK] I [MASK] [MASK] is [MASK] .*
$\mathbf{R_{RAND}}$**@.5**	*[MASK] [MASK] can say is [MASK] .*

Appendix B shows further examples of sentiment expressions and rationales.

[6] Sufficiency is also related to Treviso and Martins (2020)'s concept of successful communication of the information necessary to replicate the prediction.

[7] The modification is restricted to the review sentences, i. e. sequence A is never masked.

[8] See the Limitations section on the influence of randomness in training and the conflation of model performance and measurement of rationale quality.

Input	Accuracy	Input	Accuracy
Full	84.6 ±0.6	**None**	69.3 ±1.7
$\mathbf{R_{RAND}}$**@.25**	72.2 ±1.0	$\neg\mathbf{R_{RAND}}$**@.25**	80.8 ±0.9
$\mathbf{R_{RAND}}$**@.5**	77.2 ±1.3	$\neg\mathbf{R_{RAND}}$**@.5**	76.3 ±1.1
$\mathbf{R_{RAND}}$**@.75**	81.2 ±0.9	$\neg\mathbf{R_{RAND}}$**@.75**	71.9 ±0.9
$\mathbf{R}_{x\nabla x}$**@.25**	79.3 ±1.7	$\neg\mathbf{R}_{x\nabla x}$**@.25**	75.1 ±0.6
$\mathbf{R}_{x\nabla x}$**@.5**	83.5 ±0.7	$\neg\mathbf{R}_{x\nabla x}$**@.5**	71.1 ±1.3
$\mathbf{R}_{x\nabla x}$**@.75**	84.5 ±0.9	$\neg\mathbf{R}_{x\nabla x}$**@.75**	69.7 ±1.8
$\mathbf{R_{LIME}}$**@.25**	82.6 ±0.6	$\neg\mathbf{R_{LIME}}$**@.25**	69.3 ±1.5
$\mathbf{R_{LIME}}$**@.5**	83.8 ±0.8	$\neg\mathbf{R_{LIME}}$**@.5**	69.4 ±1.8
$\mathbf{R_{LIME}}$**@.75**	84.1 ±0.8	$\neg\mathbf{R_{LIME}}$**@.75**	68.4 ±2.0
SE	90.1 ±0.4	**¬SE**	78.1 ±0.9
U-SE	87.5 ±0.8	**¬U-SE**	72.8 ±1.1
Test set: Laptop + Restaurant Majority baseline: 65.8			

Table 1: Test set accuracy (average and standard deviation) and effect of restricting input to random tokens (R_{RAND}), gradient-based rationales ($R_{x\nabla x}$), LIME-based rationales (R_{LIME}), SEs, the union of SEs where a sentence has multiple SEs (U-SE), and masking all other tokens (¬) for 25%, 50% and 75% lengths in both training and test data. "None" masks the review sentence completely. The review domain, aspect entity type, aspect attribute and sentence length are available to all classifiers.

5.5 Experiment 3: Agreement of Automatic and Human Rationales

For a given relative rationale length, we measure its agreement with the SE in terms of f-score (F_1) of token-level I/O tags. F-score is calculated as the geometric mean of precision, i. e. the fraction of tokens in the rationale that are also in the SE, and recall, i. e. the fraction of tokens in the SE that are in the rationale. We exclude function words from the evaluation as we observe that saliency maps focus on content words while the human SE annotation includes function words, and, at least for English, it should be straightforward to expand a rationale to cover the relevant function words if desired.[9] Results are reported using relative rationale length as a parameter, i. e. plotting f-score over relative length.

6 Results

6.1 Baselines

Figure 1 and the top section of Table 1 show the baselines masking a random fraction of tokens. The classifier performance for full review sentences (**Full**) is shown in the top row (84.6%), along with the performance for masking all review tokens (69.3%). The latter is 3.5 points higher than for the classifier that chooses the majority label of positive for all input (65.8%, bottom row), showing that useful information can be found in the review domain, the aspect category and the sentence length (number of "[MASK]" tokens).

The second to fifth row of Table 1 show what happens when we randomly mask a fraction of the input in both the training and test data. The box plots in Figure 1 suggest a close to linear relationship between the fraction of masked tokens and the accuracy of the classifier. In the range explored (25% to 75%), the standard deviation is higher than for **Full** but not as high as for **None**.

6.2 Experiment 1: Masking LIME and Gradient Rationales

Figure 2 and the middle sections of Table 1 show what happens when restricting the input to rationales (R) of a fixed relative length. Compared to random masking, rationales succeed at selecting useful tokens for classification. The .75 threshold most closely mirrors the behaviour of the **Full** classifier. These rationales seem to cover close to all useful information for polarity prediction. With 0.5 and 0.25 relative length, however, we see clear differences between the two saliency methods. Here, LIME outperforms the gradient-based method.

The results for complement masks, i. e. masking the rationales instead of the non-rationale words, are shown in the right half of Table 1 and in the first and third group of three box plots in Figure 2. For LIME, the accuracy is close to the accuracy of **None** for all three thresholds, suggesting that LIME consistently assigns high saliency scores to informative words. The gradient-based method, however, appears to leak informative words to the third score quartile, giving $\neg R_{x\nabla x}$@.25 more useful information for classification than $\neg R_{x\nabla x}$@.5.

6.3 Experiment 2: Masking Human Rationales (SEs)

The bottom section of Table 1 shows the effect of masking the SEs or their complements. Masking all but the relevant SE helps the classifier (90.1%,+5.5) and masking the SE is harmful (78.1%,-6.5).

The ¬SE classifier is still performing 8.8 points above the classifier **None** where all review tokens are masked. A possible explanation are test items with multiple aspects as they will contain multiple SEs, and it could be that the classifier is helped by

[9] A simple heuristic would be to add all words between two rationale words if there is no content word between them. False positives between two selected phrase could be avoided with the help of a syntactic parser.

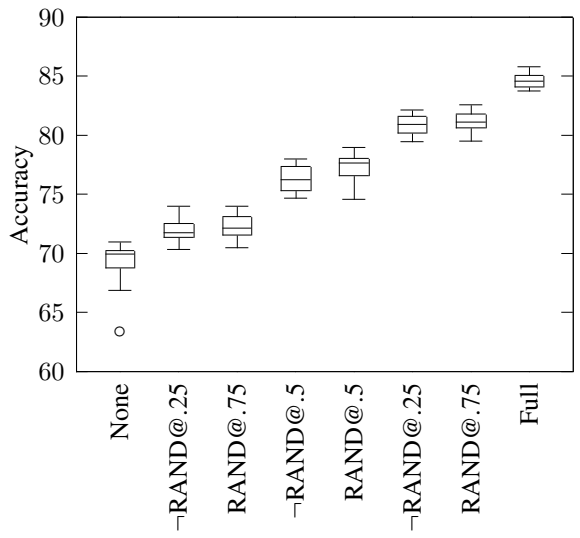

Figure 1: Accuracy distribution of sentiment polarity prediction with varying fraction of tokens of the review sentence randomly masked both at training time (same tokens masked in each epoch) and at test time. Due to rounding of the length in each sentence, the actual masked percentages from left to right are 100.0, 75.7, 74.2, 51.5, 48.5, 25.8, 24.3 and 0.0. 24 runs.

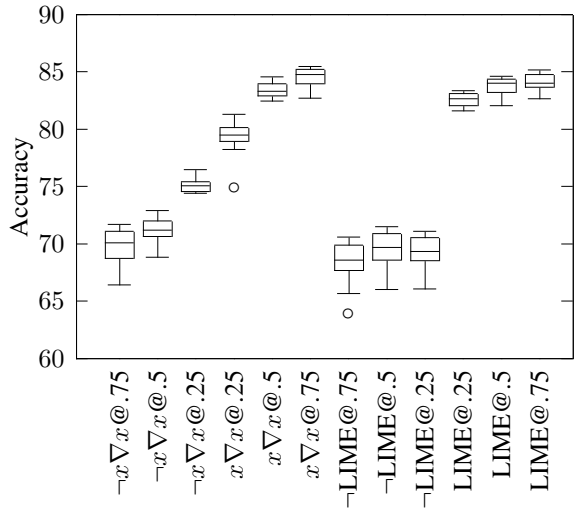

Figure 2: Accuracy distributions of sentiment polarity prediction with inputs masked according to gradient-based and LIME-based rationales and for relative rationale lengths 25%, 50% and 75%. 12 runs.

the presence of an SE for a different aspect which happens to have the same polarity, e.g.

Example 6. ***Great*** *food and the prices are* ***very reasonable***

In this example, the SE, *great*, could be masked but the other SE, *very reasonable*, for a different aspect (price) will remain. If we mask all SEs in a sentence regardless of whether they are relevant to the aspect (¬U-SE), the accuracy drops to 72.9 (-11.7), confirming that these "off-topic" SEs can indeed be helpful. Conversely, when all SEs of a test sentence are included in the input along with the SE in focus (U-SE), there is more noise and the accuracy, although still higher than the Full classifier (87.3%,+2.7) is lower than the SE classifier. These results suggest that the Full classifier does not just rely on sentiment indicators from the SE relevant to the target aspect, a strategy that can be helpful when multiple SEs have the same polarity (Ex. 6) but does not work when the polarities disagree (Ex. 1).

Even with all the SEs in the input masked, the accuracy of the ¬U-SE classifier is still above the **None** baseline (+3.8), indicating that the classifiers can pick up sentiment from outside the SEs. A large part of this difference (3.1 points) can be attributed to neutral instances where the classifier may have learned an association of empty SEs with a lack of sentiment.

6.4 Experiment 3: Agreement of Automatic and Human Rationales

Figure 3 shows f-score of rationales – measuring agreement with SEs as described in Section 5.1 – for the twelve classifiers trained without word masking (**Full**). Both $R_{x\nabla x}$ and R_{LIME} rationales only exceed the baseline of selecting all tokens as rationale (f-score 60.6 for this test set) by a small margin and not consistently over all runs. For $R_{x\nabla x}$ rationales, relative lengths just over 50% seem to work best for runs that do outperform the baseline. The curve for $R_{x\nabla x}$ shows only small changes in f-score when the rationale length is increased beyond 75%. This is likely caused by the preference for content words of the gradient-based saliency method, causing mostly function words to be picked last and these words do not count in our evaluation measure for agreement between R and SE. The picture is more complex for R_{LIME} rationales. While relative lengths just over 50% again are strong candidates for best agreement with

SEs, in some runs the best agreement is for lengths between 60% and 92%.

The sharp change of evaluation score at 25%, 50% and 75% may be related to the rounding of rationale lengths and a high number of short test sentences.[10] Future work may be able to avoid this distracting feature by stochastic rounding or carrying forward rounding errors to the next sentence when processing data instances in sequence.

Length Distribution and Length Oracle A wide range of SE lengths would be a possible explanation for low agreement with fixed-length rationales.[11] Figure 4 shows a preference for lengths between 0 and 30%. Lengths of 50% or more occur in 24.5% of test items. This length distribution certainly poses challenges for reaching high agreement with fixed length rationales but it is not clear to what extent. If we select an optimal rationale length for each test item, in other words, if we supply the rationale extraction with a length oracle, the f-score increases to 81.1 on average (range 79.5 to 82.6 over twelve runs) for $R_{x\nabla x}$ and to 84.2 (83.9 to 84.2) for R_{LIME}, suggesting that a fixed relative rationale length is not suitable for producing rationales that agree well with SEs.

Number of Spans A further difference between SEs and rationales is their distribution of the number of spans in each test item: 82.4% of SEs are continuous, 12.3% have two spans, 3.4% are empty and only one test item has more than three spans. For $R_{x\nabla x}$ with 50% relative length (highest f-score for agreement), however, only 8.3% of rationales are continuous, 14.9% have two spans, none are empty and 60.0% have more than three spans. R_{LIME} rationales are less fragmented. The respective numbers are 8.9%, 17.8% and 56.2%.

The lack of control over the number of spans produced by the rationalisation methods tested may therefore by a contributing factor to the poor agreement with human sentiment expressions. Limiting the number of spans or encouraging a low number of spans may improve agreement, potentially at the cost of no longer faithfully explaining the prediction of the classifier.

[10] At the pronounced step at 50% rationale length, the rationales of all odd-length test items gain one token in length as their raw rationale length switches from being rounded down to being rounded up.

[11] A longer SE forces len(SE)-len(R) false negatives. A shorter SE forces len(R)-len(SE) false negatives.

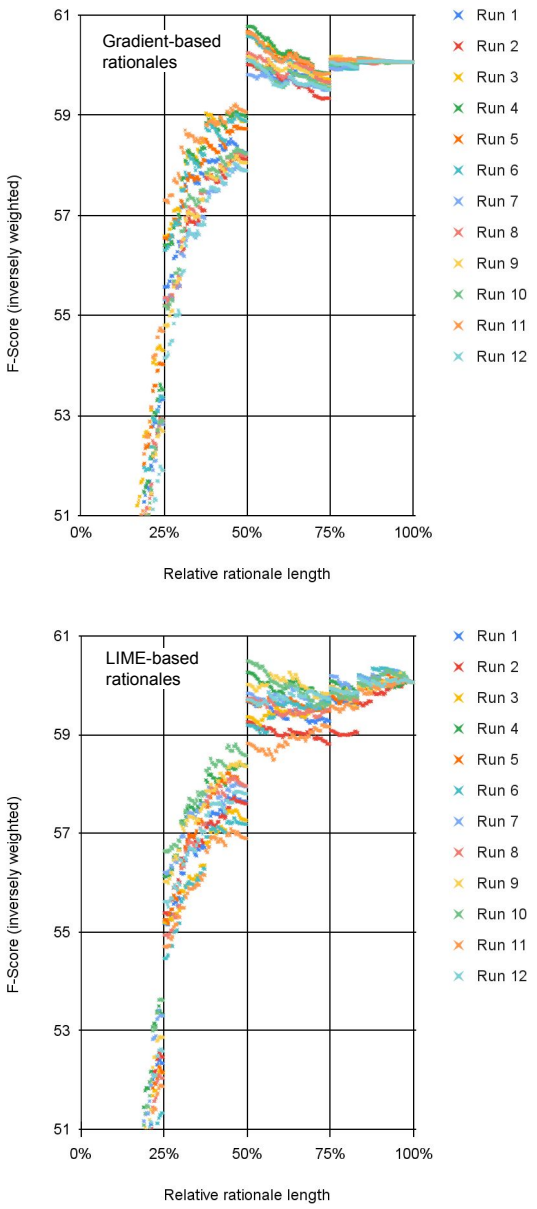

Figure 3: Agreement of the rationales with SEs for twelve sentiment classifiers; The x-axis is the rationale length and the y-axis is the inversely-weighted average f-score (F_1);

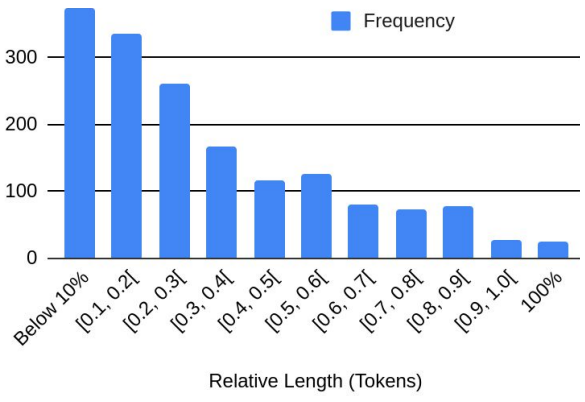

Figure 4: Sentiment expression length distribution

Table 2: Word clouds comparing sentiment expressions and rationales: Each token of the training and test data is sorted into one of the above four buckets and, for each bucket, a word cloud is created from the tokens in the bucket. The relative rationale length is 50% (@.5).

Table 3: Word clouds comparing sentiment expressions and rationales: As for Table 2 but for LIME.

Word Clouds Tables 2 and 3 show word clouds[12] for subsets of test tokens selected according to whether they belong to the SE and the rationale with 50% length. Tokens selected both as SE and rationale are dominated by sentiment words such as *great*, *love* and *best*, and the negators *not* and *n't*. Tokens selected by rationales but not by SEs seem to focus on aspect terms such as *food*, *laptop* and *service* for $R_{x\nabla x}$ but not as frequently for R_{LIME}. Aspect terms may be needed by the classifier to identify the correct SE. Also frequent in this set are the coordinating conjunctions *but* and *and* (again with different frequencies for the two types of rationales) and forms of *be*. The word clouds for tokens not selected by rationales (rightmost column) are dominated by function words.

[12] https://github.com/amueller/word_cloud

7 Related Work

Atanasova et al. (2020) compare various saliency methods on three text classification tasks (natural language inference and two non-aspect-based sentiment polarity datasets). They find that gradient-based rationales perform the best overall, and that LIME performs better than other non-gradient-based rationales. DeYoung et al. (2020) introduce a new benchmark, ERASER, which consists of seven datasets annotated with human rationales. They evaluate the faithfulness of various saliency methods on the benchmark datasets and compare the automatically extracted rationales to the human rationales. They find that LIME produces more faithful rationales than gradient or attention-based rationales, and that attention-based rationales are most likely to agree with human rationales. In our paper, we also evaluate faithfulness and agreement with human rationales, but on the ABSA task, using a basis the data that was used in the series of SemEval shared tasks from 2014 to 2016 (Pontiki et al., 2014, 2015, 2016). More recently, Bastings et al. (2022) compare the faithfulness of 16 gradient-based rationale extraction methods and seven variants of LIME-based rationale extraction with a focus on detection of shortcuts taken by a model for debugging the model, and Chrysostomou and Aletras (2022a) carry out a comparison of rationale extraction methods in out-of-domain settings.

Another relevant study to ours is that of Mousavi et al. (2022) who focus, not on English consumer reviews, but on Italian personal narratives. They compare gradient-based rationales of a sentiment polarity classifier to words identified by human annotators as "emotion carriers". Emotion carriers are defined following Tammewar et al. (2020) as "entities or actions that explain, cause or carry the emotion" (Mousavi et al., 2022, p. 62). These are distinguished from "emotion-laden words" that "explicitly express [...] sentiment polarity". They find that the rationales focus on emotion-laden words rather than the emotion-carrier words. Although there is no direct correspondence, emotion carriers in narratives are related to aspect terms in product reviews and emotion-laden words to sentiment expressions. We find that the rationales focus on both sentiment expressions and aspect terms.

Another ABSA dataset is the dataset of beer reviews with scores for particular aspects such as appearance and taste (McAuley et al., 2012). Sys-

tems trained on this dataset take as input a review and aspect, and the task is to locate the parts of the review related to the aspect and classify its sentiment polarity. In contrast, in systems trained on the dataset used in this work, the input is an individual sentence in a review whose aspect has already been identified, and the task is to classify its sentiment polarity. To test the former type of system, McAuley et al. (2012) created a small test set where each sentence in a review has been annotated with a particular aspect. These annotations serve as human/gold rationales in (Antognini and Faltings, 2021; Paranjape et al., 2020; Yu et al., 2021; Guerreiro and Martins, 2021). In our study, the original SemEval dataset already contains sentence-level aspect annotations and the human/gold rationales are the result of a further layer of annotation which highlights the sentiment expressions within these sentences. For example, the entirety of the following sentence from the beer test set would be used as a human/gold rationale (for the "overall" aspect).

Example 7. *I've had several beers of this tripel/IPA blend variety and haven't really been taken by any of them.*

whereas in our study, the human/gold rationale would be *haven't really been taken by any of them.* or its equivalent in the SemEval data. Each represents a different and valid perspective on ABSA.

The recent survey of Bibal et al. (2022) discusses explanations and evaluation of their faithfulness. User-centric explanations that try to produce what the user expects are distinguished from technical explanations that try to explain what the model does. This distinction is useful, and relates to a interesting question emerging from our study – to what extent should the sentiment expressions and the LIME/gradient rationales agree? Atanasova et al. (2020) observe that faithful rationales do not necessarily agree with human rationales, and comment that faithfulness and agreement with human rationales are two distinct properties associated with rationales that should not be conflated. Similarly, Carton et al. (2020) find that human rationales do not "have high sufficiency and comprehensiveness" when evaluating human rationales as rationales for model predictions. They attempt to address this discrepancy by introducing derived evaluation metrics that re-scale sufficiency and comprehensiveness according to performance with full information and with no information.

8 Conclusion

Using automatic rationales determined by a black-box and a gradient-based saliency method, and sourcing human rationales for a popular English ABSA dataset, we have thoroughly explored the behaviour of a BERT model fine-tuned for this task.

In answer to our first research question, we find that both saliency methods produce rationales with similar faithfulness but LIME produces more comprehensive rationales. In answer to our second research question, we find that the words in the human rationales are not being used to their full potential in our BERT-based classifier. When they are used in isolation, performance improves by 5.5%. In answer to our third research question, we do not find a high level of agreement between the human and the automatic rationales. Some of the differences can be accounted for by the fixed length of the rationales, content words related to the aspect and the continuous nature of the human rationales.

Although Kaljahi and Foster (2018) report negative results with joint learning of polarities and sentiment expressions on this dataset, the promising classifier results when all but the relevant SEs are masked suggest that ABSA systems should try to learn these prior to or in parallel with learning the polarities. The improved agreement between the human rationales and the oracle length automatic rationales suggests that future work should also explore instance-specific rationale lengths, e. g. using methods of Chrysostomou and Aletras (2022b).

Limitations

Masking Accuracy of classifiers reported in this work is for masking both the training and the test data for reasons explained in Section 5.1. Each result is for a different type of mask and hence for a different test set. Our results cannot be used to gauge performance for unmasked or differently masked inputs that can be expected in applications.

Rationale Evaluation with ROAR To evaluate rationales, ROAR uses the performance of a model trained and tested with input masked according to the rationales. The reported numbers therefore do not only reflect the quality of the rationales but also the difficulty of the task, the size of the training data and the performance of the machine learning method. Furthermore, the measurements are influenced by randomness in training as the masking of training data changes the path of the

optimisation process.[13] The values have to be seen relative to baseline performance of **Full** and $\mathbf{R_{RAND}}$ and in comparison to different types of rationales.

Domains The experiments are restricted to the two domains of the dataset, namely restaurant and laptop reviews, with just 28 aspect entity types and 14 attribute labels. We encoded these in a shared vocabulary with the review sentences.[14] We did not explore alternatives such as using reserved embedding table entries to encode the domain and aspect categories (so that tuning these embedding table entries does not affect the embedding of the tokens of the review sentence) or using more natural question templates, e. g. adding function words where appropriate and lower-casing the categories. Performance differences may change for other domains, number of aspect categories and the ratio of the training size of smallest domain and largest domain (in our work 4:5).

Task The ABSA task (see Section 1) assumes that the aspect category is already marked in the input and labelled with entity type and aspect category.

Number of Test Scores On first sight, the high number of test scores could be a concern as testing many models on test data can lead to overfitting to test data. However, only the result for the **Full** setting is a vanilla test set result. All remaining results are testing on derived (masked) test sets matching the masking applied to the training data. Therefore, these results do not leak performance information for building better classifiers on the test data. Using the test set here is convenient as the data set does not come with a validation set and the validation set we held out from the training data for selecting the training epoch is very small.

Language Experiments are for English only due to availability of SEA data. Various factors may cause different patterns for other languages, e. g. *(a)* BERT subword units, *(b)* evaluation excluding function words vs. languages that use mostly morphology instead, *(c)* freer word order may result in annotators producing more discontinuous SEs.

[13] Our reporting of averages over twelve runs, and in some cases 24 runs, compensates for the latter effect as each run shuffles the order of the training data and uses a new random initialisation of the classification head.

[14] For the aspect entity type and attribute label, sharing is reduced by using capital letters.

Ethics Statement

Kaljahi and Foster (2018) provide the annotation of sentiment expressions for the ABSA dataset, which we use as human rationales in our experiments. The annotation was carried out by expert annotators, annotation guidelines are provided with the dataset and high inter-annotator agreement was reported.

While our work uses sentiment analysis as a task of study, we do not propose improvements or changes in how this task is addressed, nor do we propose new methods for producing rationales. Therefore, we do not see any new ethical considerations compared to the approach we build on if the approaches described in this work are deployed. As to other insights and methods for understanding model behaviour presented in this study, we do not see an obvious way how these could cause harm: The intended use is by a researcher or engineer to identify issues with an existing predictive model, in our study a sentiment polarity classifier taking a single review sentence as input, and to gain actionable insights to improve the model and/or explanations. If the technology is functioning as intended this may be beneficial to the business deploying a model and/or explanation method, as well as to their customers, depending on costs and magnitude of improvements. If improvements are substantial this can have hard to predict effects. For sentiment analysis of product reviews, improvements can increase trust in automatically aggregated reviews of products and service.

A potential misuse of the results of this and similar studies is to use the methods to exaggerate the quality of explanations and the trustworthiness of predictions of a model.

The compute budget for this work is dominated by running the sentiment polarity classifiers in inference mode for LIME. For this step, we spent 93 GPU days on a fairly balanced mixture of NVIDIA GeForce RTX 2080 Ti and NVIDIA Quadro RTX 6000 cards. An additional 63 CPU hours was spent on CPU-only machines (2x Intel Xeon E5-2620 v4, 16 cores) to deduplicate LIME queries[15] and to run the LIME explainer. Model training took less than nine GPU days. An additional five to ten GPU days was used during development.

[15] The reduction in compute budget from deduplication is largest for short inputs. To let other user benefit from deduplication, we plan to submit a feature request and a workaround to the LIME project after acceptance.

Acknowledgements

This research is supported by Science Foundation Ireland (SFI) through the SFI Frontiers for the Future programme (19/FFP/6942) and the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development. We thank the reviewers for their insightful and helpful comments.

References

Diego Antognini and Boi Faltings. 2021. Rationalization through concepts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 761–775, Online. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(61):1803–1831.

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.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022a. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022b. Flexible instance-specific rationalization of NLP models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10545–10553.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2015. Extraction of salient sentences from labelled documents. ArXiv 1412.6815v2.

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. 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.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1080–1095, Online. Association for Computational Linguistics.

Nuno M. Guerreiro and André F. T. Martins. 2021. SPECTRA: Sparse structured text rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

6534–6550, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Rasoul Kaljahi and Jennifer Foster. 2018. Sentiment expression boundaries in sentiment polarity classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 156–166, Brussels, Belgium. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. ArXiv 1611.07270v1.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025.

Seyed Mahed Mousavi, Gabriel Roccabruna, Aniruddha Tammewar, Steve Azzolin, and Giuseppe Riccardi. 2022. Can emotion carriers explain automatic sentiment prediction? a study on personal narratives. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 62–70, Dublin, Ireland. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. 2020. Annotation of emotion carriers in personal narratives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1517–1525, Marseille, France. European Language Resources Association.

Marcos Treviso and André F. T. Martins. 2020. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online. Association for Computational Linguistics.

Maria Mihaela Truşcă, Daan Wassenberg, Flavius Frasincar, and Rommert Dekker. 2020. A hybrid approach for aspect-based sentiment analysis using

deep contextual word embeddings and hierarchical attention. In *Web Engineering (20th International Conference on Web Engineering, ICWE 2020)*, volume 12128, pages 365–380, Cham. Springer International Publishing.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 12822–12835. Curran Associates, Inc.

A Model Hyperparameters

For the classification head, we use two hidden layers with dimension 1536 and 256 respectively and dropout layers with dropout 0.2, 0.5 and 0.1 around and between the hidden layers. Including this classification head, the model has 111 million parameters, adding 1.6 million parameters to the base model. We fine-tune $\text{BERT}_{\text{BASE}}$ with a learning rate of 0.00001 and train the classification head with a learning rate of 0.00003. We train with a batch size of eight on 10.8 GB NVIDIA RTX 2080 Ti and of 16 on 23.7 GB NVIDIA Quadro RTX 6000 GPUs, accumulating the gradients of eight or four batches respectively (virtual batch size of 64).

B Examples of Sentiment Expressions and Rationales

Bold = SE, underline = R@.5, showing a random sample of 18 training instances per domain. (If the last sentence sampled has multiple opinion annotations in the data set all respective items are included, potentially showing more than 18 examples.)

B.1 Gradient-Based Rationales

B.1.1 Laptop Domain

First the screen **goes completely out** .

HP is **more interested in selling extended warranties** (which cost more than the netbook new) then they are in helping or fixing .

HP is more interested in selling extended warranties (which **cost more than the netbook new**) then they are in helping or fixing .

It **did not take long** to get used to the Mac OS .

My **favorite** part of this computer is that it **has a vga port** so I can connect it to a bigger screen .

My favorite part of this computer is that it has a vga port so I **can connect it** to a bigger screen .

2 months later , the battery went .

Probably **as good as you can get** in a netbook , **does everything I ask for** and has some very good unexpected pluses .

Probably as good as you can get in a netbook , does everything I ask for and has **some very good unexpected pluses** .

Seriously , **save yourself the hassle** and **purchase from a different company** .

YOU **WILL NOT BE ABLE TO TALK TO AN AMERICAN** WARRANTY SERVICE IS OUT OF COUNTRY .

I **am enjoying it** and the quality it provides is great !

I am enjoying it and the quality it provides is **great** !

the features are **great** , the only thing it needs is better speakers .

the features are great , **the only thing it needs is better speakers** .

Here we are **another year later and the computer is doing the same thing** .

They **loved it** .

LOVE IT LOVE IT LOVE IT !

B.1.2 Restaurant Domain

Great find in the West Village !

We ate at this Thai place following the reviews but **very unhappy** with the foods .

The hostess and the waitress were **incredibly rude** and **did everything they could to rush us out** .

The hostess and the waitress were **incredibly rude** and **did everything they could to rush us out** .

Very , very nice

Great sake !

We were seated outside and the waiter **spilled red wine and hot tea on myself and my date** .

So close , but **not good enough** .

This is **the kind of place you 'd like to take all your friends to and still keep a secret** .

Nevertheless , I finished my plate , and that 's when I **found a maggot in mushroom sauce** at the bottom .

This is a **nice** restaurant if you are looking for a **good** place to host an intimate dinner meeting with business associates .

All the staff is **absolutely professional** !!

And **amazingly cheap** .

Our food was **great** too !

All in all we 're **already coming up with excuses to go ahead really soon in the next few wks** !!!!

The restaurant is **cute but not upscale** .

As a Japanese native , I 've lived in the Tristate area for over 8 years , but I was just **so amazed** at this place .

$ 20 for all you can eat sushi **can not be beaten** .

B.2 LIME-Based Rationales

B.2.1 Laptop Domain

First the screen **goes completely out** .

HP is **more interested in selling extended warranties** (which cost more than the netbook new) then they are in helping or fixing .

HP is more interested in selling extended warranties (which **cost more than the netbook new**) then they are in helping or fixing .

It **did not take long** to get used to the Mac OS .

My **favorite** part of this computer is that it **has a vga port** so I can connect it to a bigger screen .

My favorite part of this computer is that it has a vga port so I **can connect it** to a bigger screen .

2 months later , the battery went .

Probably **as good as you can get** in a netbook , **does everything I ask for** and has some very good unexpected pluses .

Probably as good as you can get in a netbook , does everything I ask for and has **some very good unexpected pluses** .

Seriously , **save yourself the hassle** and **purchase from a different company** .

YOU **WILL NOT BE ABLE TO TALK TO AN AMERICAN** WARRANTY SERVICE IS OUT OF COUNTRY .

I **am enjoying it** and the quality it provides is great !

I am enjoying it and the quality it provides is **great** !

the features are **great** , the only thing it needs is better speakers .

the features are great , **the only thing it needs is better speakers** .

Here we are **another year later and the computer is doing the same thing** .

They **loved it** .

LOVE IT LOVE IT LOVE IT !

B.2.2 Restaurant Domain

Great find in the West Village !

We ate at this Thai place following the reviews but **very unhappy** with the foods .

The hostess and the waitress were **incredibly rude** and **did everything they could to rush us out** .

The hostess and the waitress were **incredibly rude** and **did everything they could to rush us out** .

Very , very nice

Great sake !

We were seated outside and the waiter **spilled red wine and hot tea on myself and my date** .

So close , but **not good enough** .

This is **the kind of place you 'd like to take all your friends to and still keep a secret** .

Nevertheless , I finished my plate , and that 's when I **found a maggot in mushroom sauce** at the bottom .

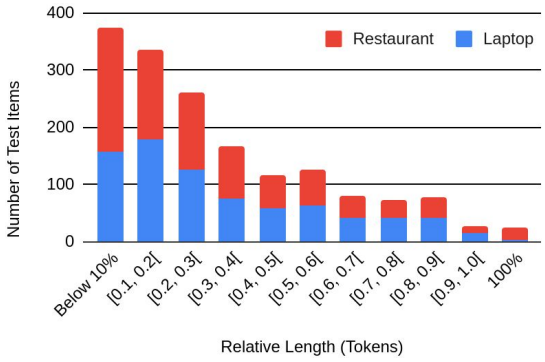

Figure 5: Sentiment expression length distribution broken down by domain

This is a **nice** restaurant if you are looking for a **good** place to host an intimate dinner meeting with business associates .

All the staff is **absolutely professional** !!

And **amazingly cheap** .

Our food was **great** too !

All in all we 're **already coming up with excuses to go ahead really soon in the next few wks** !!!!

The restaurant is **cute but not upscale** .

As a Japanese native , I 've lived in the Tristate area for over 8 years , but I was just **so amazed** at this place .

$ 20 for all you can eat sushi **can not be beaten** .

C SE Length Distribution by Domain

Figure 5 shows the domain breakdown for Figure 4. The SE length distribution is similar across the two domains. For relative lengths of 100%, we see a higher number of examples in the restaurant domain (22) than in the laptop domain (3).

D Word Cloud Tables

Tables 4 and 5 show the average frequency and rank of selected tokens in each of the word clouds of Tables 2 and 3.

E Indentification of Function Words

For the exclusion of function words from the evaluation, we use a word list extracted from English UD treebanks based on part of speech to identify function words. We include symbols and punctuation as function words as they are not content words.

F Evaluation Metrics

Accuracy Accuracy is the number of correctly classified test instances divided by the total number of test instances. This is implemented in `https://github.com/redacted/absa-rationale-eval/blob/main/scripts/train-classifier.py` per batch in line 1303 and accuracies of batches are accumulated in line 1364 weighted by batch size.

F-Score Further detail to the description of f-score F_1 in Section 5.5 is that we weight each event (false positive, false negative, true positive and true negative) inversely to the sentence length it occurs in. This way each test sentence makes the same contribution to the overall score, not letting a small number of long sentences dominate the results. Differently to a macro average, the relationship of f-score being the geometric mean of precision and recall is maintained. The implementation of this metric is in lines lines 225 to 302 of of `https://github.com/redacted/absa-rationale-eval/blob/main/scripts/evaluation.py`. (Currently, this code is duplicated in `train-classifier.py`.)

G Dataset License

While Kaljahi and Foster (2018) do not include a license file in their data repository (Footnote 2), they describe intended use of the data by other researchers in the paragraph "The dataset reported in this work" in their conclusions. They did not anticipate the use for interpretability methods but we believe that our work is covered by "extraction of sentiment expressions" and "linguistic insights".

Word	**SE** $\cap$ $\mathbf{R}_{x\nabla x}$		**SE** $\cap$ $\neg\mathbf{R}_{x\nabla x}$		$\neg$**SE** $\cap$ $\mathbf{R}_{x\nabla x}$		$\neg$**SE** $\cap$ $\neg\mathbf{R}_{x\nabla x}$	
Top 10 **SE** $\cap$ $\mathbf{R}_{x\nabla x}$ words (4634 types, 263454 tokens)								
great	489.8	1	35.6	67	404.9	10	43.1	110
n't	460.7	2	11.3	131	309.2	14	13.4	252
not	458.7	3	81.3	34	279.3	15	64.7	87
good	392.7	4	32.7	74	249.4	16	47.9	103
very	348.4	5	106.2	28	209.9	21	88.8	75
love	175.0	6	7.7	178	68.7	110	2.0	820
a	155.4	7	1044.6	2	215.2	18	1976.8	7
no	149.6	8	21.1	92	120.1	47	18.6	212
really	148.8	9	21.9	90	94.0	69	23.3	191
have	146.2	10	197.8	13	142.2	36	401.8	25
Top 10 **SE** $\cap$ $\neg\mathbf{R}_{x\nabla x}$ words (1301 types, 133866 tokens)								
the	125.8	16	1578.2	1	170.0	27	4428.7	3
a	155.4	7	1044.6	2	215.2	18	1976.8	7
to	115.6	23	983.1	3	132.8	40	1764.6	9
it	78.8	54	637.2	4	147.0	32	1798.3	8
of	46.3	106	600.3	5	76.3	99	1159.7	11
and	39.1	130	439.6	6	393.3	12	4130.7	4
,	44.2	111	394.4	7	552.9	4	6044.4	2
for	82.6	46	368.1	8	185.4	23	1118.6	12
I	41.8	116	295.6	9	397.6	11	2729.1	5
this	62.7	71	262.7	10	107.1	54	687.6	18
Top 10 $\neg$**SE** $\cap$ $\mathbf{R}_{x\nabla x}$ words (5551 types, 466194 tokens)								
.	4.2	1111	7.8	176	1435.3	1	6418.0	1
but	89.1	42	40.2	61	683.2	2	235.4	32
food	79.2	53	30.1	78	587.8	3	189.6	42
,	44.2	111	394.4	7	552.9	4	6044.4	2
is	47.0	103	143.7	20	520.0	5	2481.3	6
laptop	124.4	17	0.9	728	489.0	6	16.3	224
was	52.0	89	162.7	18	439.1	7	1048.9	13
service	45.2	109	4.1	254	431.4	8	45.9	108
!	0.4	4554	0.9	751	424.6	9	459.4	24
great	489.8	1	35.6	67	404.9	10	43.1	110
Top 10 $\neg$**SE** $\cap$ $\neg\mathbf{R}_{x\nabla x}$ words (2022 types, 552798 tokens)								
.	4.2	1111	7.8	176	1435.3	1	6418.0	1
,	44.2	111	394.4	7	552.9	4	6044.4	2
the	125.8	16	1578.2	1	170.0	27	4428.7	3
and	39.1	130	439.6	6	393.3	12	4130.7	4
I	41.8	116	295.6	9	397.6	11	2729.1	5
is	47.0	103	143.7	20	520.0	5	2481.3	6
a	155.4	7	1044.6	2	215.2	18	1976.8	7
it	78.8	54	637.2	4	147.0	32	1798.3	8
to	115.6	23	983.1	3	132.8	40	1764.6	9
The	4.7	1092	58.0	47	65.6	118	1547.8	10

Table 4: Average frequency (over twelve runs) and rank of top 10 words in each word cloud of Table 2 for $\mathrm{R}_{x\nabla x}$

Word	$\mathbf{SE} \cap \mathbf{R_{LIME}}$		$\mathbf{SE} \cap \neg\mathbf{R_{LIME}}$		$\neg\mathbf{SE} \cap \mathbf{R_{LIME}}$		$\neg\mathbf{SE} \cap \neg\mathbf{R_{LIME}}$	
Top 10 $\mathbf{SE} \cap \mathbf{R_{LIME}}$ words (4564 types, 255284 tokens)								
the	519.2	1	1184.8	1	964.9	4	3633.8	3
great	517.8	2	7.6	275	437.7	12	10.3	491
not	498.4	3	41.6	55	297.6	25	46.4	116
a	433.8	4	766.2	2	601.0	9	1591.0	5
good	404.8	5	20.6	102	284.3	26	13.0	399
n't	400.1	6	71.9	33	250.3	29	72.3	87
very	376.8	7	77.9	27	225.0	33	73.7	84
it	358.6	8	357.4	5	942.8	5	1002.6	10
to	332.7	9	766.0	3	410.9	13	1486.4	7
and	299.2	10	179.4	10	3362.8	1	1161.2	8
Top 10 $\mathbf{SE} \cap \neg\mathbf{R_{LIME}}$ words (3272 types, 142036 tokens)								
the	519.2	1	1184.8	1	964.9	4	3633.8	3
a	433.8	4	766.2	2	601.0	9	1591.0	5
to	332.7	9	766.0	3	410.9	13	1486.4	7
of	201.2	12	445.4	4	301.9	24	934.1	11
it	358.6	8	357.4	5	942.8	5	1002.6	10
,	110.0	32	328.7	6	1625.7	2	4971.7	2
for	194.3	13	256.3	7	404.2	14	899.8	12
I	101.9	40	235.4	8	742.7	6	2384.0	4
in	83.6	55	191.1	9	191.8	38	594.9	14
and	299.2	10	179.4	10	3362.8	1	1161.2	8
Top 10 $\neg\mathbf{SE} \cap \mathbf{R_{LIME}}$ words (5431 types, 474364 tokens)								
and	299.2	10	179.4	10	3362.8	1	1161.2	8
,	110.0	32	328.7	6	1625.7	2	4971.7	2
is	97.1	41	93.6	22	1494.9	3	1506.4	6
the	519.2	1	1184.8	1	964.9	4	3633.8	3
it	358.6	8	357.4	5	942.8	5	1002.6	10
I	101.9	40	235.4	8	742.7	6	2384.0	4
but	106.4	37	22.9	92	728.3	7	190.3	36
was	108.3	35	106.3	19	684.8	8	803.2	13
a	433.8	4	766.2	2	601.0	9	1591.0	5
The	22.6	216	40.1	57	488.9	10	1124.4	9
Top 10 $\neg\mathbf{SE} \cap \neg\mathbf{R_{LIME}}$ words (4562 types, 544628 tokens)								
.	2.8	1327	9.2	225	233.6	30	7619.8	1
,	110.0	32	328.7	6	1625.7	2	4971.7	2
the	519.2	1	1184.8	1	964.9	4	3633.8	3
I	101.9	40	235.4	8	742.7	6	2384.0	4
a	433.8	4	766.2	2	601.0	9	1591.0	5
is	97.1	41	93.6	22	1494.9	3	1506.4	6
to	332.7	9	766.0	3	410.9	13	1486.4	7
and	299.2	10	179.4	10	3362.8	1	1161.2	8
The	22.6	216	40.1	57	488.9	10	1124.4	9
it	358.6	8	357.4	5	942.8	5	1002.6	10

Table 5: Average frequency (over twelve runs) and rank of top 10 words in each word cloud of Table 2 for R_{LIME}

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section on pages 9-10

☑ A2. Did you discuss any potential risks of your work?
Ethics statement section on page 10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix G

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix G

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The dataset is a standard dataset for the task

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2 + Limitation section on pages 9-10

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2

C ☑ Did you run computational experiments?

3+4+5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A + Ethics statement

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Setup: Section 5; no hyper-parameter search has been performed; the hyper-parameter choices based on previous work and ad-hoc choices are reported in Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.