Investigating the saliency of sentiment expressions in aspect-based sentiment analysis

Anonymous ACL submission

Abstract

We examine the behaviour of an aspect-based sentiment classifier built by fine-tuning the BERT$_{\text{BASE}}$ model on the SemEval 2016 English dataset. In a set of masking experiments, we examine the extent to which the tokens which express the sentiment towards the aspect are being used by the classifier. The enhanced performance of a classifier that only sees the relevant sentiment expressions suggests that they are not being used to their full potential. Furthermore, sentiment expressions which are not directly relevant to the aspect in focus also appear to be used. We then use a gradient-based method to identify the most salient words. A comparison of these salient words, or 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

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*

an ASBA system should return the polarity `positive` for the “location” aspect and `negative` for the “service” aspect. Much work in ASBA 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). Recent approaches (Sun et al., 2019; Trusca et al., 2020) are underpinned by pretrained models such as BERT (Devlin et al., 2019).

Kaljahi and Foster (2018) provide an additional layer of annotation on the English SemEval 2016 laptop and restaurant data by explicitly marking the span of the sentiment expressions, i.e. those words deemed by two human judges to be actually expressing the sentiment towards the aspect. In Ex. 1, the sentiment expression (SE) for the “location” aspect is *love where it is located*. For the “service” aspect, the SE is *leaves much to be desired*. We demonstrate that SEs are useful to our BERT-based ABSA system by showing that performance degrades when they are masked from the input and improves when all but the SE tokens are masked. We also show that SEs that are not referring to the aspect in focus are also being used, for better or worse.

Saliency approaches to understanding the output of a system attempt to locate the parts of the input that contribute the most to the system’s decision. We investigate whether the rationale (the most salient words) according to a gradient-based approach corresponds to the gold labelled SEs. The highest overlap with the SEs (approx 50%) is obtained by a rationale length of 50% of the input sequence. Some of the words that are being used by the classifier that are not in the SEs 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 salient. We also note structural differences – the SEs have been mostly annotated as continuous spans whereas the rationales do not have this property; the rationales are a fixed proportion of the input length whereas the SEs are not.

2 Dataset

We use the SemEval 2016 English ASBA dataset, focusing on SubTask B where the label to be predicted is the sentiment polarity, 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 to this dataset by marking the spans of the SEs in each sentence.

3 Sentiment Classifier

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

Example 2. **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, prefixing the question with a domain label to help the model to adjust to domain-specific patterns. Hyperparameters are detailed in the appendix.

4 Saliency Method

For our experiments, we employ a gradient-based saliency method as recent discussion suggests caution with popular attention-based saliency methods (Jain and Wallace, 2019; Bastings and Filippova, 2020) and as results of Chrysostomou and Aletras (2021) show superior faithfulness of gradient-based rationales over attention-based rationales for seven of eight settings tested. Using the gradient for saliency has been introduced to NLP by 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. The basic method described by Li et al. (2016) uses the absolute value of the derivative directly as a saliency score. We implement a variant of gradient-based saliency for the input point P that combines the idea of integrating the gradient over a line (Sundararajan et al., 2017) with using the average gradient in a sample of points from a Gaussian distribution centred around P (Smilkov et al., 2017). We integrate from $0.95P$ to $1.05P$, using 15 equidistant points.[1]

5 Experiments

5.1 Setup

We first check what happens if we provide the classifier with 1) just the SEs identified by Kaljahi and Foster (2018), 2) just the rationales identified by the saliency method and 3) random subsets of the input tokens. We then examine how well the rationale tokens align with the SEs.

Experiment 1: Word Masking We derive new training, development and test sets, with either all but the target words masked or all the target words masked. This modification is restricted to the review sentences, i. e. sequence A is never masked. We train the following:

- **Full**: Sequence B is the review sentence
- **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)
- **R@**L and ¬**R@**L: We obtain rationales with relative length L, $L \in \{0.25, 0.5, 0.75\}$ by applying the saliency method to the **Full** classifier, and keep only the tokens of the rationale (R) or its complement (¬R), e. g. $L = 0.25$ means that one quarter of input tokens are selected as rationale.
- **A@**L and ¬**A@**L: For each item in the data, we randomly select a fraction L of tokens, $L \in \{0.25, 0.5, 0.75\}$.

Below is an example with the SE (in bold) and its complement masked.

Example 3.

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] .*

We train nine classifiers for ten epochs, selecting the epoch with highest accuracy according to development data. In each setting, we use the same set of nine seeds for random initialization.

[1] This limits the maximal relative distance of samples from P and therefore can use a small number of samples without the noise that sampling from a Gaussian distribution would produce, especially in a high dimensional space.

Target	Training and Test Setting	
	SE/R/A	¬SE/¬R/¬A
SE	89.8 ±0.4	77.8 ±0.5
U-SE	86.5 ±1.2	72.4 ±1.3
R@.25	79.8 ±0.8	74.4 ±0.8
R@.5	82.8 ±0.7	70.6 ±0.9
R@.75	84.1 ±0.9	69.0 ±2.2
A@.25	72.5 ±0.8	81.1 ±1.2
A@.5	76.9 ±1.1	76.6 ±0.9
A@.75	81.6 ±1.3	71.1 ±1.5

Test set: Laptop + Restaurant
Majority baseline: 65.8
Unmasked (Full): 84.2 ±0.6

Table 1: Test set accuracy (x100, average and standard deviation over nine runs) and effect of restricting input to SEs, the union of all SEs where a sentence has multiple opinions (U-SE), rationales (R), random tokens (A) and masking all other tokens (¬SE, ¬R, and ¬A) for 25%, 50% and 75% lengths.

Experiment 2: Agreement of Saliency Method with SEs For each possible rationale length k, we select the top-k tokens[2] according to the saliency method as the rationale of the prediction and measure its agreement with the SE in terms of precision, recall and f-score of token-level I/O tags. Precision is the fraction of tokens in the rationale that are also in the SE, and recall is the fraction of tokens in the SE that are in the rationale.[3] 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. Results are reported using relative rationale length as a parameter, e. g. the result for 20% rationale length is obtained by evaluating for each test item the rationale for which its length divided by the input length is closest to 20%.

5.2 Results

Experiment 1: Word Masking Table 1 shows the word masking results. The classifier performance for full review sentences (Full) is shown at the bottom (84.2%), along with the performance of the classifier that chooses the majority label of positive (65.8%) The top row shows the effect of masking the SEs or their complements. Masking all but the relevant SE helps the classifier (89.8%,+5.6) and masking the SE is harmful (77.8%,-6.4).

The ¬SE classifier is still performing 12 points above the majority classifier. Test items with multiple aspects will contain multiple SEs, and it could be that the classifier is helped by the presence of an SE for a different aspect which happens to have the same polarity, e.g.

Example 4. ***Great*** *food and the prices are* ***very reasonable***

If we mask all SEs regardless of whether they are relevant to the aspect (¬U-SE), the accuracy drops to 72.4 (-11.8), 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 (86.5%,+2.3) is lower than the SE classifier (-3.3). 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. 4) but doesn't 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 majority baseline (+6.6), indicating that the classifiers can pick up sentiment from outside the SEs. 3.1 points of this difference can be attributed to neutral instances where the classifier has learned an association of empty SEs with a lack of sentiment.

The middle and bottom sections of Table 1 show what happens when restricting the input to rationales (R) of a fixed relative length, and randomly masking input tokens (A). Compared to random masking, rationales succeed at selecting useful tokens for classification. The .75 length threshold most closely mirrors the behaviour of the Full classifier, but performance still falls short suggesting that some useful information is missing.

Experiment 2: Agreement of Saliency Method with SEs Figure 1 shows f-score of rationales, as described in Section 5.1, for the nine classifiers trained without word masking (Full). Any length from 50.0% to 50.6% is an optimal relative length

[2] When a token consists of two or more BERT tokens, we use the saliency score of the highest scoring BERT token.

[3] We report inversely weighted average f-score where we count each TP, FP, FN and TP only as $1/N$ instead of 1, where N is the input length of the test item in which the event occurs. This is similar to micro-average but each test item can contribute at most 1 to the four counts.

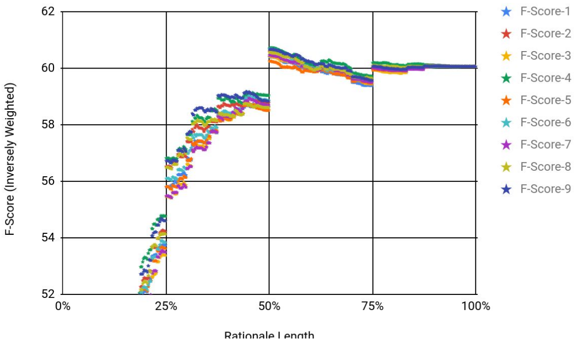

Figure 1: Agreement of the rationales with SEs for nine sentiment classifiers; The x-axis is the rationale length and the y-axis is the inversely-weighted average f-score

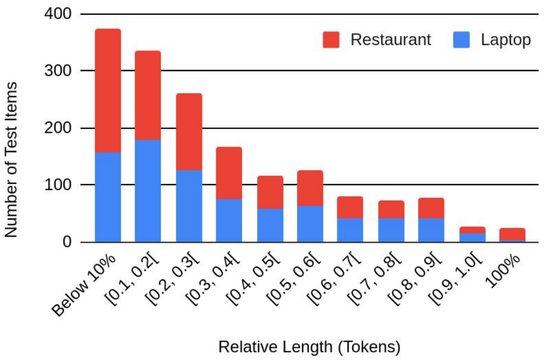

Figure 2: Sentiment expression length distribution

for all nine classifiers.[4] The curve 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 when measuring agreement.

A wide range of SE lengths would be a possible explanation for low agreement with fixed-length rationales. Figure 2 shows a preference for lengths between 0 and 30%. Lengths of 50% or more occur in 24.5% of test items. 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 82.3 (81.8 to 83.0 over all nine runs) suggesting that a fixed relative rationale length is not suitable for producing rationales that agree well with SEs.

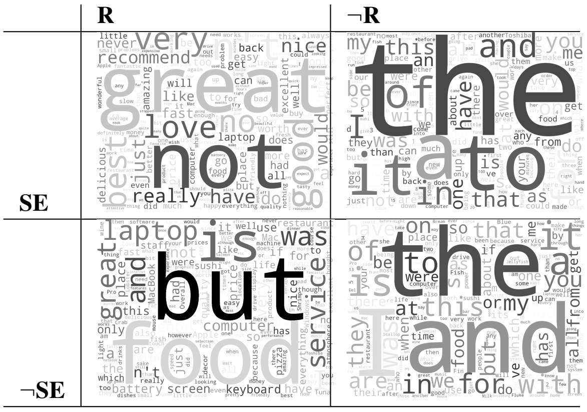

Table 2: Word clouds comparing sentiment expressions and rationales: Rationale length 50%.

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 rationales with 50% relative length (highest f-score for agreement), however, only 8.2% of rationales are continuous, 14.4% have two spans, none are empty and 60.7% have more than three spans.

Table 2 shows word clouds[5] 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 negator *not*. Tokens selected by rationales but not by SEs seem to focus on aspect terms such as *food*, *laptop* and *service*. These may be needed by the classifier to identify the correct SE. Also frequent in this set are the coordinating conjunction *but* and forms of *be*. The word clouds for tokens not selected by rationales (right-most column) are dominated by function words.

6 Conclusion

Using manually annotated SEs and rationales determined by a gradient-based saliency method, we have explored the behaviour of a BERT model fine-tuned on a popular English ABSA dataset. There is some overlap between the SEs and the rationales, and differences can be accounted for by the fixed length of the rationales, content words in the input related to the aspect and the continuous nature of the SEs compared to the rationales. The promising results when all but the relevant SEs are masked suggest that future systems should try to learn these prior to or in parallel with the polarities.

[4] At the pronounced step at 50% rationale length, the rationales of 52.0% of test items gain one token in length as for sentences with an odd number of tokens, the raw rationale length ends in .5 and switches from being rounded down to being rounded up.

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

References

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.

George Chrysostomou and Nikolaos Aletras. 2021. Flexible instance-specific rationalization of NLP models. ArXiv 2104.08219v2, accepted at AAAI22.

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.

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.

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.

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.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smooth-Grad: removing noise by adding noise. ArXiv 1706.03825v1.

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.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Maria Mihaela Trusca, Daan Wassenberg, Flavius Frasincar, and Rommert Dekker. 2020. A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention.

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. We fine-tune BERT with a learning rate of 0.00001 and train the classification head with a learning rate of 0.00003. We train for ten epochs, keeping BERT parameters frozen for the first three epochs, and select the best model according to development accuracy as the final model. We train with a batch size of eight on a 11 GB NVIDIA RTX 2080 Ti GPU, accumulating the gradients of eight batches (virtual batch size of 64). We did not tune the above hyper-parameters as we immediately outperformed the LSTM and CNN baselines of Kaljahi and Foster (2018), the gap to Sun et al. (2019)'s results can be explained with the combination of domains in a single model and differences between the 2014 and 2016 SemEval test sets, and our goal is to obtain a competitive classifier rather than one that outperforms the state-of-the-art.