Attention as a Model Explainer for Text Classification. Faithfulness and Plausibility

Abstract

This report investigates the utility of attention mechanisms as tools to explain neural network predictions through comparative experiments with SHAP, an alternative model-agnostic explanation methods. In a series of experiments, it examines whether Attention can be used to produce **plausible** and **faithful** explanations of BERT based text classifiers.

1 Introduction

When it comes to explaining predictions of Machine Learning Models, we can distinguish between *model agnostic* and *model specific* methods. An example of a model agnostic method is SHAP (SHapley Additive exPlanations), which can be applied to any machine learning model to calculate a *feature importance score*, i.e. the contribution that each feature has to the final prediction. But the calculation of SHAP values is computationally very expensive. A computationally less expensive alternative would be to read of feature importance scores directly from the model. In case of Neural Networks, a suitable candidate for this would be the (or an) attention layer, which can be understood as providing a view of where the model is "focusing," that *may* imply feature importance. This report explores attention's role in interpretability, with a specific focus on text classification. It investigates whether attention be reliably used as an importance measure. And compares its suitability for delivering explanations of predictions in comparison to SHAP.

2 Model-Agnostic Methods and SHAP

SHAP (SHapley Additive exPlanations) offers an explanation model grounded in Shapley values from cooperative game theory, which aims to measure each feature's contribution to the output. The Shapley value ϕ_i for each feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \big(v(S \cup \{i\}) - v(S)\big)$$

where N is the set of all features, S is any subset of N excluding i, and $v(S)$ represents the model output when only features in S are considered. As the calculation only involves the models predictions and the input features, it can be applied to any Machine Learning model.

SHAP is theoretically well motivated, as the equation of the SHAP values can be derived from a formalisation of principles such as accuracy, consistency and relevance [5]. However, the calculation of SHAP values is computationally very expensive: the calculation of one value involves iteration of over the power set of all the features. In practice, this is approximated, but the calculation of shapley values still remains very time intensive [2].

3 Attention as an Importance Measure

Given that attention mechanisms direct a model's focus, they might be considered as proxies for feature importance.

The basic set-up follows [1]: Focusing on classification tasks, I use $\texttt{bert} - \texttt{tiny}$[1], a pretrained light version of BERT from Huggingface, to which I append an attention layer and a classification layer. During training, only the final attention and classification layers are updated.

Attention-Based Explainer Attention is a probability distribution and therefore only contains positive values. But the explainer should also output negative values for features that count against a certain classification. In order to account for this, I use the following **Attention-Based Explainer** [1] which produces an attention-based explainability score $\hat{a}_t$ for each token t in a sequence is as:

$$\hat{a}_t = a_t \cdot \text{sign}(W_{\text{classifier}} \cdot (a_t h_t))$$

where a_t is the respective attention value, h_t the output of the previous layer, and $W_{\text{classifier}}$ the weights of the following classification layer.

This approach offers computational simplicity, as these values are calculated almost as a by-product from making the prediction. This is a decisive advantage over SHAP.

4 Evaluating Attention: Plausibility and Faithfulness

However, [8] and [4] have questioned whether attention is a reliable model explainer. They argue that it lacks two features that are required for intelligible explanations: *plausibility* and *faithfulness*.

Plausibility is the degree to which attention-based explanations pick out features in an intuitively understandable manner. While the determination of what my count as a plausible features is somewhat open, there are, arguably, some cases which don't count as plausible explanations, such as emphasis on POS tokens. More generally, [8], argue that an observed lack of correlation with other explainer methods such as SHAP effects plausibility negatively, as well.

Faithfulness represents the degree to which the a high feature importance score correlates with a high predictive value. [4] argues against attention explainers faithfulness by constructing adversarial distributions that yield identical predictions, but vastly different explanations. If two vastly different distributions can lead to very similar predictions, why should any of them have explanatory value?

The question is, though, why does the attention explainer lack plausibility and faithfulness?

5 Hypothesis

As suggested by [9], one reason for the possibility of adversarial attention distributions is that the attention mechanism *does not make a contribution to the model prediction in the first place.* This leads to the following

Hypothesis: If attention mechanisms contribute meaningfully to model predictions, then both faithfulness and plausibility should improve, thus providing a more reliable attention based explainer, as well.

Therefore, when investigating whether attention is a reliable model explainer, one should first check:

1. Does Attention make a contribution to predictions?

If the answer is positive, we can assess plausibility and faithfulness by looking into the following:

2. Does Attention mimic the classifier? (Plausibility)
3. Does Attention correlate with SHAP explanations? (Faithfulness)
4. Does Attention emphasize irrelevant tokens? (Faithfulness)

These questions suggest the following experiments:

[1] https://huggingface.co/prajjwal1/bert-tiny

Experiment 1: Relevance To determine if attention contributes to predictions, replace the attention module of a trained classifier with a uniform attention matrix and compare results. A significant performance drop (in practice, by more than 2 percent) would indicate that attention contributes to prediction.

Experiment 2: Interpretability The most important question is whether attention-based and SHAP-based explanations actually mimic the classifier's behaviour. This is also called *Fidelity* [3]. If the model predicts a certain class, say 'positive sentiment', then the feature importance scores of the explainer should reflect this, i.e. the explainer should highlight more positive feature importance scores than negative ones for the respective prediction instance. In order to see if that is the case, I sum up all the feature importance scores of the instance and check whether the resulting value is positive (matching the positive prediction).

$$y = \sum_i s_i$$

where s_i is the importance score of the i-th feature. Comparing the sum of the feature importance score with the output of the classifier lets us calculate the accuracy with which the explainer mimics the behaviour of the classifier.

For **SHAP** we can do something similar. But due to the nature of the SHAP value we have to add a baseline to the sum

$$p = \text{baseline} + \sum_i \varphi_i$$

and the explainer delivers a positive result if $p > 0.5$. The DeepSHAP explainer that is used to implement SHAP has as a baseline the average model prediction.

Experiment 3: Plausibility To assess the plausibility of attention, I calculate the correlation between attention and SHAP scores and calculate the average attention scores assigned to irrelevant tokens, such as punctuation, to see if they have negligible importance.

6 Diagnosis

The are three possible reasons why attention might not make a contribution to the prediction, which are connected to what happens *before*, *during*, and *after* the attention layer.

Before: If the vectors that are being fed into the Attention layer are highly similar, then which of those is emphasised by the attention layer has little effect on the resulting prediction. [6] call this "conicity" (similar vectors form a narrow cone in vector space).

During: Attention is a probability distribution and therefore dense. If every feature receives a value, then it might be that a single attention layer is not able to make a decisive emphasis on particular positions.

After: If the task is very simple, like a binary classification, then there a many ways of aggregating the outputs of the attention layer into a single number. This suggest that with more complex tasks, attentions suitability as an explainer might rise [9].

If the hypothesis that attention might have explanatory value *if* it makes a contribution to the prediction is correct, then improving one of those features should result in a better interpretation of attention as a model explainer. In another round of experiments, this hypothesis is tested by focusing on the attention layer itself.

Experiment 4: Sparsity Replace the attention layer with a sparse attention [7], by replacing softmax with sparsemax. Sparsemax produces sparse probability distributions by solving the following optimization:

$$\Pi_\Omega(z) = \underset{p \in \Delta^L}{\operatorname{argmax}} \left(z^\top p - \frac{1}{2}\|p\|_2^2 \right)$$

where:

- $z \in \mathbb{R}^L$ is the input vector, or logits, representing scores for each element,
- p is the resulting probability distribution,
- $\Delta^L = \{p \in \mathbb{R}^L \mid p \geq 0, \sum p_i = 1\}$ is the L-dimensional probability simplex.

This formulation maximizes the linear term $z^\top p$, which assigns higher probabilities to elements with higher logits z, while the regularization term $\frac{1}{2}\|p\|_2^2$ encourages sparsity by penalizing the probability distribution's ℓ_2-norm.

Using sparsemax instead of softmax should improve the results we get from experiments 1 till 3.

7 Discussion

Datasets I use the jigsaw toxic comment classification dataset from kaggle[2], which presents a multinomial classification task, the imdb[3] and yelp-polarity[4] datasets from huggingface which contain binary labels for sentiment classification, and the dbpedia[5] dataset also from huggingface, which is labelled for a multiclass classification of wikipedia articles.

Results First of all it was observed that computing the SHAP values on average took longer than computing the attention scores *by a factor of 1000*! But this gain in efficiency is only fruitful if the attention scores fare well in the experiments.

Figure 3 shows the results of experiments 1 and 2. The first two columns indicate the accuracy of the model with the trained attention layer and of the uniform adversary. Columns 3 till 5 indicate the fidelity of the explainer, i.e. its accuracy in mimicking the model's behaviour. Figure 4 shows the results of experiment 3.

Both the jigsaw and the dbpedia dataset seem to be highly imbalanced (less than 1 percent are positive predictions). The jigsaw dataset is an example where attention does not make a significant contribution. (Is the attention layer in the jigsaw dataset sparse?) The case of dbpedia shows that even though attention can be seen to make a contribution, its fidelity score can be quite low.

Positive results are observable with respect to imdb and yelp. Here, attention is seen to make a contribution to the prediction, and the explainer mimics the classifier particularly well. In both imdb and yelp, we have a lower share of irrelevant tokens than in dbpedia and jigsaw. In the case of yelp, there is a significant average correlation between the SHAP and the attention explainer. In case of imdb there is very little correlation. However, given that the SHAP explainer in this case also has a significant lower accuracy value and fidelity score than the attention explainer, this might not be seen as speaking against the plausibility of the attention explainer.

The results of experiment 4 are somewhat surprising. Using sparse attention does not lead to significant improvements as expected. But there might be reasons for that. In case of imbd and yelp, manually sampling suggested that the sparse attention layer actually produced 1-hot attention vectors. This was confirmed to some extend by manually sampling the dense attention layer from the models of the previous experiments, which were observed to be quite sparse already, with less than three values above 0.0001. The jigsaw case does not seem to be significant due to its imbalance. In the case of dbpedia the sparse attention explained turned out to be even worse than the dense one.

[2] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification
[3] https://huggingface.co/datasets/stanfordnlp/imdb
[4] https://huggingface.co/datasets/fancyzhx/yelp_ polarity
[5] https://huggingface.co/datasets/fancyzhx/dbpedia_14

	Dense Att.	**Uniform Att.**	**Dense Fidelity**	**Uniform Fidelity**	**SHAP Fidelity**
jigsaw	0.917	0.914	0.932	0.966	0.977
imdb	0.622	0.56	0.84	0.92	0.64
yelp	0.66	0.62	0.825	0.795	0.85
dbpedia	0.91	0.88	0.59	0.73	0.63

Table 1: Results of Experiments 1 and 2.

	Dense/ SHAP Corr.	**Uniform/ SHAP Corr.**	**Dense irrel. tokens**	**SHAP irrel. tokens**
jigsaw	-0.166	0.08	0.2	0.13
imdb	0.014	0.2	0.04	0.08
yelp	0.33	0.2	0.02	0.07
dbpedia	0.05	0.22	0.42	0.06

Table 2: Results of Experiment 3.

	Sparse Att.	**Uniform Att.**	**Sparse Fidelity**	**Uniform Fidelity**	**SHAP Fidelity**
jigsaw	0.92	0.91	0.97	0.99	0.99
imdb	0.62	0.72	0.99	0.97	0.68
yelp	0.62	0.64	0.97	0.9	0.84
dbpedia	0.91	0.88	0.23	0.61	0.63

Table 3: Results of Experiments 4a).

	Sparse/ SHAP Corr.	**Uniform/ SHAP Corr.**	**Sparse irrel. tokens**	**SHAP irrel. tokens**
jigsaw	-0.01	0.02	1(!)	0.1
imdb	0.045	0.2	0.19	0.091
yelp	0.02	0.2	0.6	0.08
dbpedia	0.02	0.08	0.05	0.21

Table 4: Results of Experiment 4b).

8 Conclusion

The results highlight the following two tentative conclusions:

- In some cases (e.g., dbpedia), attention appears to contribute to predictions but does not enhance faithfulness. This seems to speak against the hypothesis that attention being making a significant contribution to the model prediction is enough. However, in this case, the SHAP explainer does not fare very well, either.
- On tasks like imdb and yelp, attention shows potential for plausibility when its fidelity score is high. Double checking after the surprising results of the experiments with one-hot (or discrete) attention, we observed that dense attention turned out to have been sparse after all. In this respect, going discrete can be seen as overkill. In this regard, we may assume that sparse(!) attention works. But this does not tell us much yet about dense attention distributions that don't turn out to be sparse.

Immediately following up, there can be two further investigations:

- **Conicity of Attention Distributions**: Investigate the conicity of the output of the BERT classifier into the attention layer. This would shed some more light on the above situation and might further explain the numbers observed.
- **Complex NLP Tasks**: Extend analysis to tasks such as text summarization and question answering to evaluate attention's efficacy in more complex setups.

References

[1] Francesco Bodria et al. "Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis". In: *Sistemi Evoluti per Basi di Dati*. 2020.

[2] Hugh Chen, Scott M. Lundberg, and Su-In Lee. "Explaining a series of models by propagating Shapley values". In: *Nature Communications* 13.1 (2022), p. 4512.

[3] Riccardo Guidotti et al. *A Survey Of Methods For Explaining Black Box Models*. 2018. arXiv: `1802.01933 [cs.CY]`.

[4] Sarthak Jain and Byron C. Wallace. "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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 06/2019, pp. 3543–3556.

[5] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: `1705.07874 [cs.AI]`.

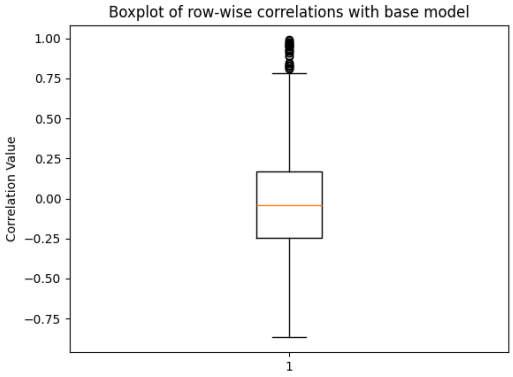
[6] Akash Kumar Mohankumar et al. *Towards Transparent and Explainable Attention Models*. 2020. arXiv: `2004.14243 [cs.CL]`.

[7] Ben Peters, Vlad Niculae, and André F. T. Martins. "Interpretable Structure Induction via Sparse Attention". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, 11/2018, pp. 365–367.

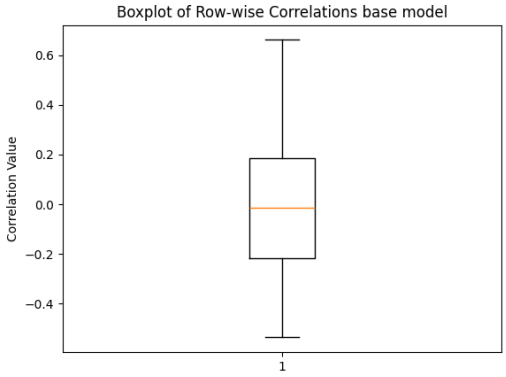
[8] Sofia Serrano and Noah A. Smith. *Is Attention Interpretable?* 2019. arXiv: `1906.03731 [cs.CL]`.

[9] Sarah Wiegreffe and Yuval Pinter. *Attention is not not Explanation*. 2019. arXiv: `1908.04626 [cs.CL]`.

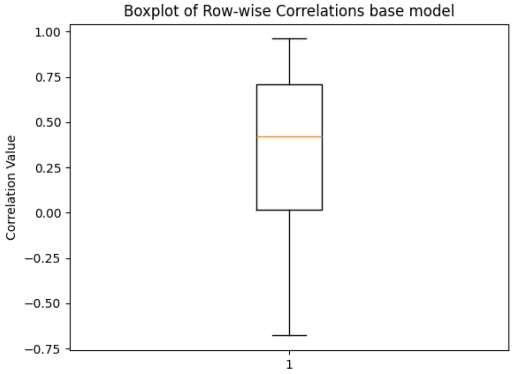
9 Appendix. Sample Predictions and Correlations between SHAP and Attention Based Explanations.



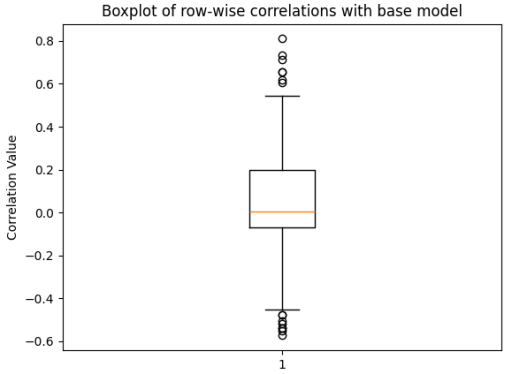
(a) Jigsaw



(b) imdb

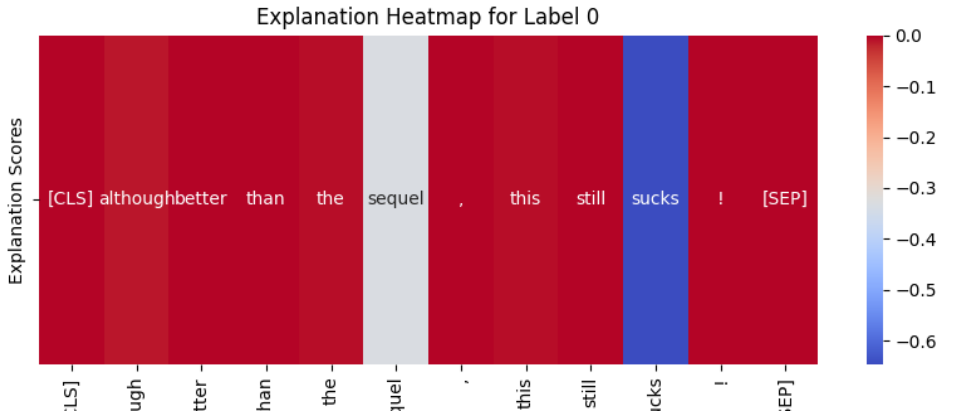


(c) yelp

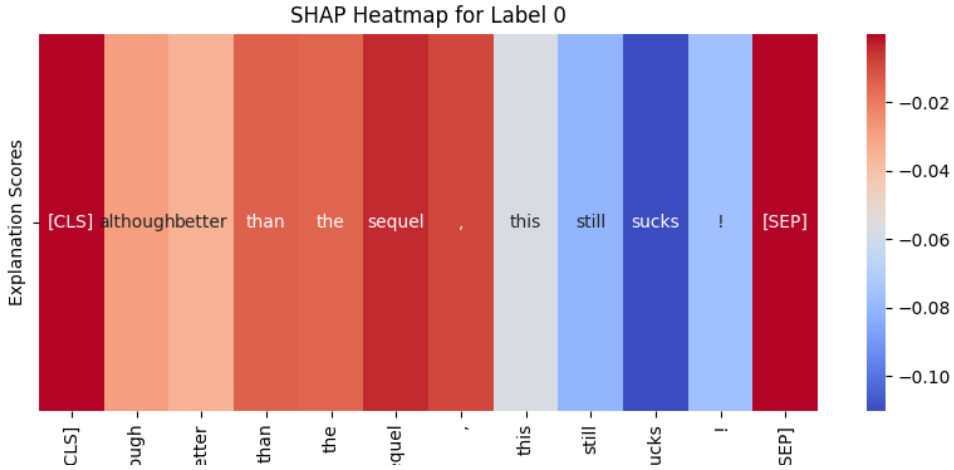


(d) dbpedia

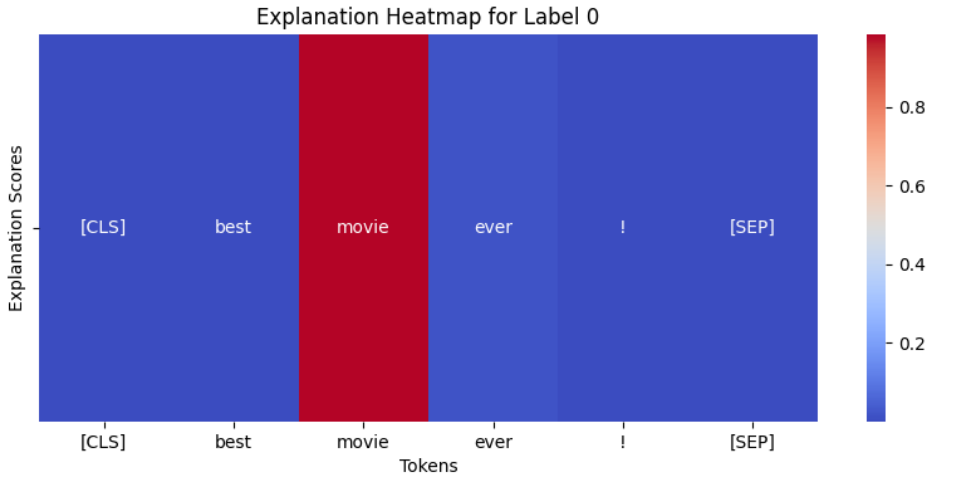
Figure 1: Correlations between the dense attention explainer and SHAP values.



(a) Attention Explainer IMDB

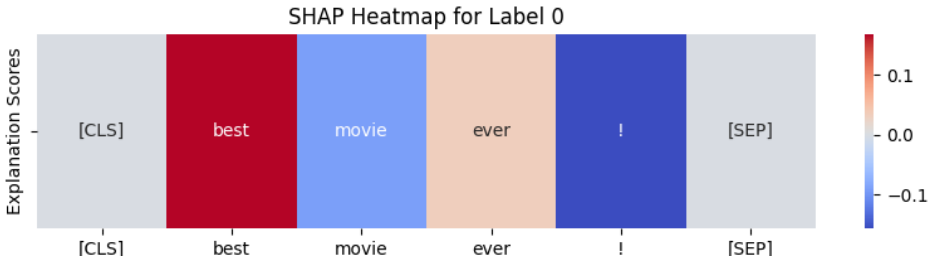


(b) SHAP Explainer IMDB

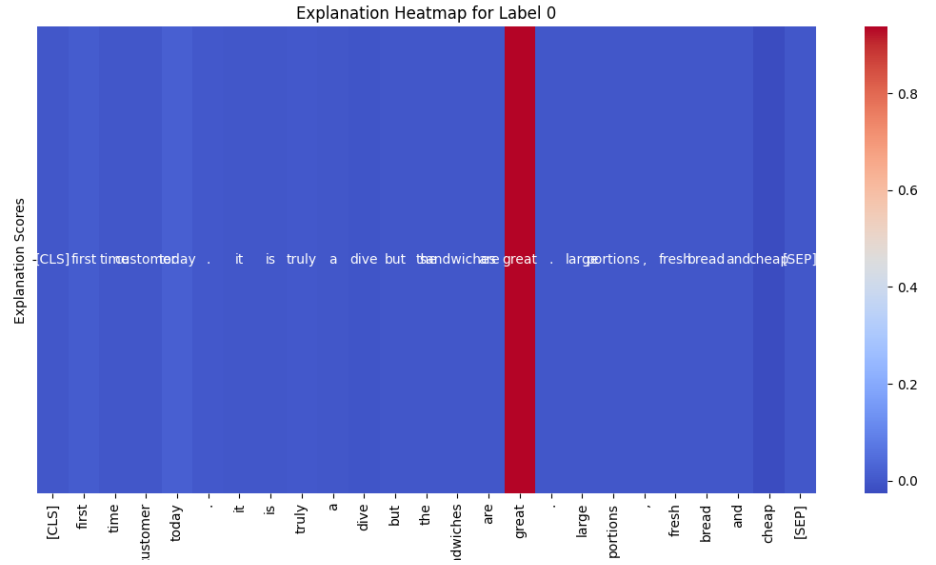


(c) Attention Explainer IMDB

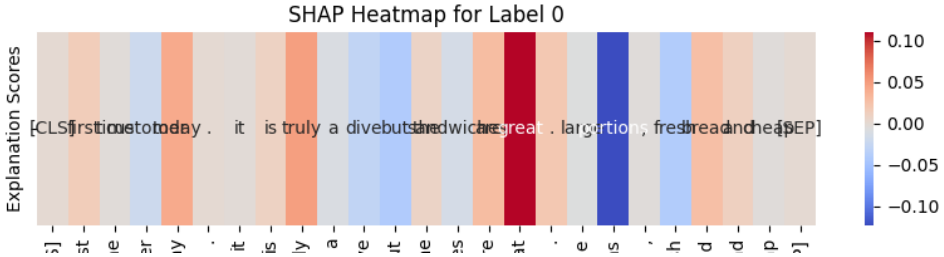
Figure 2: Part 1 of Attention and SHAP heatmaps for IMDB dataset. Predictions were made with the dense attention layer.



(a) SHAP Explainer



(b) Attention Explainer Yelp



(c) SHAP Explainer Yelp

Figure 3: Part 2 of Attention and SHAP heatmaps for Yelp dataset. Predictions were made with the dense attention layer. As one can see, its feature importance scores are already very sparse, indicating that sparse attention would lead to one-hot or discrete attention.