Feature Attribution methods

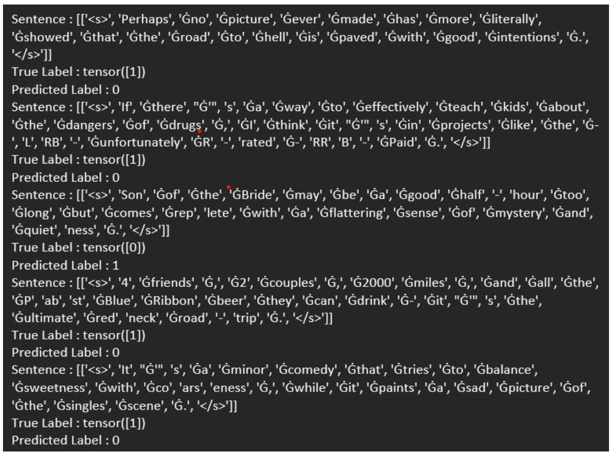
Payanshi Jain
14230143
payanshi.jain@student.uva.nl

1 Introduction

In this paper, the faithfulness evaluation and quantification of a subset of attribution methods are investigated. The main hypothesis posits that attribution evaluation methods may yield lower scores on sentences with complex linguistic interactions, such as negation and adverbials, due to the inherent challenges involved in capturing the true impact of these interactions. To test this hypothesis, a qualitative analysis is conducted on a sentiment classification model using the RoBERTa model(Yinhan Liu, 2019), which is an extended version of BERT trained on a large corpus with longer sequences. The evaluation is performed on the Stanford Sentiment Treebank dataset(SST2)(Richard Socher and Potts, 2020), which comprises 11,855 movie reviews annotated with sentiment labels. This dataset proves to be suitable for attribution extraction due to the explicit words indicating the sentiment expressed in each sentence. Through the utilization of these attribution methods and the subsequent in-depth analysis, the primary objectives are to gain a deeper understanding of the reasoning underlying the sentiment classification model and to evaluate the validity of attribution evaluation methods in accurately measuring faithfulness.

2 Dataset & Model

To evaluate the performance of RoBERTa on the SST2 dataset, an analysis was conducted on the test dataset, specifically targeting the identification of the top 5 sentences where the model made incorrect predictions. The SST2 dataset, consisting of annotated movie reviews, was selected as an appropriate resource for this sentiment analysis task. The findings presented in Figure 1 indicates that the model encountered difficulties in accurately classifying sentences with conflicting sentiments or themes. Notably, it struggled with the interpretation of sarcastic statements, movies that incorporate a combination of sweetness and darkness, as well as road trip narratives involving beer. These mispredictions shed light on the intricate challenge of comprehending subtle nuances and contextual cues in language understanding.



```
Sentence : [['<s>', 'Perhaps', 'Ġno', 'Ġpicture', 'Ġever', 'Ġmade', 'Ġhas', 'Ġmore', 'Ġliterally',
'Ġshowed', 'Ġthat', 'Ġthe', 'Ġroad', 'Ġto', 'Ġhell', 'Ġis', 'Ġpaved', 'Ġwith', 'Ġgood', 'Ġintentions', 'Ġ.',
'</s>']]
True Label : tensor([1])
Predicted Label : 0
Sentence : [['<s>', 'If', 'Ġthere', "Ġ'", 's', 'Ġa', 'Ġway', 'Ġto', 'Ġeffectively', 'Ġteach', 'Ġkids', 'Ġabout',
'Ġthe', 'Ġdangers', 'Ġof', 'Ġdrugs', 'Ġ,', 'ĠI', 'Ġthink', 'Ġit', "Ġ'", 's', 'Ġin', 'Ġprojects', 'Ġlike', 'Ġthe', 'Ġ-
', 'L', 'RB', '-', 'Ġunfortunately', 'ĠR', '-', 'rated', 'Ġ-', 'RR', 'B', '-', 'ĠPaid', 'Ġ.', '</s>']]
True Label : tensor([1])
Predicted Label : 0
Sentence : [['<s>', 'Son', 'Ġof', 'Ġthe', 'ĠBride', 'Ġmay', 'Ġbe', 'Ġa', 'Ġgood', 'Ġhalf', '-', 'hour', 'Ġtoo',
'Ġlong', 'Ġbut', 'Ġcomes', 'Ġrep', 'lete', 'Ġwith', 'Ġa', 'Ġflattering', 'Ġsense', 'Ġof', 'Ġmystery', 'Ġand',
'Ġquiet', 'ness', 'Ġ.', '</s>']]
True Label : tensor([0])
Predicted Label : 1
Sentence : [['<s>', '4', 'Ġfriends', 'Ġ,', 'Ġ2', 'Ġcouples', 'Ġ,', 'Ġ2000', 'Ġmiles', 'Ġ,', 'Ġand', 'Ġall', 'Ġthe',
'ĠP', 'ab', 'st', 'ĠBlue', 'ĠRibbon', 'Ġbeer', 'Ġthey', 'Ġcan', 'Ġdrink', 'Ġ-', 'Ġit', "Ġ'", 's', 'Ġthe',
'Ġultimate', 'Ġred', 'neck', 'Ġroad', '-', 'trip', 'Ġ.', '</s>']]
True Label : tensor([1])
Predicted Label : 0
Sentence : [['<s>', 'It', "Ġ'", 's', 'Ġa', 'Ġminor', 'Ġcomedy', 'Ġthat', 'Ġtries', 'Ġto', 'Ġbalance',
'Ġsweetness', 'Ġwith', 'Ġco', 'ars', 'eness', 'Ġ,', 'Ġwhile', 'Ġit', 'Ġpaints', 'Ġa', 'Ġsad', 'Ġpicture', 'Ġof',
'Ġthe', 'Ġsingles', 'Ġscene', 'Ġ.', '</s>']]
True Label : tensor([1])
Predicted Label : 0
```

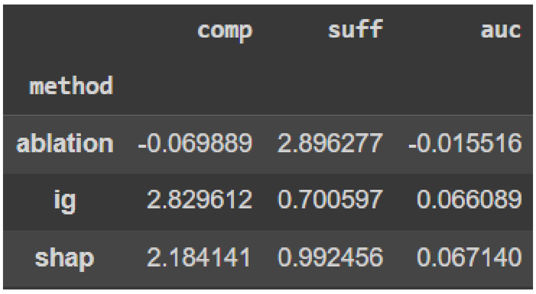
Figure 1: Five sentences found on which the model makes an incorrect prediction.

3 Experiments

In this experiment, three attribution methods are evaluated: Feature Ablation, Shapley Values(KernelShap), and Integrated Gradients. Feature Ablation replaces individual tokens in a sentence with a baseline token and quantifies the resulting difference in model output. KernelShap applies Shapley values from game theory to estimate token contributions in a sentence. Integrated Gradients measures gradients along the trajectory from a baseline input to the original input for attribution assignment. The performance of these methods is assessed using evaluation metrics: Comprehensiveness, Sufficiency, and Area.

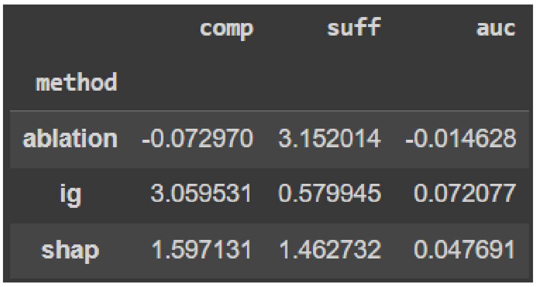
In Figure 2, we evaluate the attribution methods using the <unk> baseline. For the Feature Ablation method, the scores indicate a negative comprehensiveness, suggesting that the replaced tokens do not contribute significantly to the model's output. However, the sufficiency score is high, indicating that the modified sentences with replaced tokens can still retain the original sentiment classification. Integrated Gradients (IG) shows positive scores for comprehensiveness and sufficiency, implying that the contributions of different tokens are adequately captured. The area score indicates a moderate level of attribution. Shapley Values yields results similar to IG, with relatively high scores for comprehensiveness and sufficiency, suggesting effective attribution estimation.

In Figure 3 presents the attribution evaluation scores using the <pad> baseline. Feature Ablation exhibits similar patterns to the <unk> baseline, with low comprehensiveness and high sufficiency scores. IG maintains its high sufficiency score, suggesting that modified sentences can still capture the original sentiment classification. However, the comprehensiveness score is higher compared to the <unk> baseline. Shapley Values shows a relatively balanced performance in terms of comprehensiveness and sufficiency. In Figure 4, we utilize a zero-valued baseline to evaluate the attribution methods. The scores for Feature Ablation and Shapley Values remain consistent with the previous baselines. IG demonstrates similar performance as in the previous tables, with a high sufficiency score and a moderate comprehensiveness score.



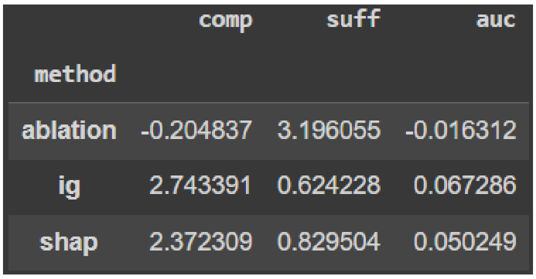
	comp	suff	auc
method			
ablation	-0.069889	2.896277	-0.015516
ig	2.829612	0.700597	0.066089
shap	2.184141	0.992456	0.067140

Figure 2: The <unk> baseline shows mixed results across attribution methods, with IG having the highest comprehensiveness and area scores, while ablation exhibits negative comprehensiveness and area scores.



	comp	suff	auc
method			
ablation	-0.072970	3.152014	-0.014628
ig	3.059531	0.579945	0.072077
shap	1.597131	1.462732	0.047691

Figure 3: The <pad> baseline shows more consistent results across attribution methods, with IG again demonstrating the highest comprehensiveness and area scores, while ablation and SHAP show mixed scores across the evaluation methods.



	comp	suff	auc
method			
ablation	-0.204837	3.196055	-0.016312
ig	2.743391	0.624228	0.067286
shap	2.372309	0.829504	0.050249

Figure 4: The zero-valued baseline yields similar results to the <unk> baseline, with IG demonstrating the highest comprehensiveness and area scores, while ablation shows negative comprehensiveness and area scores.

Comparing the three tables, we observe that Feature Ablation consistently yields negative comprehensiveness scores, indicating limited contribution of replaced tokens to the model's output. However, sufficiency scores remain consistently high, suggesting that the sentiment classification can still be retained. IG shows relatively balanced scores for comprehensiveness and sufficiency across all baselines. Shapley Values consistently performs well in terms of comprehensiveness and sufficiency. In terms of area scores, all three methods demonstrate moderate attribution estimation.

4 Conclusion & Discussion

In conclusion, this study has provided a detailed analysis of attribution evaluation methods and their performance in capturing the true impact of linguistic interactions. The findings support the hypothesis that attribution evaluation methods may yield lower scores on sentences with complex linguistic interactions. Overall, the results of the analysis provide valuable insights into the functioning of the sentiment classification model and the effectiveness of attribution evaluation methods. However, further research is needed to validate the findings on a larger scale and explore additional attribution methods. Nonetheless, this study contributes to the understanding of faithfulness evaluation in sentiment analysis models, emphasizing the importance of considering complex linguistic interactions when interpreting attribution scores.

References

Jean Wu Jason Chuang Christopher Manning Andrew Ng Richard Socher, Alex Perelygin and Christopher Potts. 2020. Stanford sentiment treebank.

Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.