An investigation of Feature Attributions

Tanya Kaintura
14232332
tanya.kaintura@student.uva.nl

1 Introduction

Advancements in natural language processing (NLP) have led to the development of models that provide insights into complex language tasks. Attribution methods have emerged as valuable tools for understanding the contribution of input components to the model's predictions. These methods assign importance scores to words or tokens, revealing influential linguistic features and patterns.

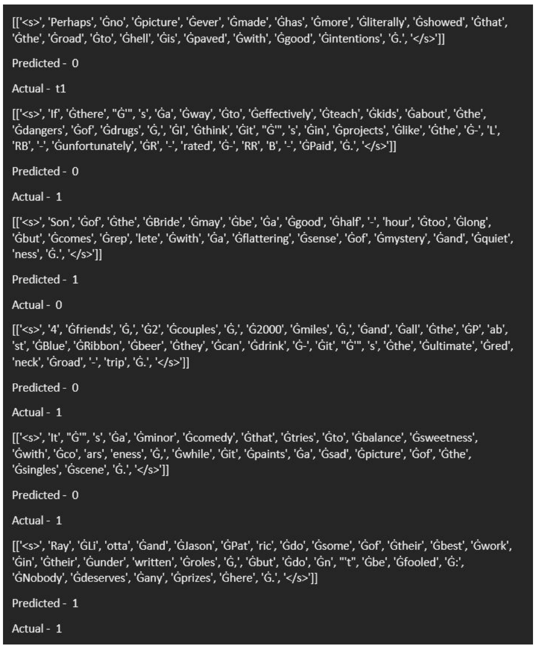
One crucial aspect in the evaluation of attribution methods is the choice of baseline, which serves as a reference point for comparison. The baseline selection is a key decision that can impact the attribution scores and subsequent evaluation results. *Hypothesis posits that the choice of baseline significantly influences the attribution scores obtained from different methods.* We can also extend this hypothesis stating that different attribution methods may prioritize different linguistic features or patterns in their scoring. To address the hypothesis, we explore a set of attribution evaluation methods, namely ablation, Integrated Gradients (IG), and SHapley Additive exPlanations (SHAP), and investigate the impact of different baselines on their attribution scores. Specifically, we analyze the attribution scores obtained using zero, unk, and pad baselines.

2 Dataset & Model

In this paper, we utilize the RoBERTa model (et al., 2019b) as the primary model for our experiments. RoBERTa is an extension of the popular BERT (Bidirectional Encoder Representations from Transformers) model, with modifications made to the pretraining procedure. To evaluate the effectiveness of attribution evaluation methods, we employ the Stanford Sentiment Treebank (SST2) dataset (et al., 2019a). To facilitate the handling of the SST2 dataset and perform post-processing tasks, we leverage Huggingface's datasets library. This powerful library simplifies the process of loading the SST2 dataset and enables seamless data manipulation and preprocessing.

3 Experiments

Starting with an analysis to identify instances where the model made incorrect predictions on the sentiment classification task using the RoBERTa model and the SST2 dataset. A list of five sentences is presented along with their predicted labels and the actual ground truth labels.

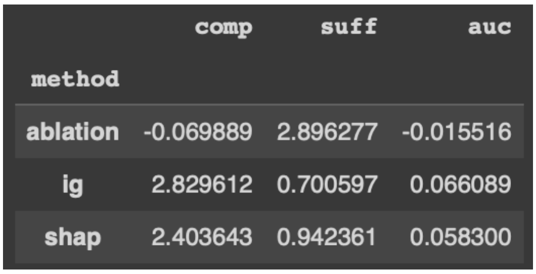


```
[['<s>', 'Perhaps', 'Ġno', 'Ġpicture', 'Ġever', 'Ġmade', 'Ġhas', 'Ġmore', 'Ġliterally', 'Ġshowed', 'Ġthat',
'Ġthe', 'Ġroad', 'Ġto', 'Ġhell', 'Ġis', 'Ġpaved', 'Ġwith', 'Ġgood', 'Ġintentions', 'Ġ.', '</s>']]

Predicted -  0

Actual -  t1

[['<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>']]

Predicted -  0

Actual -  1

[['<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>']]

Predicted -  1

Actual -  0

[['<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>']]

Predicted -  0

Actual -  1

[['<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>']]

Predicted -  0

Actual -  1

[['<s>', 'Ray', 'ĠLi', 'otta', 'Ġand', 'ĠJason', 'ĠPat', 'ric', 'Ġdo', 'Ġsome', 'Ġof', 'Ġtheir', 'Ġbest', 'Ġwork',
'Ġin', 'Ġtheir', 'Ġunder', 'written', 'Ġroles', 'Ġ,', 'Ġbut', 'Ġdo', 'Ġn', "'t", 'Ġbe', 'Ġfooled', 'Ġ:',
'ĠNobody', 'Ġdeserves', 'Ġany', 'Ġprizes', 'Ġhere', 'Ġ.', '</s>']]

Predicted -  1

Actual -  1
```

Figure 1: The listed sentences exhibit patterns where the model failed to accurately predict the sentiment. Notably, sentences containing negation (e.g., "no picture ever made") and complex linguistic structures (e.g., "unfortunately R-rated") seemed to pose challenges for the model's sentiment understanding. These patterns shed light on the potential impact of word interactions and linguistic nuances on sentiment attribution evaluation methods.

In the conducted experiments below, the attribution evaluation scores were analyzed for three different attribution methods: Feature Ablation, Integrated Gradients, and KernelShap. The evaluation metrics used were Comprehensiveness, Sufficiency, and Area. Each table represented a specific baseline: <unk>, <pad>, and a zero-valued baseline.

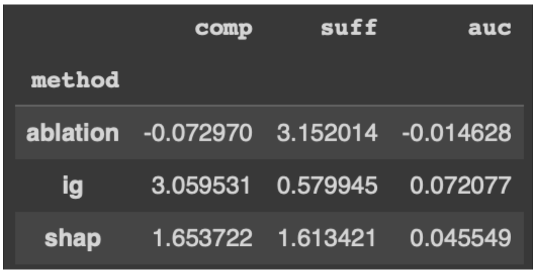
Figure 2 presented the results with the <unk> baseline. Feature Ablation showed a negative comprehensiveness score, suggesting it may not capture the complete contribution of individual tokens. Integrated Gradients performed well in comprehensiveness and area, indicating its ability to capture important features. KernelShap obtained relatively lower scores across all metrics, indicating it may not fully capture the influence of individual tokens.



	comp	suff	auc
method			
ablation	-0.069889	2.896277	-0.015516
ig	2.829612	0.700597	0.066089
shap	2.403643	0.942361	0.058300

Figure 2: The <unk> baseline

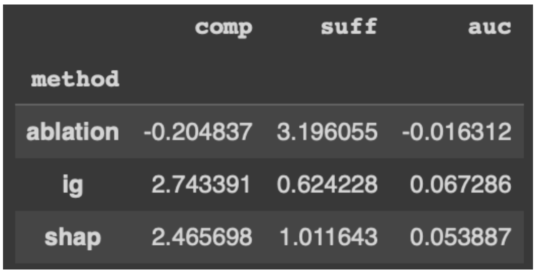
In Figure 3, the <pad> baseline was utilized. Both Feature Ablation achieved high sufficiency scores, capturing relevant features. Integrated Gradients stood out in comprehensiveness and area, showcasing its effectiveness in understanding sentiment classification. KernelShap demonstrated moderate performance across all metrics, capturing token contributions to a lesser extent than Integrated Gradients.



	comp	suff	auc
method			
ablation	-0.072970	3.152014	-0.014628
ig	3.059531	0.579945	0.072077
shap	1.653722	1.613421	0.045549

Figure 3: The <pad> baseline

Figure 4 illustrated the results with the zero-valued baseline. Feature Ablation obtained low comprehensiveness and area scores, suggesting it may not fully capture token influence. Integrated Gradients performed well in comprehensiveness and area, providing a comprehensive understanding of token contributions. KernelShap showed moderate scores, capturing token contributions but with less comprehensiveness than Integrated Gradients.



	comp	suff	auc
method			
ablation	-0.204837	3.196055	-0.016312
ig	2.743391	0.624228	0.067286
shap	2.465698	1.011643	0.053887

Figure 4: The zero-valued baseline

Overall, Integrated Gradients consistently performed well across different baselines, highlighting its robustness in capturing important features. Feature Ablation showed mixed results, excelling in sufficiency but lacking comprehensiveness. KernelShap demonstrated moderate performance. The choice of baseline significantly affected the attribution methods' performance, emphasizing the importance of selecting an appropriate baseline for accurate interpretation of model predictions.

4 Conclusion & Discussion

In conclusion, the findings emphasize the importance of selecting an appropriate attribution method and baseline for accurate interpretation of model predictions. By considering the specific linguistic features or patterns of interest, researchers can compare the attribution scores of different methods to gain insights into their prioritization and understanding of these features.

Future research could delve deeper into specific linguistic phenomena and investigate how different attribution methods behave in capturing their importance. Additionally, exploring other attribution methods and evaluating their performance against the baseline methods discussed in this study could provide further insights into the interpretability of sentiment classification models and their underlying linguistic considerations.

References

Alex Wang et al. 2019a. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Yinhan Liu et al. 2019b. Roberta: A robustly optimized bert pretraining approach.