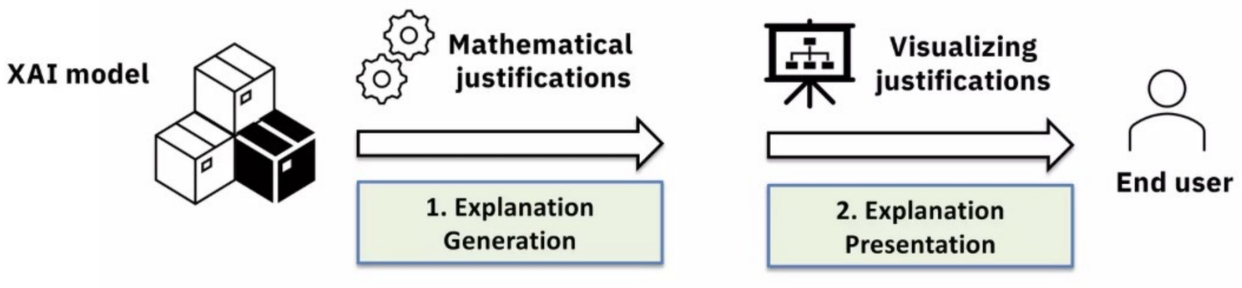
Exploring the Impact of Sentence Length on Attribution Evaluation Scores

Universiteit van Amsterdam

Tanya Kaintura, Payanshi Jain

Motivation

- **Hypothesis:** Longer sentences tend to receive lower attribution evaluation scores due to the presence of complex linguistic interactions.
- Word phrases and idiomatic expressions often found in longer sentences convey meaning or sentiment as a whole, making semantics of a sentence independent of individual words[1][2].
- Importance of word phrases: The presence of word phrases as shortcuts in longer sentences highlights the need to understand their impact on attribution methods, as they can significantly influence the evaluation of feature importance.



Methodology

1) RoBERTa Model

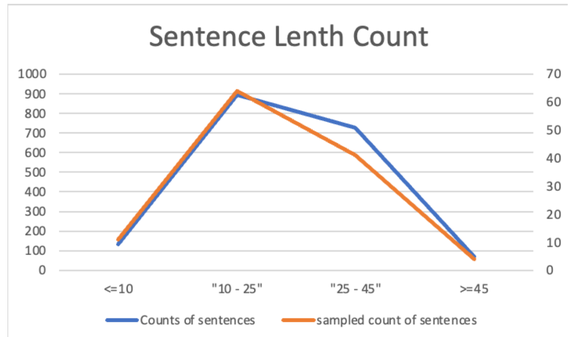
- RoBERTa is an extended version of BERT trained on a large corpus with longer sequences.
- RoBERTa has the same architecture as BERT but uses a byte-level BPE as a tokenizer and uses a different pretraining scheme.

Stanford Sentiment Treebank Dataset

Characteristics:
- 11,855 movie reviews
- Each annotated with sentiment

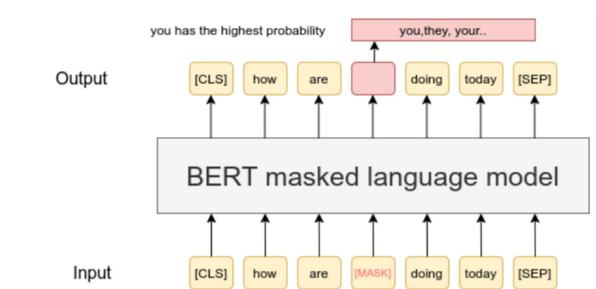
2) Weighted Random Sampling Approach:

1. **I**terate through the corpus and **calculate the length** of each sentence**.**
2. **Creating in four buckets:** based on length range(<=10, >10 & <=25, >25 & <=45, and <45).
3. **Weighted Radom sampling:** Based on the above distributions applied weighted random sampling to obtain **120 samples**.

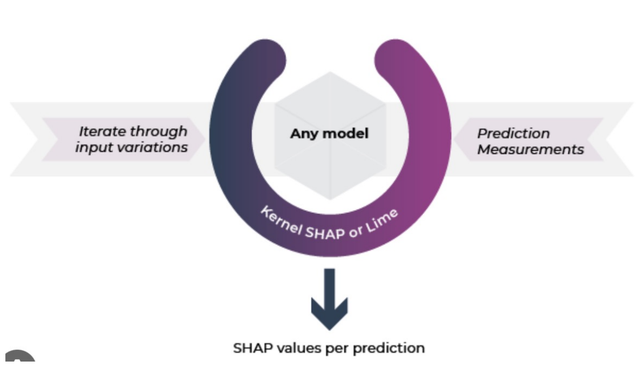


3) Feature Attribution Methods:

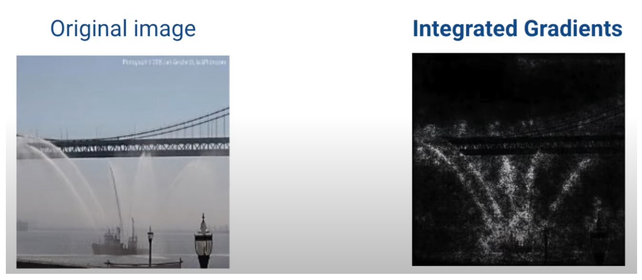
1. Feature Ablation:



2. Shapley Values (KernelSHAP):



3. Integrated Gradients:



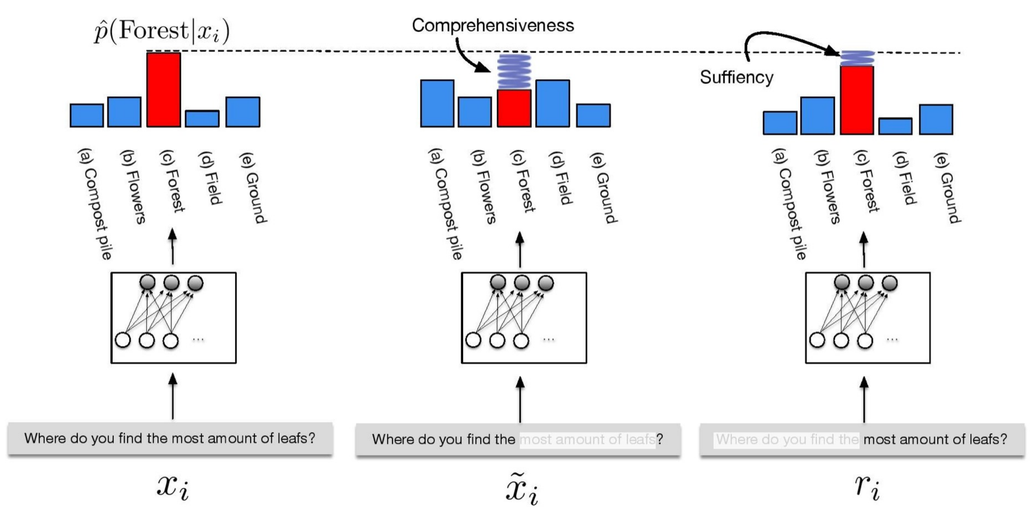
4) Evaluation:

1. Comprehensiveness

Comprehensiveness is an evaluation metric used to assess the attribution methods' ability to capture the overall contribution of the features or tokens in a sentence.

2. Sufficiency

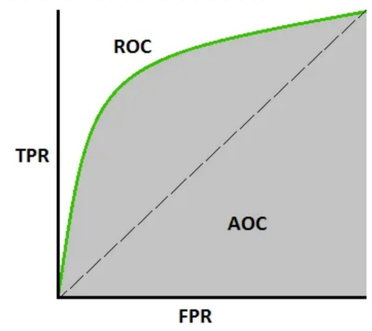
Sufficiency is an evaluation metric that measures whether the attribution methods can accurately identify the minimal set of features necessary to make a decision.



Rationale comprehensiveness is the measure of decrease in a model's confidence in the predicted output ('Forest') when the tokens comprising the provided rationale are erased. Rationale sufficiency is the measure of the model's ability to come to the same output prediction('Forest') using only the rationales.

3. Area:

Area is an evaluation metric used to assess the overall quality of the attribution scores by considering the entire range of possible feature subsets.

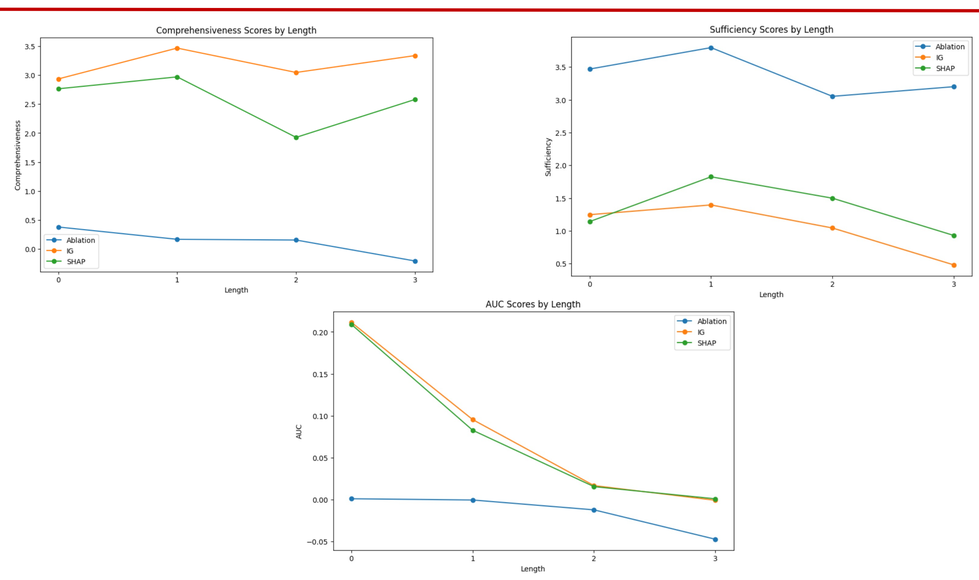


AUC - ROC Curve [Image 2] (Image courtesy: My Photoshopped Collection)

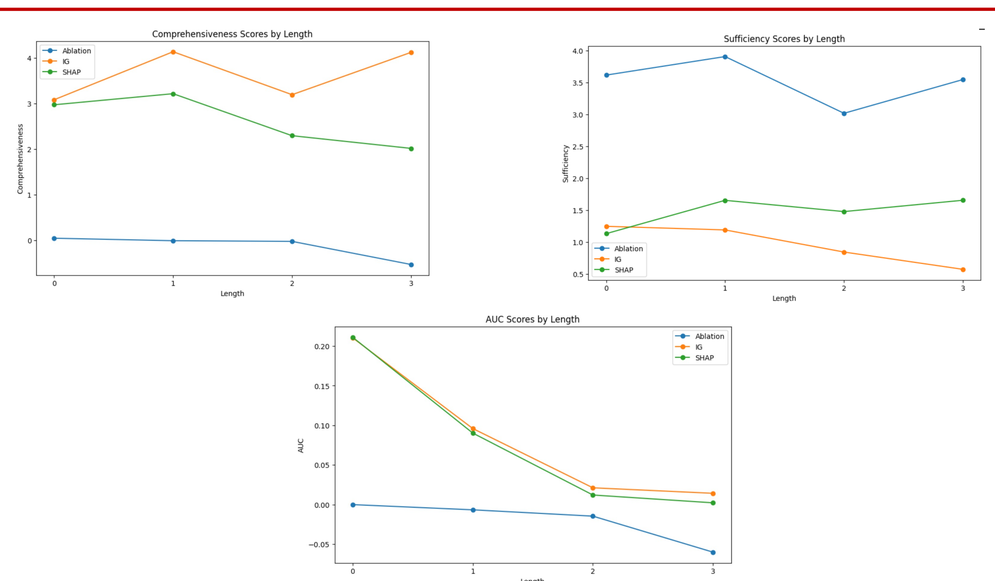
Experiments and Results

Evaluation vs Length of sentences for different Attribution method

UNK Baseline



ZERO Baseline



PAD Baseline – the graphs followed the same pattern as UNK

Limitations

- Limited Attribution Methods
- Scope of Linguistic Interactions
- Sample Size and Diversity
- Need for Further Investigation

Conclusion

Based on the provided tables and the given hypothesis, the following conclusions can be drawn
- Comprehensiveness and Sufficiency: It can be observed that longer sentences generally receive lower attribution evaluation scores. This supports the hypothesis that longer sentences tend to have lower attribution scores.
- AUC (Area Under the Curve): The AUC scores provide an overall measure of the performance of the attribution methods. The AUC scores are decreasing across baselines supporting our hypothesis.

References

[1]:Mengnan Du1∗, Varun Manjunatha2, Rajiv Jain2, Ruchi Deshpande3, Franck Dernoncourt2, Jiuxiang Gu2, Tong Sun2 and Xia Hu1. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU Models , 2021.
[2]:Yoon Kim. Convolutional Neural Networks for Sentence Classification. ArXiv preprint arXiv:1408.5882, 2014.