Social Computing Group,
Department of Informatics
Technical University of Munich

GrammarSHAP: An Efficient Model-Agnostic and Structure-Aware NLP Explainer

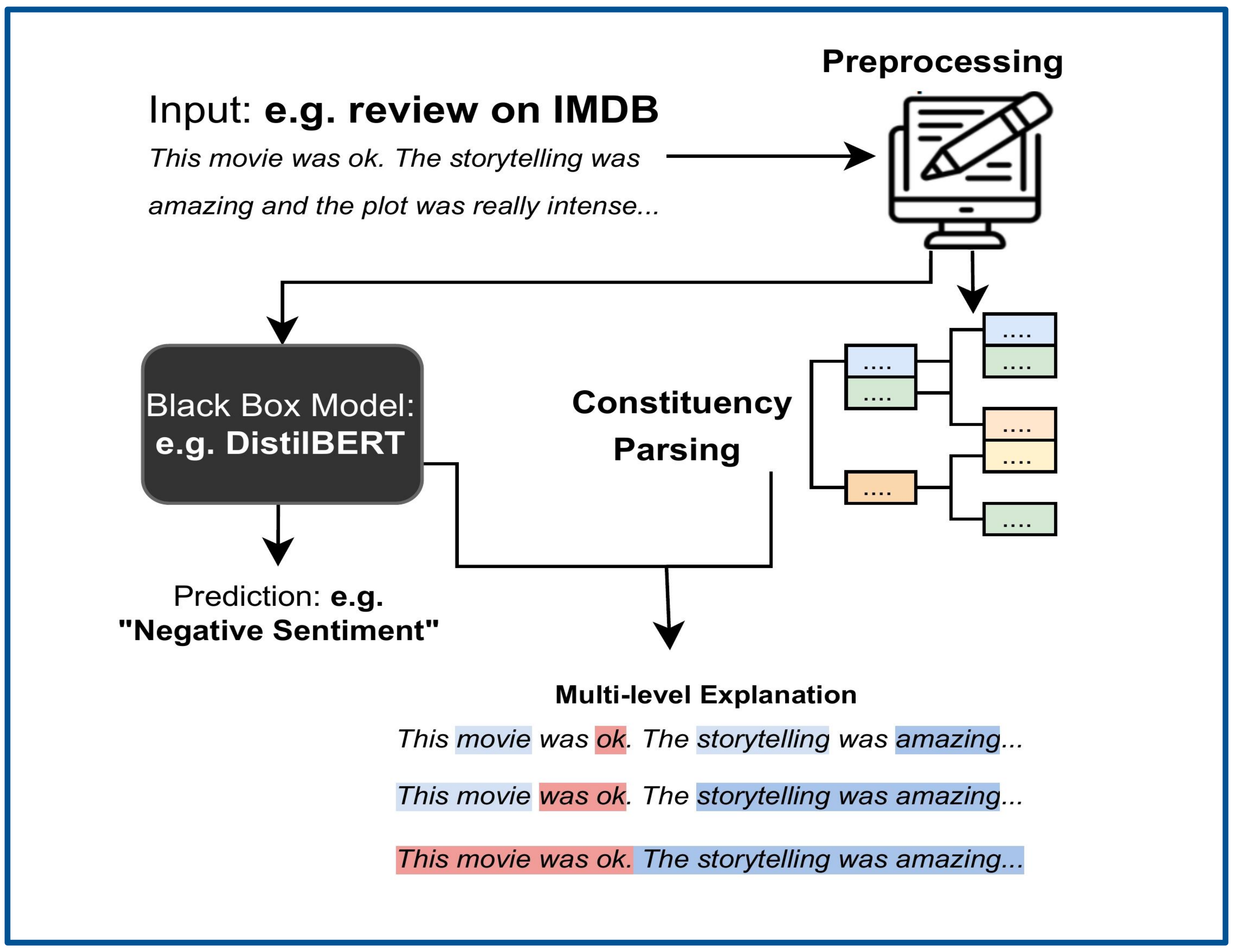
Edoardo Mosca, Defne Demirtürk, Luca Mülln, Fabio Raffagnato, Georg Groh

Motivation

- Transformers exploit **contextual information**, post-hoc explainability methods do not !!
- Most explanations provide **importance scores only at the word level**.

Contribution

- We design ***GrammarSHAP***, a **model-agnostic** approach to generate **multi-level explanations** that consider the **text's structure**.
- We **extend the SHAP framework** with an efficient method **tailored for NLP use cases**.

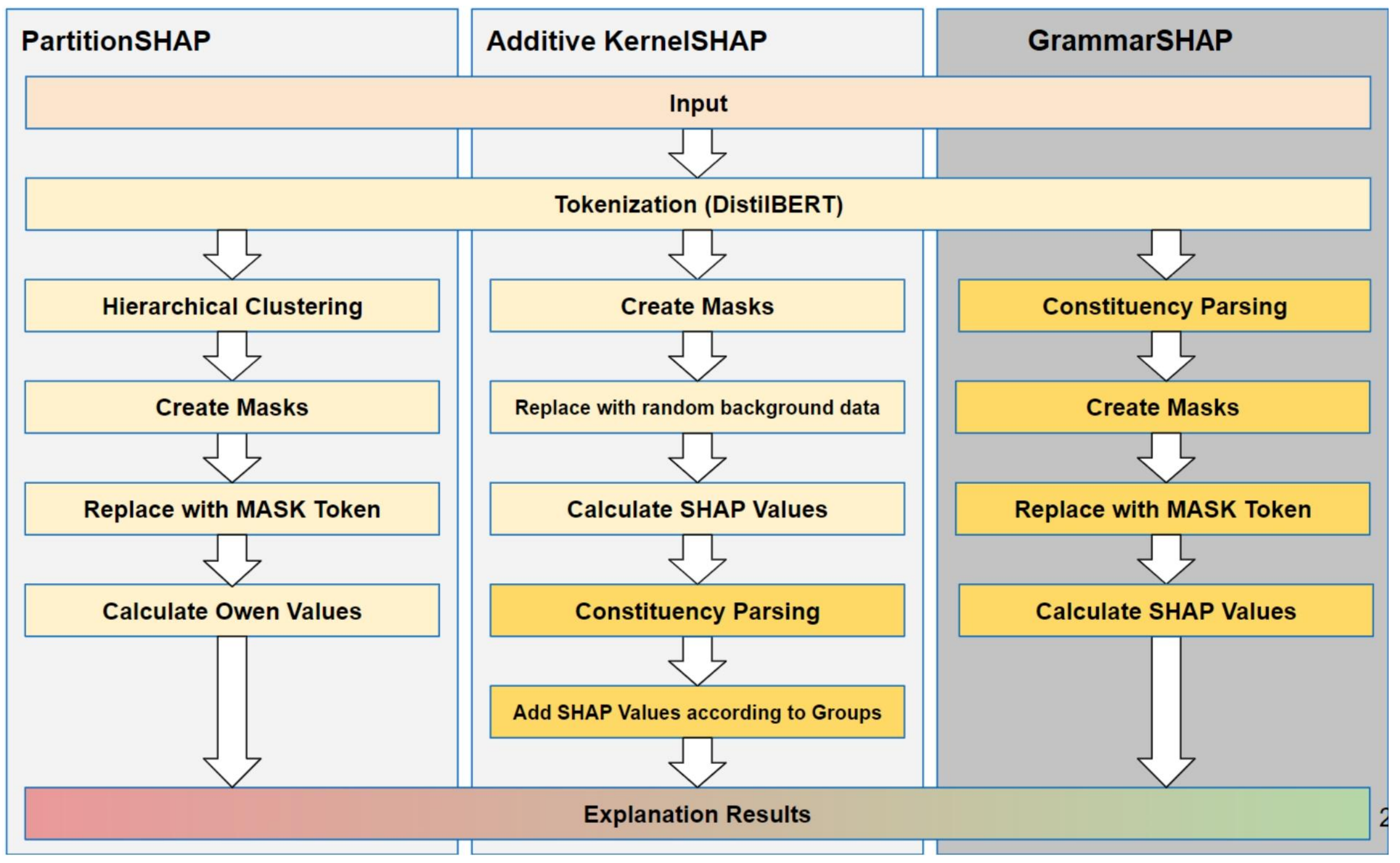

Preprocessing
Input: **e.g. review on IMDB**
This movie was ok. The storytelling was amazing and the plot was really intense...
Black Box Model: **e.g. DistilBERT**
Constituency Parsing
....
Prediction: **e.g. "Negative Sentiment"**
Multi-level Explanation
This movie was ok. The storytelling was amazing...
This movie was ok. The storytelling was amazing...
This movie was ok. The storytelling was amazing...

1 Use a constituency parsing layer to hierarchically select multi-word tokens.

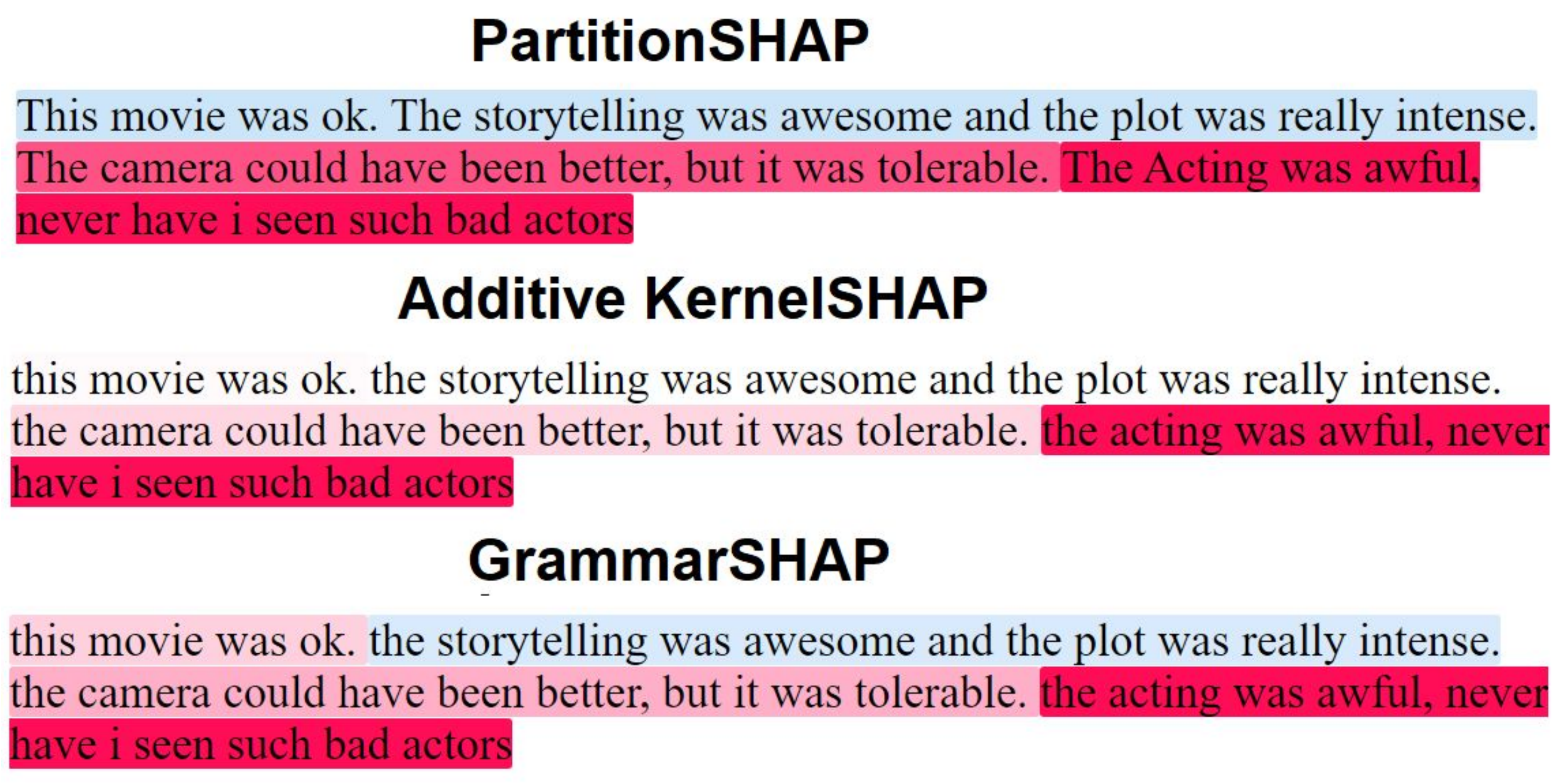
2 Adapt KernelSHAP for multi-words tokens and use [MASK] tokens for improved efficiency and run-time.

3 Obtain group-level feature importance.

Methods for grouped explanations



Evaluation and Results


PartitionSHAP

This movie was ok. The storytelling was awesome and the plot was really intense. The camera could have been better, but it was tolerable. The Acting was awful, never have i seen such bad actors

Additive KernelSHAP

this movie was ok. the storytelling was awesome and the plot was really intense. the camera could have been better, but it was tolerable. the acting was awful, never have i seen such bad actors

GrammarSHAP

this movie was ok. the storytelling was awesome and the plot was really intense. the camera could have been better, but it was tolerable. the acting was awful, never have i seen such bad actors

Explanations produced by the three compared methods.

Method	Running Time
PartitionSHAP	2sec
Add. KernelSHAP	~1h
GrammarSHAP	~3min

Average running time for GrammarSHAP compared to the existing SHAP baselines.

Datasets

IMDb
SST-2

Target Models

DistilBERT
BiLSTM

Takeaways

- GrammarSHAP can identify more **fine-grained contributors**, especially if the sentence contains contrastive sentiments.
- The usage of masking tokens instead of a background dataset considerably speeds up the execution, thus GrammarSHAP is suitable for long texts.

Limitations and Future Work

- A quantitative evaluation for faithfulness is required.
- Improving efficiency → adapting other explainers to the grouping method.
- More word-grouping functions can be implemented via dependency parsing.