SHAP-Based Explanation Methods: A Review for NLP Interpretability

Edoardo Mosca[1]

Ferenc Szigeti[1]

Stella Tragianni[1]

Daniel Gallagher[2]

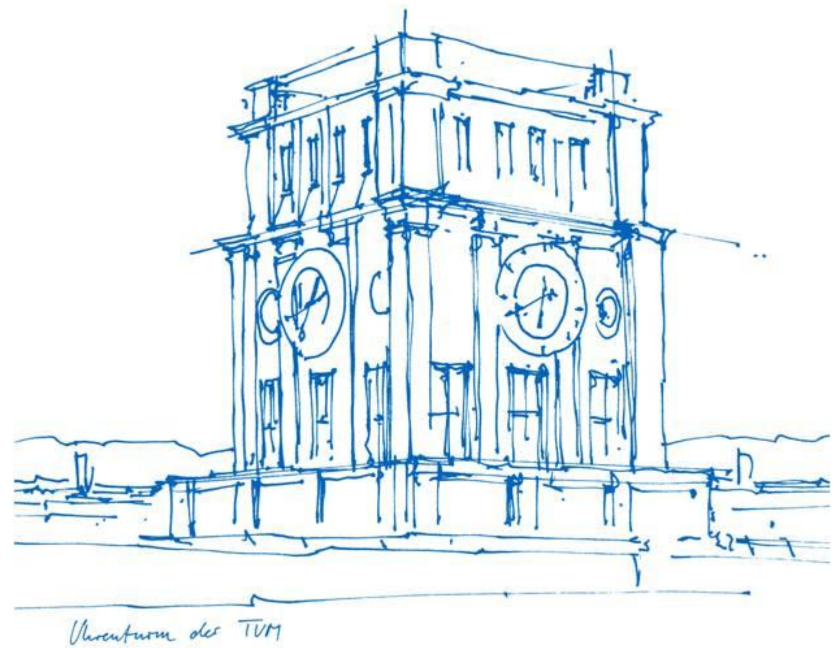
Georg Groh[1]


[1] *Technical University of Munich, Germany*

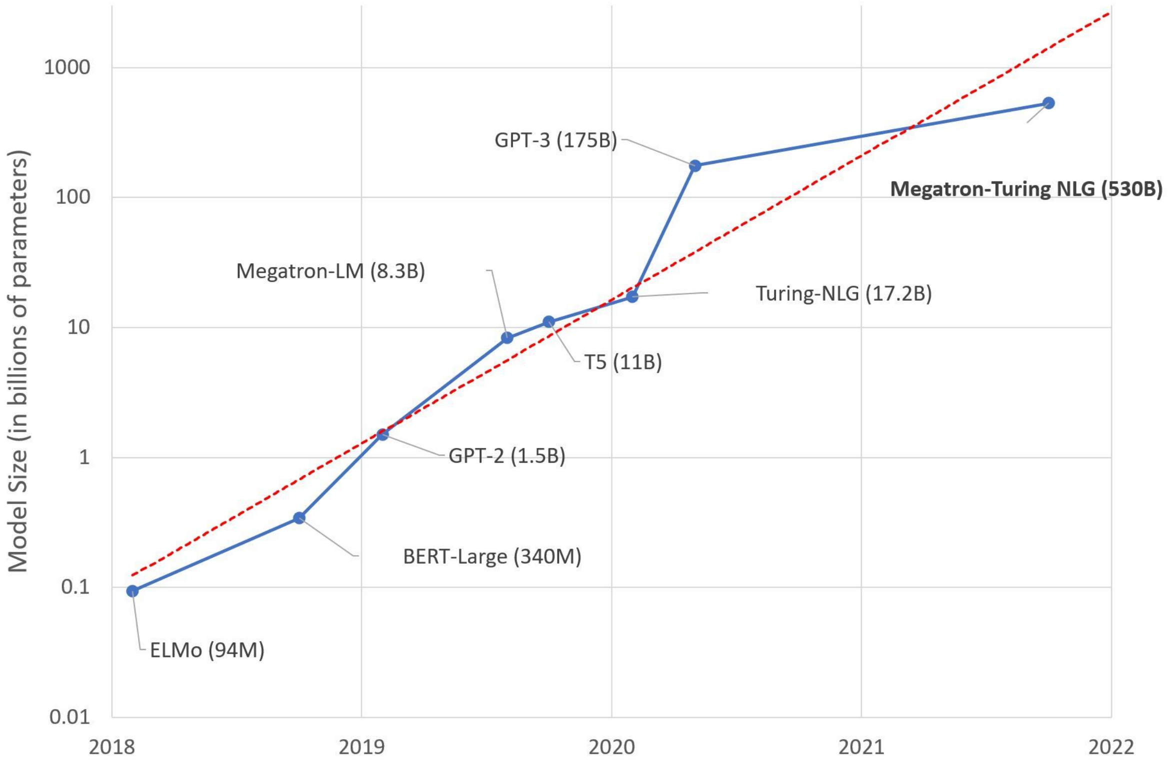
[2] *University College Dublin, Ireland*

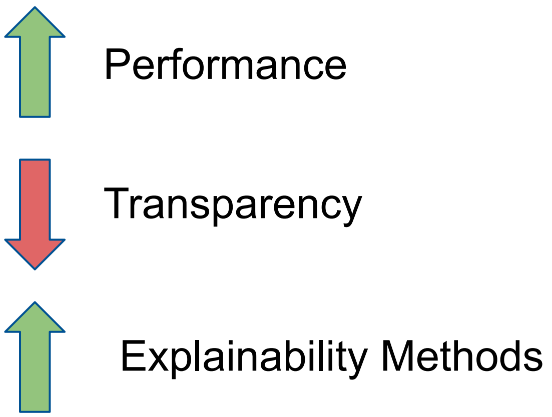
COLING 2022

12nd-17th October | Gyeongju



Explainability is in High Demand



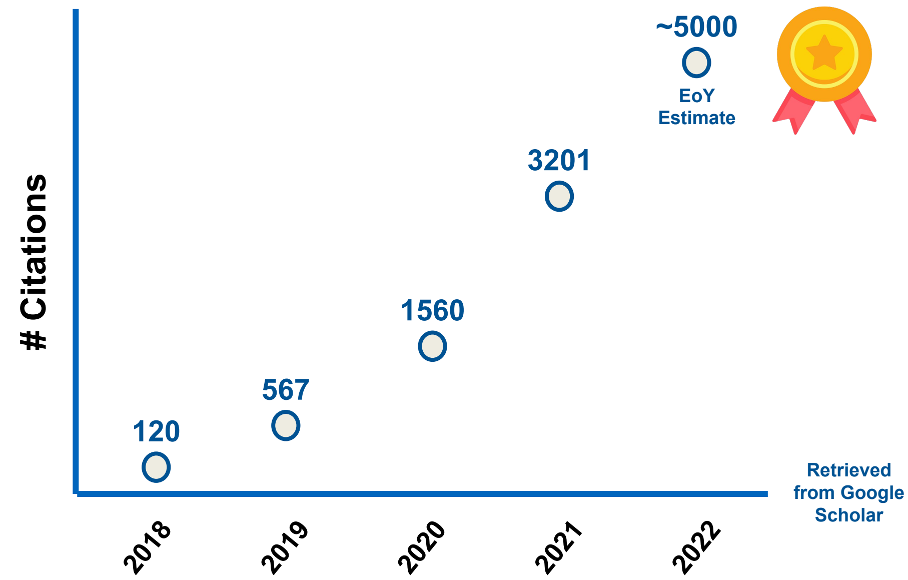


SHAP is in Business

Lundberg et al. (2017)

SHAP

- Has become a gold standard
- Many follow-up methods.
- Relevance for NLP?



Our contribution

1. **Identify five research directions inspired by SHAP**
   Newer methods have focused on different models, data, assumptions, etc..

2. **Review available SHAP-based approaches**
   How each approach addresses existing issues and to what directions belongs.

3. **Investigate the relevance for NLP applications**
   Method-by-method assessment + use-case-based recommendations.

Selection Criteria and Previous Reviews

Previous works:

- Only brief mention of SHAP derivates
- Usually between 5 and 9 methods considered
- Relevance for NLP applications and researcher not addressed

Our Review:

40+ presented approaches

Papers already known to us

+

All papers citing SHAP

↓

Filter out the ones with less than 15 cit.
(10 cit. for last two years)

↓

Consider only papers introducing new expl. methods

Before we start…

Don’t know SHAP yet?

Based on the game-theoretic concept of Shapley Values (1953).

Fairly distribute a ~~reward~~ among a set of ~~players~~ contributing to ~~an outcome~~.

importance score | input features | a model prediction

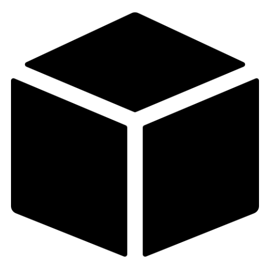
set of all features

$$\mathbf{F} = \{1, 2, \ldots, p\}$$

$$\cup\!|$$

$$S$$

coalition



$$f(S)$$

model output

Shapley Value $\phi_f(i)$: “Clever” avg. of marginal contribution across all possible coalitions.

Don’t know SHAP yet?

- # of coalitions is exponential: approximations are necessary.
- Solid theoretical foundations, versatile, easy to extend.

SHAP

KernelSHAP

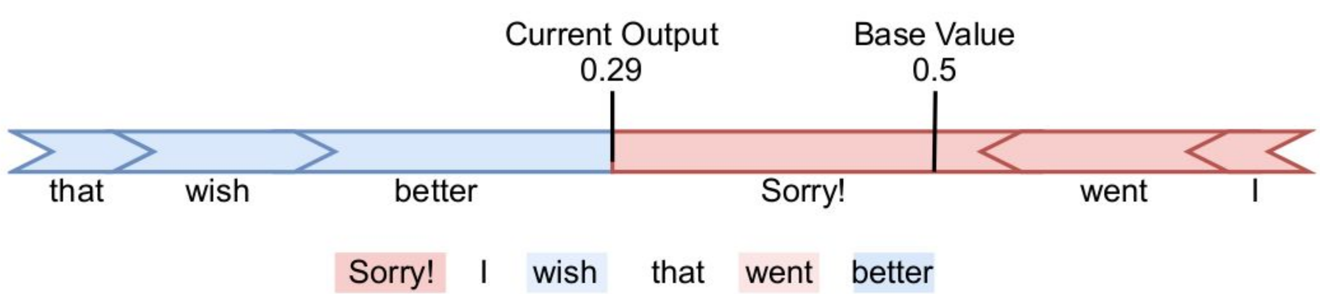
PartitionSHAP

TreeSHAP

DeepSHAP

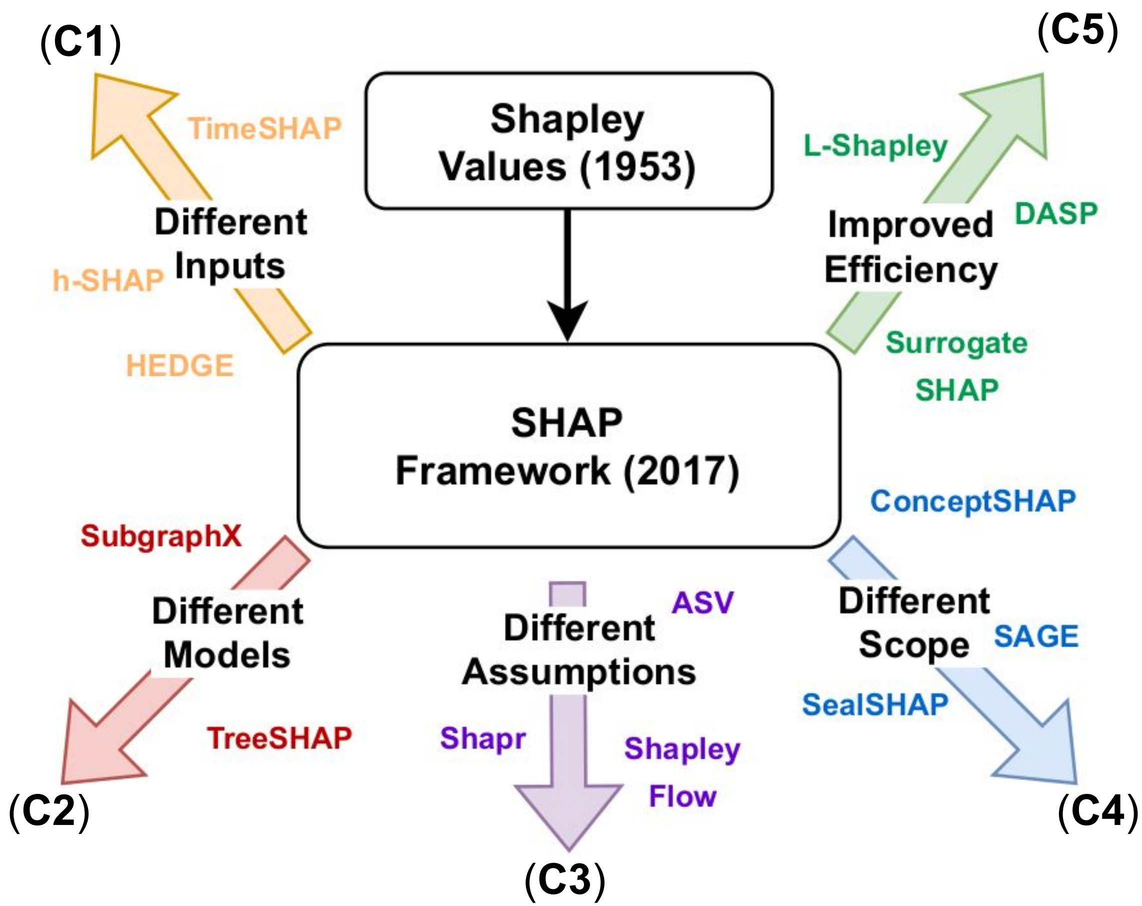
GradientSHAP

LinearSHAP



Our Review

Identification of Research Direction



Categories overlap:
A method can belong to multiple ones.

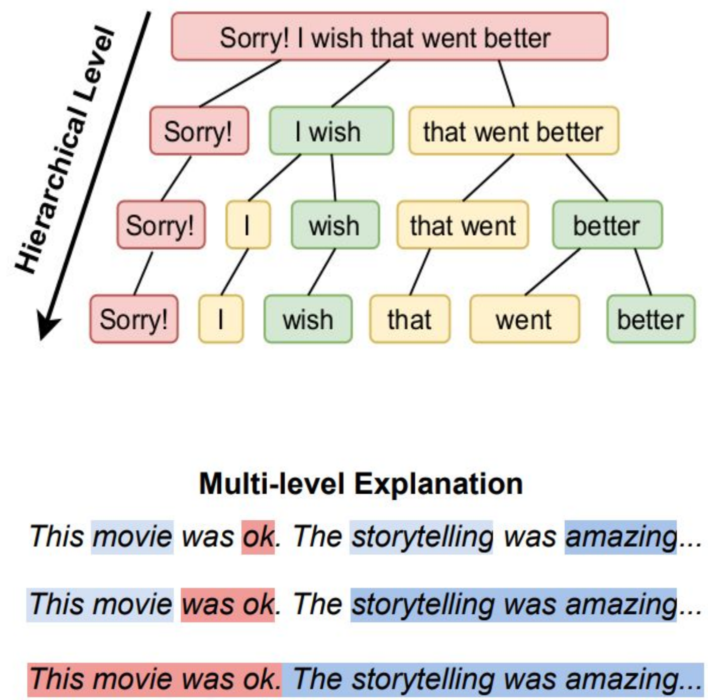
OUR REVIEW

Approaches Tailored to Different Inputs (C1)

- Knowing your input allows stronger assumptions. Plain SHAP oversimplifies.
- Words have strong interactions and their meaning is context dependent.

HEDGE: top-down breaks tokens on weakest interactions.

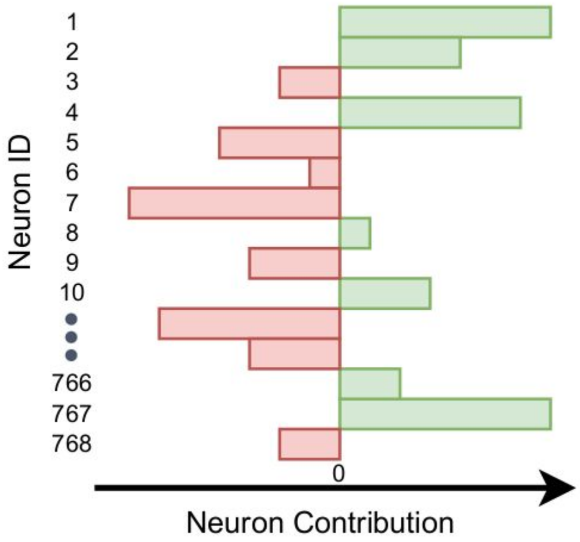
GrammarSHAP: bottom-up merges tokens based on grammar constituents.



Approaches Explaining Different Models (C2)

- Model-agnostic methods are flexible. But stricter model assumptions are a great recipe for faster, more accurate, and more fine-grained explanations.
  (This can be seen already in DeepSHAP and TreeSHAP vs KernelSHAP)

Neuron Shapley: target DNNs to quantify how each neuron contributes to a single prediction and overall model performance.



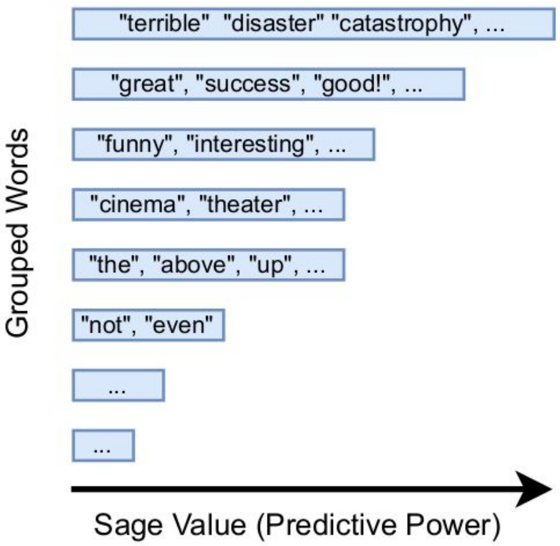
Producing Different Explanation Types (C4)

- SHAP is made for local explanations based on feature attribution.
- The broad applicability of Shapley Values suits also different settings.

SAGE: the version of SHAP for global explanation (about entire dataset).
Caveat: for NLP feature set is huge.
Trick: group tokens based on relations.

Honorable Mention

ConceptSHAP: SHAP-based method for concept explanations. Unsupervised + offers completeness score.



Modifying Core Assumptions (C3)

- Some SHAP assumptions can be at times simplistic and/or restrictive.

Causal Shapley & Shapley Flow:
leverage causal graph and causal ordering to encode feature dependencies.

More Efficient Shapley Values Estimation (C5)

- SHAP addresses the unfeasibility of computing exact Shapley Values. However…

C-Shapley: reduces the number of coalitions considered by only grouping up tokens that interact (e.g. adjacent words/nodes)

NLP Relevance and Recommendations

- We assess each reviewed method based on *availability of implementations*, *suitability for text data*, and *conceptual complexity*.

Tier 1	Tier 2	Tier 3	Tier 4
Ready Off-the-Shelf	*Adaptable*	*Potentially Applicable*	*Not Relevant*

- Based on our findings, we provide recommendations for NLP use cases

Debugging / Biases? Neuron Shapley

Concept Explanations? ConceptSHAP

Causal Dependencies in Data? Shapley Flow / ASV, ..

Need Hierarchical Structure? HEDGE

Global Understanding? SAGE

OUR REVIEW

Everything in one table!

Method	Categories	Description	NLP Applicability / Implementation
SHAP (Lundberg and Lee, 2017)		The original SHAP framework including the methods: KernelSHAP, LinearSHAP, DeepSHAP, etc.	*Ready Off-the-Shelf* Python
AVA (Bhatt et al., 2020)	**(C5)**	Combines the explanations of nearest neighbors to explain a given instance	*Adaptable* n.a.
ASV (Frye et al., 2019)	**(C1) (C3)**	Relaxes the symmetry axiom of Shapley values to incorporate causal structure into explanations	*Potentially Applicable* R
BShap (Sundararajan and Najmi, 2020)	**(C4) (C5)**	Baseline approach to facilitate comparison between different Shapley value based methods	*Adaptable* n.a.
C- and L-Shapley (Chen et al., 2018)	**(C3) (C5)**	Efficient feature attribution method that models data as a graph by considering only neighboring features	*Ready Off-the-Shelf* TensorFlow
CASV (Singal et al., 2019)	**(C1) (C2) (C3) (C4)**	Shapley value adaptation to account for counterfactuals by adhering to the Rubin Causal Model	*Not Relevant* n.a.
Causal Shapley (Heskes et al., 2020)	**(C1) (C3)**	Computing feature importance on data with (partial) causal ordering using Pearl's do-calculus	*Potentially Applicable* R
ConceptSHAP (Yeh et al., 2020)	**(C4)**	Unsupervised discover of concepts inherent to the data and model based on Shapley values	*Ready Off-the-Shelf* PyTorch
DASP (Ancona et al., 2019)	**(C3) (C5)**	Polynomial-time approximation of Shapley values in DNNs	*Adaptable* TensorFlow
Data Shapley (Ghorbani and Zou, 2019)	**(C4)**	Shapley-based importance attribution method for individual data instances in the training set	*Potentially Applicable* TensorFlow
DeepSHAP v2 (Chen et al., 2021)	**(C2) (C5)**	Computes efficiently SHAP values for DNNs with an extension to explain stacks of mixed model types	*Adaptable* n.a.
GrammarSHAP (Mosca et al., 2022a)	**(C1) (C3)**	Hierarchical explanations for text inputs based on the sentence grammatical structure	*Adaptable* n.a.
gSHAP (Tan et al., 2018)	**(C4)**	Generates intuitive Shapley-based global by aggregating local explanations	*Potentially Applicable* n.a.
h-SHAP (Teneggi et al., 2021)	**(C1) (C5)**	Hierarchical implementation of Shapley values for their efficient computation in image data	*Potentially Applicable* PyTorch
HEDGE (Chen et al., 2020)	**(C1) (C3)**	Hierarchical explanations based on feature interaction detection specifically for text data	*Ready Off-the-Shelf* PyTorch
Integrated Hessians (Janizek et al., 2021)	**(C5)**	Extension of Integrated Gradients to explain pairwise feature interactions in NNs	*Ready Off-the-Shelf* PyTorch

Takeaways and Future Work

- We reviewed 40+ SHAP- and Shapley-values-based explainability methods
- Identified five XAI research directions **+** classified each method
- Relevance of each method for NLP **+** use-case-based recommendations

Complete summary in one table!

Method	Categories	Description	NLP Applicability / Implementation
SHAP (Lundberg and Lee, 2017)		The original SHAP framework including the methods: KernelSHAP, LinearSHAP, DeepSHAP, etc.	*Ready Off-the-Shelf* Python
AVA (Bhatt et al., 2020)	**(C5)**	Combines the explanations of nearest neighbors to explain a given instance	*Adaptable* n.a.
ASV (Frye et al., 2019)	**(C1) (C3)**	Relaxes the symmetry axiom of Shapley values to incorporate causal structure into explanations	*Potentially Applicable* R
BShap (Sundararajan and Najmi, 2020)	**(C4) (C5)**	Baseline approach to facilitate comparison between different Shapley value based methods	*Adaptable* n.a.
C- and L-Shapley (Chen et al., 2018)	**(C3) (C5)**	Efficient feature attribution method that models data as a graph by considering only neighboring features	*Ready Off-the-Shelf* TensorFlow
CASV (Singal et al., 2019)	**(C1) (C2) (C3) (C4)**	Shapley value adaptation to account for counterfactuals by adhering to the Rubin Causal Model	*Not Relevant* n.a.
Causal Shapley (Heskes et al., 2020)	**(C1) (C3)**	Computing feature importance on data with (partial) causal ordering using Pearl's do-calculus	*Potentially Applicable* R
ConceptSHAP (Yeh et al., 2020)	**(C4)**	Unsupervised discover of concepts inherent to the data and model based on Shapley values	*Ready Off-the-Shelf* PyTorch
DASP (Ancona et al., 2019)	**(C3) (C5)**	Polynomial-time approximation of Shapley values in DNNs	*Adaptable* TensorFlow
Data Shapley (Ghorbani and Zou, 2019)	**(C4)**	Shapley-based importance attribution method for individual data instances in the training set	*Potentially Applicable* TensorFlow
DeepSHAP v2 (Chen et al., 2021)	**(C2) (C5)**	Computes efficiently SHAP values for DNNs with an extension to explain stacks of mixed model types	*Adaptable* n.a.
GrammarSHAP (Mosca et al., 2022a)	**(C1) (C3)**	Hierarchical explanations for text inputs based on the sentence grammatical structure	*Adaptable* n.a.
gSHAP (Tan et al., 2018)	**(C4)**	Generates intuitive Shapley-based global by aggregating local explanations	*Potentially Applicable* n.a.
h-SHAP (Teneggi et al., 2021)	**(C1) (C5)**	Hierarchical implementation of Shapley values for their efficient computation in image data	*Potentially Applicable* PyTorch
HEDGE (Chen et al., 2020)	**(C1) (C3)**	Hierarchical explanations based on feature interaction detection specifically for text data	*Ready Off-the-Shelf* PyTorch

Thank you!!

Edoardo Mosca

Ferenc Szigeti

Stella Tragianni

Daniel Gallagher

Georg Groh