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

Anonymous ACL submission

Abstract

Interpreting NLP models is fundamental for their development as it can shed light on hidden properties and unexpected behaviors. However, while transformer architectures exploit contextual information to enhance their predictive capabilities, most of the available methods to explain such predictions only provide importance scores at the word level. This work addresses the lack of feature attribution approaches that also take into account the sentence structure. We extend the SHAP framework by proposing GrammarSHAP—a model-agnostic explainer leveraging the sentence's constituency parsing to generate hierarchical importance scores.

1 Introduction

Deep learning models have raised the bar in terms of performance in a variety of *Natural Language Processing* (NLP) tasks (Vaswani et al., 2017; Devlin et al., 2019). However, also model complexity has been steadily increasing, which in turn hinders the interpretability of their predictions. This is particularly true for transformer architectures, currently established as the state of the art in various applications but at the same time containing billions of parameters (Brown et al., 2020).

Local explanations have become a popular tool to understand and interpret models' decisions (Madsen et al., 2021; Arrieta et al., 2020). These—besides increasing the public's trust in machine learning systems—can uncover unwanted behaviors such as unintended bias (Madsen et al., 2021; Dixon et al., 2018).

Feature attribution explanations are the most commonly used and can highlight parts of the input text that are relevant for the obtained outcome (Lundberg and Lee, 2017; Ribeiro et al., 2016). Almost all available methods, however, can only attribute a relevance score to single words. This is highly unintuitive as natural language in human communication can be very articulated and context-dependent. Indeed, a word's neighborhood can drastically alter its intended message and sentiment.

Our work focuses on generating explanations that account for the language structure. More specifically, we build hierarchical explanations that attribute relevance scores to sentence constituents at multiple levels. In contrast to previous work addressing the same issue (Chen et al., 2020; Chen and Jordan, 2020), we build our approach as an extension of SHAP (Lundberg and Lee, 2017)—a local explainability framework renowned for its solid theoretical background. Our contribution can be summarized as follows:

(1) We design GrammarSHAP, a model-agnostic approach for generating multi-level explanations that consider the text's structure and its constituents. More specifically, a constituency parsing layer for multi-word tokens selection is added before an adapted KernelSHAP explainer.

(2) We propose to drop the SHAP standard background dataset and use masking tokens instead. This reduces unwanted artifacts in the generated explanations and speeds up the approach's run time.

(3) We qualitatively compare our method to existing ones in terms of explanation quality and necessary computational effort.

2 Related Work

Several local explainability techniques exist to interpret predictions produced by NLP models (Arrieta et al., 2020). Among them, *features attribution* (or *feature relevance*) approaches quantify each input component's contribution to the model's output, i.e. how each feature affects the observed prediction. Methods in this category are available in a large variety: gradient-based (Simonyan et al., 2014; Sundararajan et al., 2017), neural-

network specific e.g. LRP (Bach et al., 2015) and DeepLIFT (Shrikumar et al., 2017), and model-agnostic e.g. LIME (Ribeiro et al., 2016). SHAP (Lundberg and Lee, 2017)—particularly relevant for our methodology—is by many considered to be a gold standard thanks to its solid theoretical background and broad applicability. This framework builds a unified view of methods like LIME, LRP, and DeepLIFT and the game-theoretic concept of Shapley values (Shapley, 1953).

More recent works address the limitations of word-level relevance scores by focusing on phrase-level and hierarchical explanations. The proposed approaches analyze and quantify words' interactions through exhaustive search (Tsang et al., 2018), combining their contextual decomposition scores (Singh et al., 2018), or via measuring SHAP interaction values along a predefined tree structure (Lundberg et al., 2018). Chen and Jordan (2020) combines a linguistic parse tree with Banzhaf values (Banzhaf III, 1964) to capture meaningful interactions in text inputs. (Chen et al., 2020), instead, propose to detect directly feature interaction without resorting to external structures. They propose a hierarchical explainability method that, in a top-down fashion, breaks down text components in shorter phrases and words based on the weakest detected interactions.

3 Methodology

We extend the SHAP framework (Lundberg and Lee, 2017) by proposing a model-agnostic explainer that considers the text's structural dependencies to generate importance scores at multiple levels. In particular, we couple a constituency parsing layer to hierarchically select multi-word tokens with a custom version of KernelSHAP adapted for improved efficiency and run-time. Figure 1 presents an overview of the methodological pipeline proposed in this work.

3.1 Token Selection via Constituency Parsing

To hierarchically construct multi-word tokens in a way that reflects the sentence structure, we leverage constituency parsing to group together tokens based on their grammatical interactions. To this end, we choose a state-of-the-art constituency parser: the Berkeley Neural Parser (Kitaev and Klein, 2018).

We iterate over parsed sentences from the single-word level ($depth = 0$) until the complete sentences are grouped up as a single token ($depth = N$). Additionally, we provide a library to retrieve groups of words at any depth, constituents, and combinations thereof. Our implementation also handles inconsistencies between the word-tokenization of the constituency parser and BERT. This is necessary as BERT's tokenizer uses sub-word tokens to represent OOV words and the Berkley Neural Parser[1] only allows full words as input.

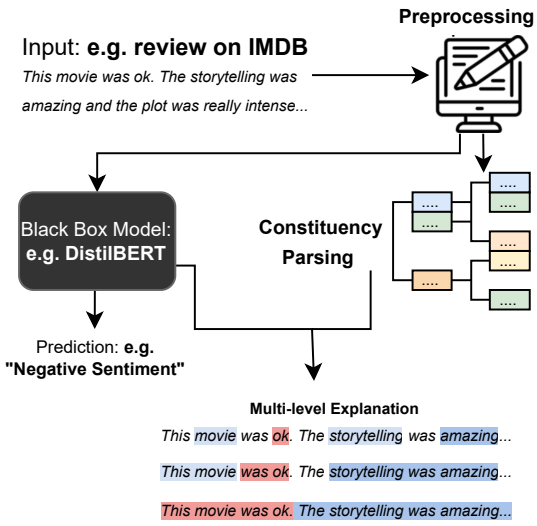

Figure 1: Overview of the proposed methodology.

3.2 Efficient Multi-Token Explainer

Our GrammarSHAP explainer directly extends the KernelSHAP method from Lundberg and Lee (2017). As parsed sentences already provide a hierarchical structure of grammatically coherent tokens, our extension is not required to compute tokens interaction to construct importance scores for multi-word tokens.

KernelSHAP takes an input sample x, a predicting model f, and a background set of samples to be used when replacing tokens to compute feature importance. Tokens belonging to the background dataset are fed to the explainer during initialization. At explanation time, a linear system of all perturbed sentences and their corresponding model predictions is solved to determine the effect of each single feature.

The extension to multi-word tokens consists in feeding the explainer—i.e. KernelSHAP—with the indices corresponding to the features to be grouped. In the case of constituency parsed sentences, indices representing multi-token groups are always

[1] spacy.io/universe/project/self-attentive-parser

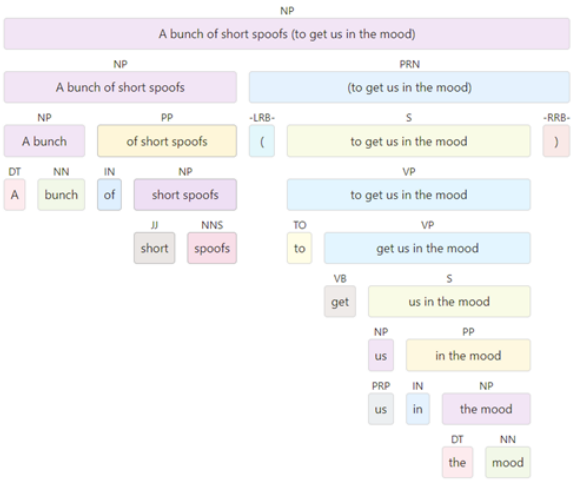

Figure 2: Example of sentence parsed with the Berkeley Neural Parser (Kitaev and Klein, 2018). Tokens are hierarchically grouped from single words (bottom level) to the whole sentence (top level)

adjacent in the input sentence. However, this is not a strict requirement for the following steps of our extension. To obtain group-level feature importance, we constrain the extended explainer to always replace a complete group of words with elements of the background dataset. Analogous to KernelSHAP, the expected effect of each feature group—i.e. its (multi-token) SHAP value—is calculated by solving the linear system of all perturbed sentences with their corresponding outcomes. In summary, our extension behaves like KernelSHAP but treats groups of tokens as single features.

While the calculation of SHAP values on multi-words tokens is a straightforward extension, it leads to several issues:

- **Computationally Expensive:** Computing importance scores for multiple levels further slows down the already inefficient KernelSHAP.
- **Unidirectional:** The explainer only highlights groups with the same sentiment as the overall sentence.
- **High Attribution for [SEP]:** The separation token changes the sentence length when used as replacement from the background data. This causes it to have high relevance for the classifier.

We address these limitation by replacing the background data with **[MASK]** tokens. This leads to a 60-folds speed up of the explainer that is not required to iterate over the background data. Moreover, **[SEP]** does causes explanation artifacts as it is excluded from the background data.

4 Empirical Findings

4.1 Data and Model to be Explained

To test and compare our method in practice, we pick a DistilBERT model (Sanh et al., 2019). Our choice is motivated by transformer architectures being established as the current state of the art in a variety of NLP applications.

Concerning the data, we pick the IMDb movie reviews (Maas et al., 2011) and the SST-2 datasets (Socher et al., 2013). For both, the *Hugging Face*[2] library provides a version of DistilBERT pre-trained on the task of binary sentiment analysis. The accuracy achieved is 93.7% and 91.3% respectively.

4.2 Existing SHAP Baselines

We compare explanations generated with Grammar-SHAP with two existing baselines from the SHAP framework (Lundberg and Lee, 2017):

(1) PartitionSHAP, i.e. the library's current recommended method for sentiment analisys on text data. Similarly to our method, it also utilizes **[MASK]** tokens for efficient word removal. However, features are only grouped via a binary tree and thus only token pairs are considered at a given hierarchical level.

(2) KernelSHAP, i.e. the library's standard for model-agnostic explanations. KernelSHAP only produces word-level explanations by default. But thanks to the additive nature of Shapley values, these can be added together according to the constituency parsing tree. We will refer to this custom hierarchical version of KernelSHAP as *Additive KernelSHAP*.

4.3 Comparison

The three methods substantially differ both in terms of generation times and explanation quality. Table 1 reports the average running time to produce an explanation. Figures 3 and 4 show—starting from the same input text—the explanations generated with each method. The text sample is particularly instructive as it contains both positive- and negative-sentiment sentences.

[2]https://huggingface.co/textattack/distilbert-base-uncased-imdb

Method	Running Time
PartitionSHAP	2
Add. KernelSHAP	3554 (∼1h)
GrammarSHAP	183 (∼3min)

Table 1: Average running time (in seconds) for GrammarSHAP compared to the existing SHAP baselines. The running time has been measured on 20 randomly selected samples (10 from IMDb and 10 from SST-2). Results were measured on a laptop machine: AMD Ryzen 5 CPU, Nvidia GPU GeForce GTX 1650, 16 GB DDR4 RAM.

PartitionSHAP is very efficient and the fastest method among the compared ones. However, it is quite coarse in grouping together tokens and fails to identify fine-grained contributions at the sub-sentence level. Additive KernelSHAP has an extremely long execution time and is the slowest of the three approaches. Moreover, it does not identify contributions opposite to the sample's overall sentiment. In contrast, GrammarSHAP is able to identify both negative and positive sentiments at different (hierarchical) levels of granularity. In terms of efficiency, GrammarSHAP does not match the performance of PartitionSHAP. However, its running time is still reasonable and does not raise issues for most applications.

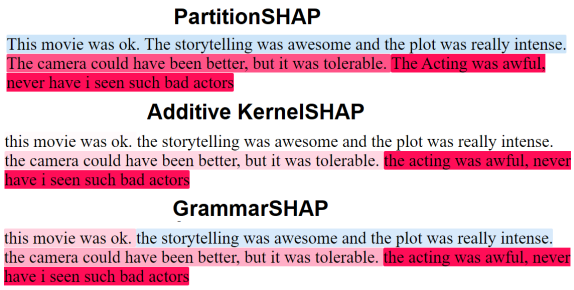

Figure 3: Comparison of three explanation methods for grouped features relevance (5th level). DistilBERT predicted the sample's sentiment as negative with a 79.5% confidence.

More examples of hierarchical GrammarSHAP explanation on (long) texts are provided in the appendix (see A). There, we also focus on presenting the explanations at different levels of granularity.

5 Limitations and Future Work

GrammarSHAP meaningfully extends the SHAP framework by providing efficient hierarchical explanations that reflect the sentence structure. However, limitations of our methodology and experimentation need to be acknowledged and motivate our future work.

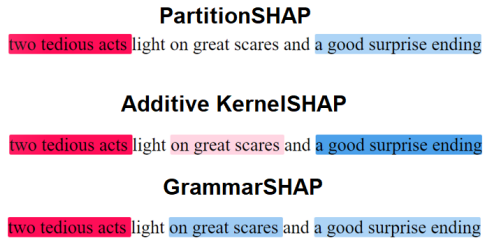

Figure 4: Comparison of three explanation methods for grouped features relevance (5th level). DistilBERT predicted the sample's sentiment as negative with a 81.8% confidence.

Regarding the explanation quality, our evaluation process is based on the introduced methodological improvements and on a qualitative analysis of the produced explanations. Although evaluation metrics for explanations are complex to define and have not been standardized yet, our comparison would considerably benefit from the usage of quantitative diagnostic properties (Atanasova et al., 2020) and word-level level metrics (Nguyen, 2018; Samek et al., 2016).

In terms of execution time, our method is still reasonable considering the granularity of contributions that it can detect. However, the necessity for further improvements in terms of efficiency becomes apparent when producing real-time explanations on the large scale.

6 Conclusion

In this work we proposed GrammarSHAP: a model-agnostic explainer for text data that accounts for the sentence structure and the existing grammatical relationships between the text tokens. Our approach leverages constituency parsing to extend the SHAP framework by providing hierarchical explanations that go beyond word-level attribution scores.

Our qualitative analysis of the produced explanation yields promising results as GrammarSHAP appears to identify more fine-grained contribution in structured text than its existing SHAP counterparts. At the same time, the usage of masking tokens instead of a background dataset considerably speeds up its execution in comparison with KernalSHAP. These properties make GrammarSHAP also suitable for long texts, especially if they contain sentences carrying different types of sentiment.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):130–140.

John F Banzhaf III. 1964. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19:317.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593.

Jianbo Chen and Michael Jordan. 2020. Ls-tree: Model interpretation when the data are linguistic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3454–3461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *NeurIPS 2017*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *NeurIPS 2017*.

Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games 2.28*, page 307–317.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency

maps. In *2nd International Conference on Learning Representations, ICLR 2014*.

Chandan Singh, W James Murdoch, and Bin Yu. 2018. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

Michael Tsang, Youbang Sun, Dongxu Ren, and Yan Liu. 2018. Can i trust you more? model-agnostic hierarchical explanations. *arXiv preprint arXiv:1812.04801*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS 2017*.

A Explanations Examples

Figure 5 shows an example of hierarchical GrammarSHAP explanation on a long text while 6 rather focuses on a shorter text. More examples can be found in the code repository attached to our submission. These are in the *Graphics Interchange Format* (GIF) format to visualize the transformation of the relevance scores through the various hierarchical levels.

depth = 1

although the actors do a convincing job playing the losers that parade across the screen , the fact that these characters are impossible to identify with had me looking at my watch a mere minutes into the film (and more than once after that) . the plot development is disjointed and slow , the verbal diarrhoea of the main character's only friend is practically insufferable , the base quality of most of the characters actions and the cavalier way in which they are treating is annoying . it is typical of ventura pons to put forth crass psychologically handicapped characters . however , this faux sociological analysis is a big step down from caricias or caresses , where the characters maltreat and despise each other for well founded reasons that play out during that film . in amor idiota we are forced to follow the meanderings of a truly subnormal intelligence as he stalks a severely depressed and detached woman . supposedly this is due to his own depression but the script doesn't support that . i won't give away the rest of the story just in case there are any masochists out there is he cured through his obsession or is the woman shocked out of her own depression through his unwavering attention ? even though i watched the whole thing i wasn't made to care even for a moment about either of them . if you can sit through all this prejudice , ignorance , betrayal , bad dialogue , flimsy philosophy , etc the camera work was pretty good and seems to be something inspired by the dogma group . the makeup also seemed to aim at showing these players in a raw and gritty light as it is the worst i've seen cayetana guillen cuervo in any of her movies (while in person she is actually attractive) .

depth = 5

although the actors do a convincing job playing the losers that parade across the screen , the fact that these characters are impossible to identify with had me looking at my watch a mere minutes into the film (and more than once after that) . the plot development is disjointed and slow , the verbal diarrhoea of the main character's only friend is practically insufferable , the base quality of most of the characters actions and the cavalier way in which they are treating is annoying . it is typical of ventura pons to put forth crass psychologically handicapped characters . however , this faux sociological analysis is a big step down from caricias or caresses , where the characters maltreat and despise each other for well founded reasons that play out during that film . in amor idiota we are forced to follow the meanderings of a truly subnormal intelligence as he stalks a severely depressed and detached woman . supposedly this is due to his own depression but the script doesn't support that . i won't give away the rest of the story just in case there are any masochists out there is he cured through his obsession or is the woman shocked out of her own depression through his unwavering attention ? even though i watched the whole thing i wasn't made to care even for a moment about either of them . if you can sit through all this prejudice, ignorance, betrayal, bad dialogue, flimsy philosophy, etc the camera work was pretty good and seems to be something inspired by the dogma group . the makeup also seemed to aim at showing these players in a raw and gritty light as it is the worst i've seen cayetana guillen cuervo in any of her movies (while in person she is actually attractive) . i suppose if the idea is that we should be

depth = 8

although the actors do a convincing job playing the losers that parade across the screen , the fact that these characters are impossible to identify with had me looking at my watch a mere minutes into the film (and more than once after that) . the plot development is disjointed and slow, the verbal diarrhoea of the main character's only friend is practically insufferable, the base quality of most of the characters actions and the cavalier way in which they are treating is annoying. it is typical of ventura pons to put forth crass psychologically handicapped characters. however , this faux sociological analysis is a big step down from caricias or caresses , where the characters maltreat and despise each other for well founded reasons that play out during that film . in amor idiota we are forced to follow the meanderings of a truly subnormal intelligence as he stalks a severely depressed and detached woman . supposedly this is due to his own depression but the script doesn't support that. i won't give away the rest of the story just in case there are any masochists out there is he cured through his obsession or is the woman shocked out of her own depression through his unwavering attention ? even though i watched the whole thing i wasn't made to care even for a moment about either of them. if you can sit through all this prejudice, ignorance, betrayal, bad dialogue, flimsy philosophy, etc the camera work was pretty good and seems to be something inspired by the dogma group . the makeup also seemed to aim at showing these players in a raw and gritty light as it is the worst i've seen cayetana guillen cuervo in any of her movies (while in person she is actually attractive).

Figure 5: Explanation generated with GrammarSHAP on a long IMDB review with negative-sentiment prediction of 91.7%. From top to bottom, relevance scores at the 1st, 5th and 8th hierarchical level.

depth = 2

klein , charming in comedies like american pie and dead on in election , delivers one of the saddest action hero performances ever witnessed

depth = 4

klein , charming in comedies like american pie and dead on in election , delivers one of the saddest action hero performances ever witnessed

depth = 8

klein, charming in comedies like american pie and dead on in election, delivers one of the saddest action hero performances ever witnessed

Figure 6: Explanation generated with GrammarSHAP on a short SST-2 review with negative-sentiment prediction of 91.6%. From top to bottom, relevance scores at the 2nd, 4th and 8th hierarchical level.