BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers

Enja Kokalj
Jožef Stefan International
Postgraduate School
Jožef Stefan Institute
enja.kokalj@ijs.si

Blaž Škrlj
Jožef Stefan International
Postgraduate School
Jožef Stefan Institute

Nada Lavrač
Jožef Stefan International
Postgraduate School
Jožef Stefan Institute

Senja Pollak
Jožef Stefan International
Postgraduate School
Jožef Stefan Institute

Marko Robnik-Šikonja
Faculty for Computer and Information Science
Ljubljana

Abstract

Transformer-based neural networks offer very good classification performance across a wide range of domains, but do not provide explanations of their predictions. While several explanation methods, including SHAP, address the problem of interpreting deep learning models, they are not adapted to operate on state-of-the-art transformer-based neural networks such as BERT. Another shortcoming of these methods is that their visualization of explanations in the form of lists of most relevant words does not take into account the sequential and structurally dependent nature of text. This paper proposes the TransSHAP method that adapts SHAP to transformer models including BERT-based text classifiers. It advances SHAP visualizations by showing explanations in a sequential manner, assessed by human evaluators as competitive to state-of-the-art solutions.

1 Introduction

Recent wide spread use of deep neural networks (DNNs) has increased the need for their transparent classification, given that DNNs are black box models that do not offer introspection into their decision processes or provide explanations of their predictions and biases. Several methods that address the interpretability of machine learning models have been proposed. Model-agnostic explanation approaches are based on perturbations of inputs. The resulting changes in the outputs of the given model are the source of their explanations. The explanations of individual instances are commonly visualized in the form of histograms of the most impactful inputs. However, this is insufficient for text-based classifiers, where the inputs are sequential and structurally dependent.

We address the problem of incompatibility of modern explanation techniques, e.g., SHAP (Lundberg and Lee, 2017), and state-of-the-art pretrained transformer networks such as BERT (Devlin et al., 2019). Our contribution is twofold. First, we propose an adaptation of the SHAP method to BERT for text classification, called TransSHAP (Transformer-SHAP). Second, we present an improved approach to visualization of explanations that better reflects the sequential nature of input texts, referred to as the TransSHAP visualizer, which is implemented in the TransSHAP library.

The paper is structured as follows. We first present the background and motivation in Section 2. Section 3 introduces TransSHAP, an adapted method for explaining transformer language model such as BERT, which includes the TransSHAP visualizer for improved visualization of the generated explanations. Section 4 presents the results of an evaluation survey, followed by the discussion of results and the future work in Section 5.

2 Background and motivation

We first present the transformer-based language models, followed by an outline of perturbation-based explanation methods, in particular the SHAP method. We finish with the overview of visualizations for prediction explanations.

BERT (Devlin et al., 2019) is a large pretrained language model based on the transformer neural network architecture (Vaswani et al., 2017). Nowadays, BERT models exist in many mono- and multilingual variants. Fine-tuning BERT-like models to a specific task produces state-of-the-art results in many natural language processing tasks, such as text classification, question answering, POS-

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 16–21
April 19, 2021 © Association for Computational Linguistics

tagging, dependency parsing, inference, etc.

There are two types of explanation approaches, general and model specific. The general explanation approaches are applicable to any prediction model, since they perturb the inputs of a model and observe changes in the model's output. The second type of explanation approaches are specific to certain types of models, such as support vector machines or neural networks, and exploit the internal information available during training of these methods. We focus on general explanation methods and address their specific adaptations for use in text classification, more specifically, in text classification with transformer models such as BERT.

The most widely used perturbation-based explanation methods are IME (Štrumbelj and Kononenko, 2010), LIME (Ribeiro et al., 2016), and SHAP (Lundberg and Lee, 2017). Their key idea is that the contribution of a particular input value (or set of values) can be captured by 'hiding' the input and observing how the output of the model changes. In this work, we focus on the state-of-the-art explanation method SHAP (SHapley Additive exPlanations) that is based on the Shapley value approximation principle. Lundberg and Lee (2017) noted that several existing methods, including IME and LIME, can be regarded as special cases of this method.

We propose an adaptation of SHAP for BERT-like classifiers, but the same principles are trivially transferred to LIME and IME. To understand the behavior of a prediction model applied to a single instance, one should observe perturbations of all subsets of input features and their values, which results in exponential time complexity. Štrumbelj and Kononenko (2010) showed that the contribution of each variable corresponds to the Shapley value from the coalition game, where players correspond to input features, and the coalition game corresponds to the prediction of an individual instance. Shapley values can be approximated in time linear to the number of features.

The visualization approaches implemented in the explanation methods LIME and SHAP are primarily designed for explanations of tabular data and images. Although the visualization with LIME includes adjustments for text data, the resulting explanations are presented in the form of histograms that are sometimes hard to understand, as Figure 1 shows. The visualization with SHAP for the same sentence is illustrated in Figure 2. Here, the features with the strongest impact on the prediction correspond to longer arrows that point in the direction of the predicted class. For textual data this representation is non-intuitive.

Various approaches have been proposed to interpret neural text classifiers. Some of them focus on adapting existing SHAP based explanation methods by improving different aspects, e.g., the word masking (Chen and Ji, 2020), or reducing feature dimension (Zhao et al., 2020), while others explore the complex interactions between words (contextual decomposition) that are crucial when dealing with textual data but are ignored by other post-hoc explanation methods (Jin et al., 2019; Chen et al., 2020).

3 TransSHAP: The SHAP method adapted for BERT

Many modern deep neural networks, including transformer networks (Vaswani et al., 2017) such as BERT-like models, split the input text into subword tokens. However, perturbation-based explanation methods (such as IME, LIME, and SHAP) have problems with the text input and in particular subword input, as the credit for a given output cannot be simply assigned to clearly defined units such as words, phrases, or sentences. In this section, we first present the components of the new methodology and describe the implementation details required to make explanation method SHAP to work with state-of-the-art transformer prediction models such as BERT, followed by a brief description of the dataset used for training the model. Finally we introduce the TransSHAP visualizer, the proposed visualization method for text classification with neural networks. We demonstrate it using the SHAP method and the BERT model.

3.1 TransSHAP components

The model-agnostic implementation of the SHAP method, named Kernel SHAP[1], requires a classifier function that returns probabilities. Since SHAP contains no support for BERT-like models that use subword input, we implemented custom functions for preprocessing the input data for SHAP, to get the predictions from the BERT model, and to prepare data for the visualization.

Figure 3 shows the components required by SHAP in order to generate explanations for the

[1] We use the Kernel SHAP implementation of the SHAP method: https://github.com/slundberg/shap.

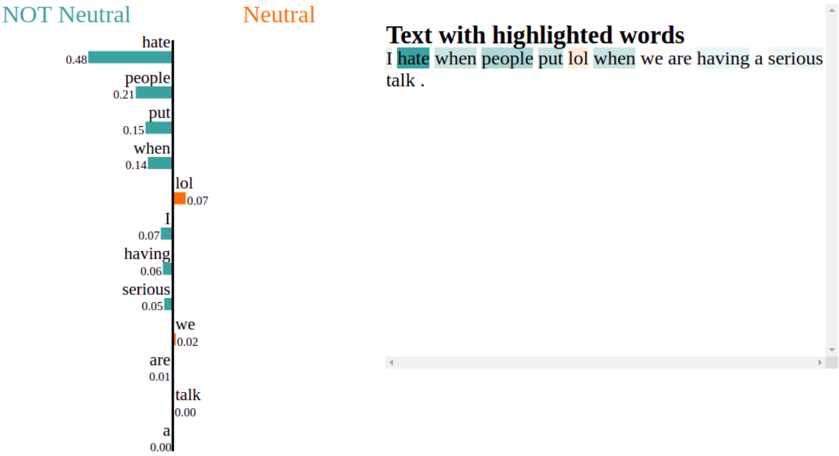

Figure 1: Visualization of prediction explanation with LIME.

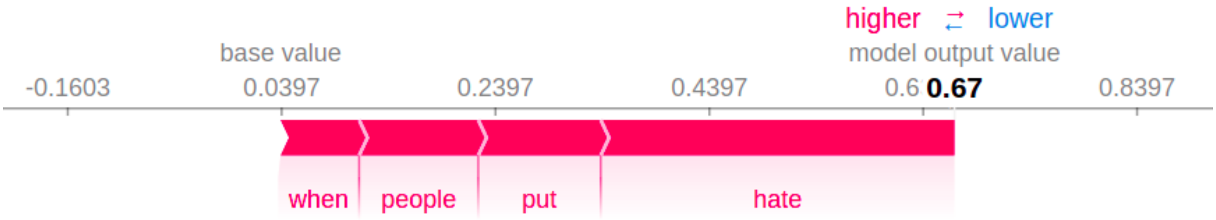

Figure 2: Visualization of prediction explanation with SHAP.

predictions made by the BERT model. The text data we want to interpret is used as an input to Kernel SHAP along with the special classifier function we constructed, which is necessary since SHAP requires numerical input in a tabular form.

To achieve this, we first convert the sentence into its numerical representation. This procedure consists of splitting the sentence into tokens and then preprocessing it. The preprocessing of different input texts is specific to their characteristics (e.g., tweets). The result is a list of sentence fragments (with words, selected punctuation marks and emojis), which serves as a basis for word perturbations (i.e. word masking). Each unique fragment is assigned a unique numerical key (i.e. index). We refer to a sentence, represented with indexes, as *an indexed instance*.

In summary, the TransSHAP's classifier function first converts each input instance into a word-level representation. Next, the representation is perturbed in order to generate new, locally similar instances which serve as a basis for the constructed explanation. This perturbation step is performed by the original SHAP. Then the perturbed versions of the sentence are processed with the BERT tokenizer that converts the sentence fragments to sub-word tokens. Finally, the predictions for the new locally generated instances are produced and returned to the Kernel SHAP explainer. With this modification, SHAP is able to compute the features' impact on the prediction (i.e. the explanation).

3.2 Datasets and models

We demonstrate our TransSHAP method on tweet sentiment classification. The dataset contains 87,428 English tweets with human annotated sentiment labels (positive, negative and neutral). For tweets we split input instances using the TweetTokenizer function from NLTK library[2], we removed apostrophes, quotation marks and all punctuation marks except for exclamation and question marks. We fine-tuned the CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020) on this classification task and the resulting model achieved the classification accuracy of 66.6%.

3.3 Visualization of a prediction explanation for the BERT model

To make a visualization of predictions better adapted to texts, we modified the histogram-based visualizations used in IME, LIME and SHAP for

[2] https://www.nltk.org

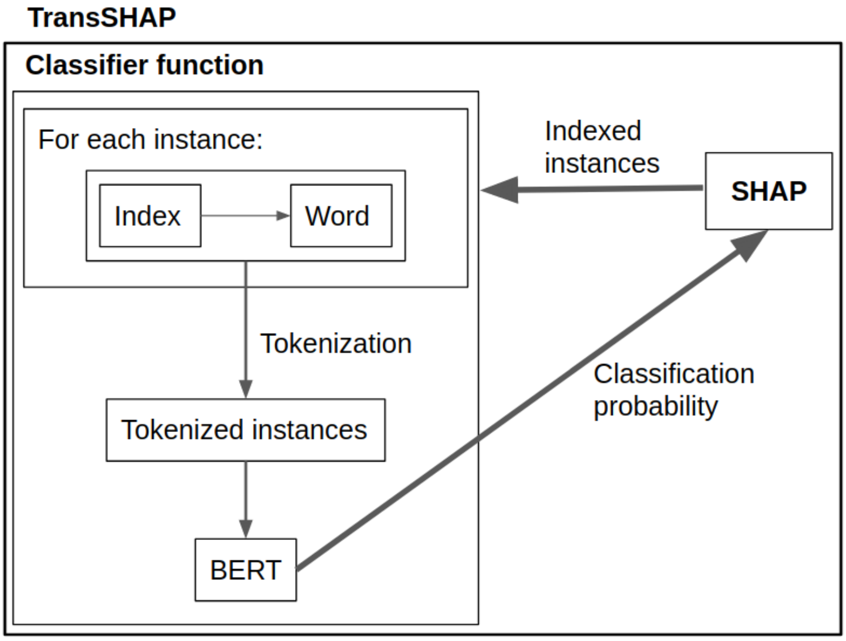

Figure 3: TransSHAP adaptation of SHAP to the BERT language model by introducing our classifier function.

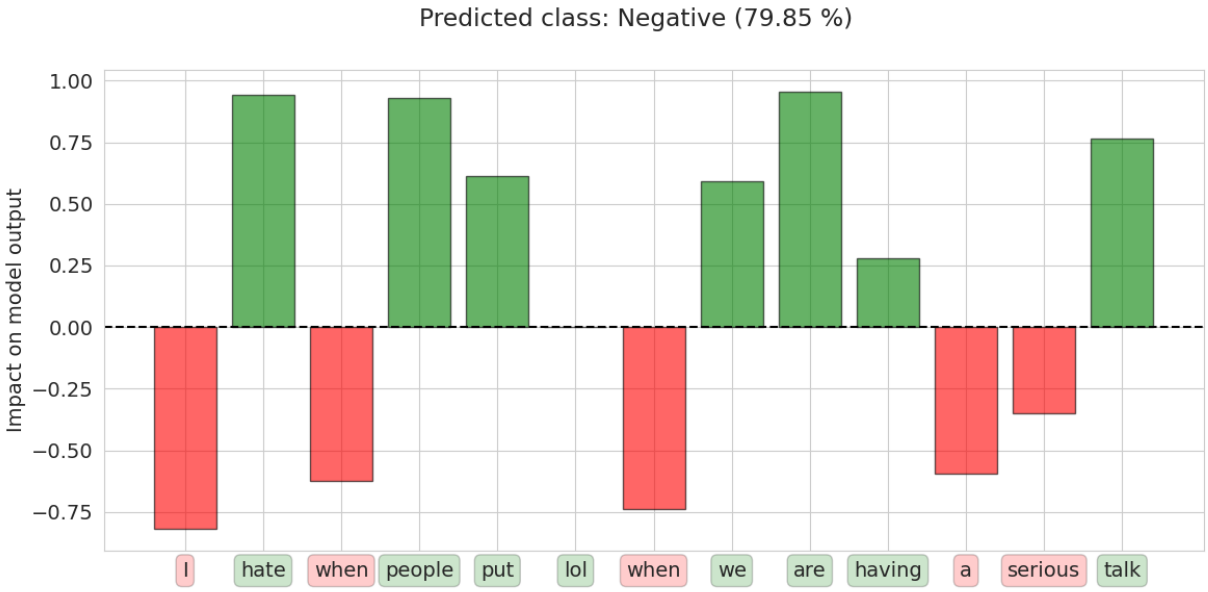

Figure 4: TransSHAP visualization of prediction explanations for negative sentiment. We obtained the features' contribution values with the SHAP method. It is evident that the word 'hate' strongly contributed to the negative sentiment classification, while the word 'lol' (laughing out loud) slightly opposed it.

tabular data. Figure 4 is an example of our visualization for explaining text classifications. It was inspired by the visualization used by the LIME method but we made some modifications with the aim of making it more intuitive and better adapted to sequences. Instead of the horizontal bar chart of features' impact on the prediction sorted in descending order of feature impact, we used the vertical bar chart and presented the features (i.e. words) in the order they appear in the original sentence. In this way, the graph allows the user to compare the direction of the impact (positive/negative) and also the magnitude of impact for individual words. The bottom text box representation of the sentence shows the words colored green if they significantly contributed to the prediction and red if they significantly opposed it.

4 Evaluation

We evaluated the novel visualization method using an online survey. The targeted respondents were researchers and PhD students not involved in the study that mostly had some previous experience with classifiers and/or their explanation methods. In the survey, the respondents were presented with three visualization methods on the same example: two visualizations were generated by existing libraries, LIME and SHAP, and the third one used our novel TransSHAP library. Respondents were asked to evaluate the quality of each visualization, suggest possible improvements, and rank the three methods.[3]

The results of 38 completed surveys are as follows. The most informative features of the visualization layout recognized by the users were the impact each word had on a prediction and the importance of the word contributions shown in a sequential view. The positioning of the visualization elements for each of the three methods was rated on the scale of 1 to 5. Our method achieved the highest average score of 3.66 (63.1% of the respondents rated it with a score of 4 or 5), second best was the LIME method with an average score of 3.13 (39.1% rated it with 4 or 5), and the SHAP method was rated as the worst with an average of 2.42 (81.5% rated it with 1 or 2). Regarding the question whether they would use each visualization method, LIME scored highest (44.7% voted "Yes"), TransSHAP closely followed (42.1% voted "Yes"), while SHAP was not praised (34.2% voted "Yes"). The overall ranking also corresponds to these results. LIME got the most votes (54.3%), TransSHAP was voted second best (40.0% of votes), and SHAP was the least desirable (5.7% of votes). In addition, we asked the participants to choose the preferred usage of the method out of the given options. The TransSHAP and SHAP methods were considered most useful for the purpose of debugging and bias detection, while the LIME method was also recognized as suitable for explaining a model to other researchers (usage in scientific articles).

5 Conclusion and further work

We presented the TransSHAP library, an extension of the SHAP explanation approach for transformer neural networks. TransSHAP offers a novel testing ground for better understanding of neural text classifiers, and will be freely accessible after acceptance of the paper (for review purposes available here: https://bit.ly/2UVY2Dy).

The explanations obtained by TransSHAP were quantitatively compared in a user survey, where we assessed the visualization capabilities, showing that the proposed TransSHAP's visualizations were simple, yet informative when compared to existing instance-based visualizations produced by LIME or SHAP. TransSHAP was scored better than SHAP, while LIME was scored slightly better in terms of overall user preference. However, in specific elements, such as positioning of the visualization elements, the visualization produced by TransSHAP is slightly better.

In further work, we plan to address problems of the perturbation-based explanation process when dealing with textual data. Currently, TransSHAP only supports random sampling from the word space, which may produce unintelligible and grammatically wrong sentences, and overall completely uninformative texts. We intend to take into account specific properties of text data and apply language models in the sampling step of the method. We plan to restrict the sampling candidates for each word based on their part of speech and general context of the sentence. We believe that better sampling will improve the speed of explanations and decrease the variance of explanations. Furthermore, the explanations could be additionally improved by expanding the features of explanations from individual words to larger textual units consisting of words that are grammatically and semantically linked.

Acknowledgements

We would like to acknowledge the Slovenian Research Agency (ARRS) for funding the first and the second author through young researcher grants and supporting other authors through the research program *Knowledge Technologies* (P2-0103) and the research project *Semantic Data Mining for Linked Open Data*. Further, we acknowledge the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

[3] The survey questions are available here: https://forms.gle/icpYvHH78oE2TCJt7.

References

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xisen Jin, Junyi Du, Zhongyu Wei, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Marco Túlio 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, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Proceedings of Text, Speech, and Dialogue, TSD 2020*. Accepted.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Erik Štrumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18.

Wei Zhao, Tarun Joshi, Vijayan Nair, and Agus Sudjianto. 2020. Shap values for explaining cnn-based text classification models.