Explainability Methods for Legal Judgment Prediction in Switzerland

Joel Niklaus, Bohua Chen, XinWei Li, YingYing Lee
University of Bern - FS 2022 Natural Language Processing (NLP) Seminar
Keywords: SHAP, Attention-head, Sequence Classification, LIME, Integrated Gradient, Counterfactual

Introduction

As machine learning algorithms continue to advance, their practical applications in industry, finance and healthcare are increasing. However, the effectiveness and credibility of these algorithms are limited if they are unable to provide valid explanations for the decisions they make. That is, the existence of a 'black box' makes it difficult to explain the rationale of the outcomes and leads to an opaqueness between input and output. Therefore, the model interpretability is what determines whether the model is reliable or not.

We will explore 7 explainable methods which are popular in the explainable AI, namely Attention-head, SHAP, Integrated Gradient, Counterfactual, LIME and Sequence Classification explainer. Each method has its own capability to help us in understanding the vectors which are representative of the words (i.e. description of court case) intuitively.

To be more precise, we are trying to figure out whether the LJP model can capture the important information from the vectors and how does it influence the outcome. Hence, we pay more attention to use the explanation methods to analyze 4 main aspects in NLP, namely Tokenization, Stop word removal, Lemmatization & Stemming and Part-of-speech tagging to validate the accuracy of the LJP model.

Data Description

We recently presented a dataset for legal judgment prediction including 85k Swiss federal supreme court decisions in three languages: DE,FR & IT. although we achieved up to 70% Macro- F1 Score, the models still work as black boxes and are thus not interpretable.

Acknowledgements / Reference

1. https://arxiv.org/pdf/1905.07188.pdf
2. https://www.aclweb.org/anthology/P19-3007.pdf
3. https://christophm.github.io/interpretable-ml-book/counterfactual.html
4. Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E., Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. (2015) Journal of Computational and Graphical Statistics, 24(1): 44-65:
5. https://arxiv.org/abs/1309.6392

Conclusion

Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production. There are plenty of explainability methods available, but NOT every explainability methods are suitable for Legal Judgement Prediction (LJP) model. We have explored 6 explainability methods as presented in the Poster and come to a conclusion:

Attention-Head view, **Shapley Additive exPlanations** and **Sequence Classification explainer** are easy-to-use for LJP model explainability, and their algorithm characteristics such as position embeddings, efficiency property of features and attributions sensitivity are intuitively interpreting the predictions of LJP model respectively.

LIME, **Integrated Gradient**, **Counterfactual**-explainability methods are feasible, however, it requires more effort for exploration such as a negate database construction

Methodology

Attention-head view

For this task, it may be useful to know which words the model finds similar (or different) between the two sentences.
Attention heads that draw connections between input sentences would thus be highly relevant.

The model view makes it easy to find these inter-sentence patterns, which are recognizable by their cross-hatch shape.

These heads can be further explored by clicking on them or accessing the attention-head view.

Shap-Additive

The method is a post-hoc interpretation of the model, in which the core of the interpretation is to calculate the Shapley Value of each of the characteristic variables.The SHAP value is the value assigned to each feature through that sample. To get the optimal prediction model and the optimal set of indicators it requires to explore the impact of each feature on the outcome.
Step1: Random sampling from the training data;
Step2:Randomly replace the features in sample x with those in z to get two new vectors.
Step 3: Calculate the marginal returns for each time and average them to get the Shapley value of the feature.

Sequence Classification

Transformers Interpret - Sequence Classification Explainer brings explainable AI to the transformers package with just 2 lines of code. It allows you to get word attributions and visualizations for those attributions simply.

This said method is suitable for NLP model explainer as it is built based on sequence of inputs over space or time

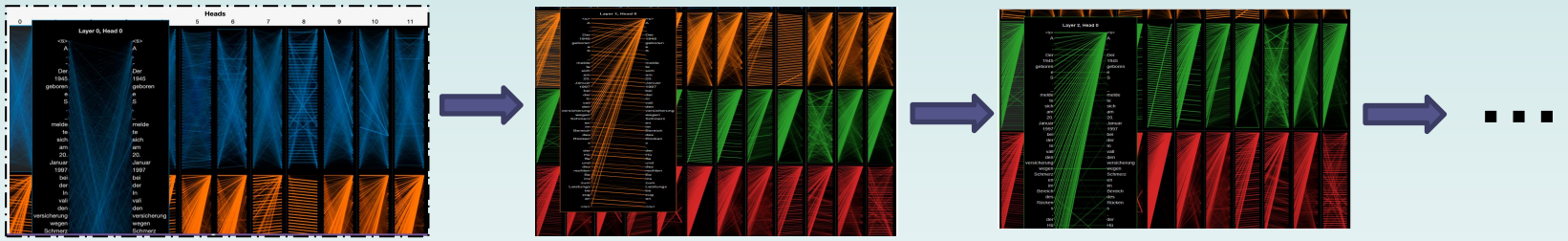
Positive attribution numbers indicate a word contributes positively towards the predicted class, while negative numbers indicate a word contributes negatively towards the predicted class.

Feature-based methods first transform sequences into feature vectors and then apply existing vectorial data classification methods

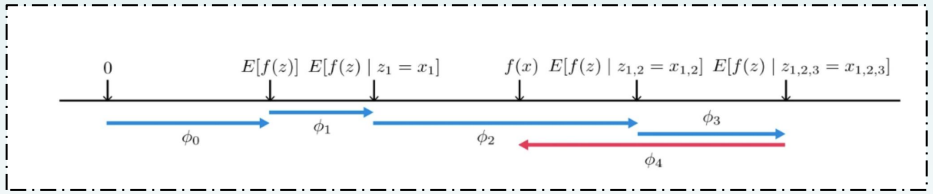
Legend: Negative Neutral Positive

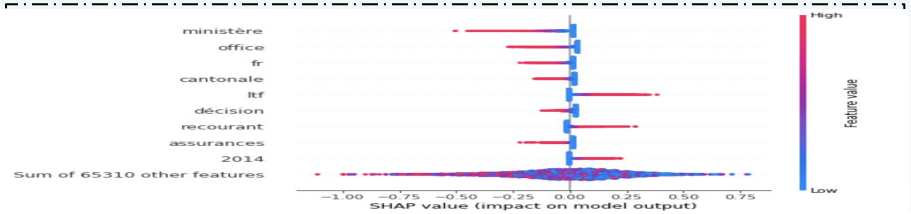
Results Interpretation

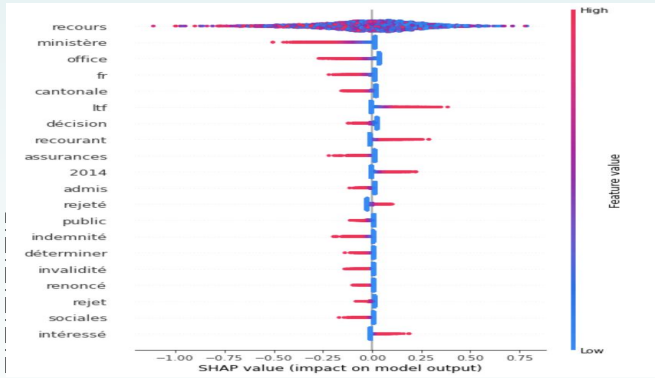
A.- Der 1945 geborene S._ meldete sich am 20. Januar 1997 bei dern validenversicherung wegen Schmerzen im Bereich des Rückens, der Hüfte und des rechten Beins zum Leistungsbezug an.



Layer 0, header 0 | Layer 1, header 0 | Layer 2, header 0 | Layer 11, header 0







Word Importance

[CLS] Sa ##ch ##ver ##halt : A . Mit Urteil 9 ##C _ 605 / 2014 vom 17 . September 2014 wies das Bundes ##gericht eine Be ##sch ##wer ##de des A . _ gegen einen Ent ##scheid des Vers ##icher ##ung ##sgericht ##s des Kantons Solo ##thur ##n vom 18 . Juni 2014 ab , so ##weit es darauf ein ##trat . B . A . _ ers ##ucht um Revision dieses Ent ##scheid ##s . Die Be ##sch ##wer ##de in jen ##em Verfahren sei gut ##zu ##hei ##ssen und die Sa ##che an das kanton ##ale Gericht zur ##Ã ##¼ ##ck ##zu ##weisen , damit dieses in kor ##rekt ##er Zusammen ##setzung neu ent ##scheid ##e . Es sei ein Schriften ##wechsel durch ##zu ##f ##Ã ##¼ ##hren und bei Ober ##richter C . _ eine Stellung ##nahme ein ##zu ##holen . [SEP]

True Label	Predicted Label	Attribution Label	Attribution Score
1	dismissal (0.99)	dismissal	4.05

"Urteil", "Bundesgericht", "Revision", "Es sei ein Schriftenwechsel" and "holen" contribute positively to "dismissal"
"C .", "Stellung" and "zu" contribute negatively to "dismissal"

Word Importance

[CLS] Sa ##ch ##ver ##halt : A . Mit St ##raf ##be ##fe ##hl vom 21 . Juli 2011 b ##Ã ##¼ ##ss ##te die Staat ##san ##walt ##schaft Len ##z ##burg - Aa ##rau X . _ wegen Wide ##r ##handlung gegen das [UNK] mit Fr . 200 . - - . Auf das hier ##ge ##gen erhoben ##e Revision ##sg ##esu ##ch vom 25 . Juni 2014 trat das Ober ##gericht des Kantons Aa ##rgau nicht ein . B . X . _ be ##ant ##rag ##t mit Be ##sch ##wer ##de in St ##raf ##sache ##n , der Be ##schluss des Ober ##gericht ##s sei auf ##zu ##heben und dieses an ##zu ##weisen , das Revision ##sg ##esu ##ch mater ##iel ##l zu be ##urt ##eile ##n . Die Oberst ##aat ##san ##walt ##schaft und das Ober ##gericht des Kantons Aa ##rgau verzi ##chten auf eine Verne ##hm ##lassung . [SEP]

True Label	Predicted Label	Attribution Label	Attribution Score
0	approval (0.87)	approval	3.78

"Die Oberstaatsanwaltschaft", "Obergericht des Kantons Aargau" and "verzichten" contribute positively to "approval"
"erhobene" and "lassung" contribute negatively to "approval"

Word Importance

[CLS] Fat ##ti : A . Nell ' ambito di un pro ##ced ##imento av ##viato dina ##nzi al Consiglio di Stato per den ##ega ##ta giustizia che vede ##va una cliente dell ' av ##v . A . _ opp ##osta al Municipio di X . _ la rappresenta ##nte del Governo ti ##cines ##e , l ' av ##v . l . _ , ha inde ##tto il 15 marzo 2012 un ' ud ##ienza ist ##rut ##toria . Ad ud ##ienza con ##clusa la pat ##roc ##inata dell ' av ##v . A . _ ha chi ##esto di ac ##cer ##tare la null ##it ##Ã di un alle ##gato della contro ##parte e dell ' ud ##ienza stessa , critica ##ndo ##ne lo svo ##lgi ##mento non ##ch ##Ã ##© la condotta dell ' av ##v . l . _ , la quale avrebbe agit ##o a suo sva ##nta ##ggio e d ' inte ##sa con la contro ##parte . [SEP]

True Label	Predicted Label	Attribution Label	Attribution Score
1	dismissal (0.99)	dismissal	3.45

"denegata", "giustizia", "critica", "suo", and "ggio" contribute positively to "dismissal"