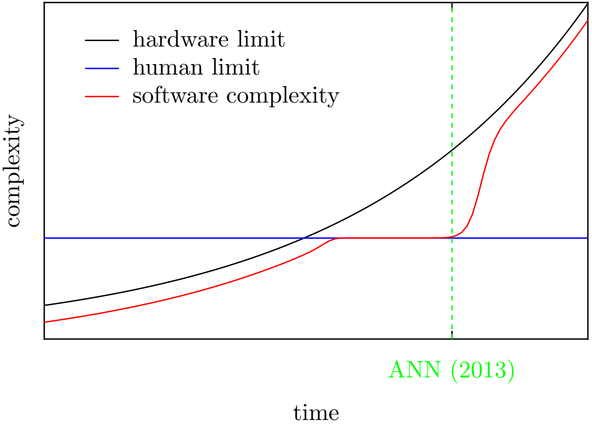
Tuning of Explainable Artificial Intelligence tools in the field of text analysis

Philipp Weinmann | 11. Juni 2021

Inhaltsverzeichnis

1. **Motivation**
2. **Related Works**
3. **Interpretability**
4. **Shap**
5. **Dataset**
6. **Results**
7. **Framework**
8. **Outlook**

Motivation: XAI in general



Motivation: XAI in general

XAI can be used to:

1. provide a hint to solve the problem without ANN

Motivation: XAI in general

XAI can be used to:

1. provide a hint to solve the problem without ANN
2. disqualify

Motivation: XAI in general

XAI can be used to:

1. provide a hint to solve the problem without ANN
2. disqualify
3. increase trust

Interpretability

- Simulatability

Interpretability

- Simulatability
- Decomposability

Interpretability

- Simulatability
- Decomposability
- Algorithmic transparency

[3]

Interpretability

- Simulatability
- Decomposability
- Algorithmic transparency

[3]

=> need for an explanation model

Interpretability: Example

Feature name	Feature importance
rutgers	0.040224
athos	0.036232
geneva	0.030274
1993	0.025009
christ	0.022898
article	0.021479
writes	0.019735
com	0.019473
paul	0.016807
don’t	0.014403

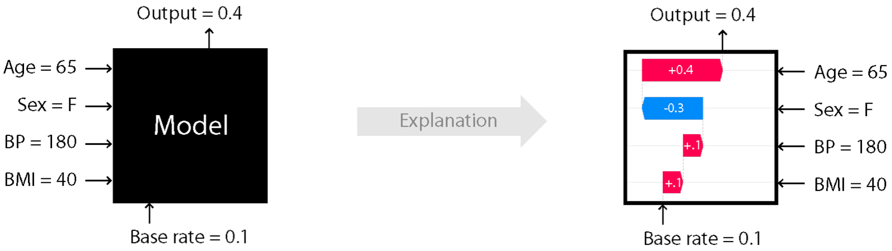
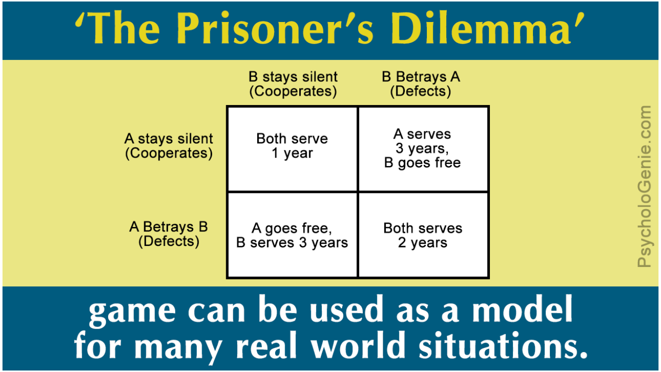


Figure: Visualization of a shap explanation [4]

Shap: Competitive Game theory



‘The Prisoner’s Dilemma’

	B stays silent (Cooperates)	B Betrays A (Defects)
A stays silent (Cooperates)	Both serve 1 year	A serves 3 years, B goes free
A Betrays B (Defects)	A goes free, B serves 3 years	Both serves 2 years

game can be used as a model for many real world situations.

Figure: Prisoners Dilemma [5]

Competitive Game theory: text example

- “**Jesus** word is the word of God” => 100% christian
- “word is the word of God” => 100% christian

Competitive Game theory: text example

- “**Jesus** word is the word of God” => 100% christian
- “word is the word of God” => 100% christian

- “word is the word of”

Our Dataset

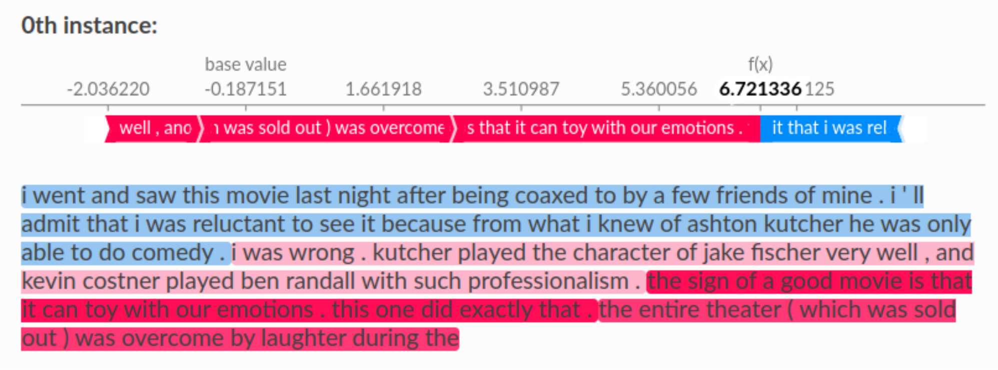
- 20Newsgroups[2]: Atheist and Christian emails.
- Binary tfidf-classifier

Results: Parameter

Text-Hierarchy
Word
Sentence
Paragraph
2-gram
...

Table: Parameter: Text Hierarchies

Results: Parameter



0th instance:

i went and saw this movie last night after being coaxed to by a few friends of mine . i ' ll admit that i was reluctant to see it because from what i knew of ashton kutcher he was only able to do comedy . i was wrong . kutcher played the character of jake fischer very well , and kevin costner played ben randall with such professionalism . the sign of a good movie is that it can toy with our emotions . this one did exactly that . the entire theater (which was sold out) was overcome by laughter during the

Figure: Explanation with different text hierarchy example[4].

Results: Calculation method

Context Influence

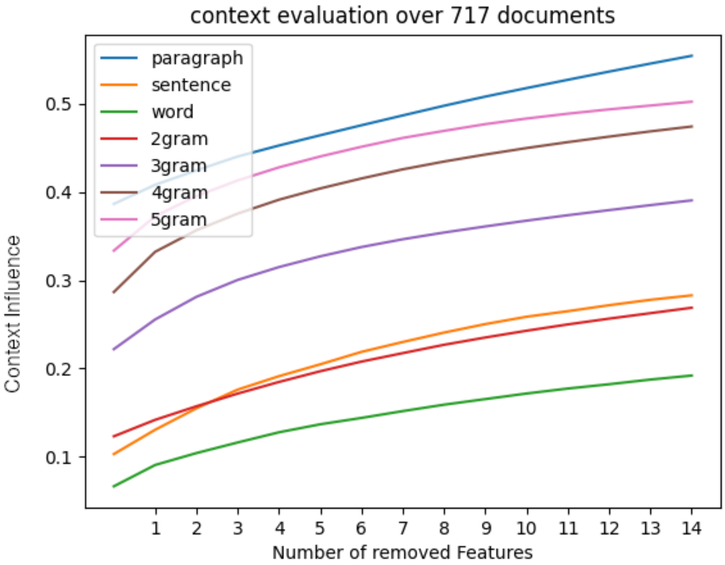
$$contextInfluence(w') = \\ |classificationScore(W) - classificationScore(W \backslash w') - shapFeatureImportance(w')|$$

$classificationScore(W)$: The classification score the classifier gives the text W

$classificationScore(W \backslash w')$: *The value the classifier gives the text W without the word w'*

$shapFeatureImportance(w')$: *The feature importance shap gives to the word w'*

$contextInfluence(w')$: *The importance of the context of the word w' according to shap*


context evaluation over 717 documents
paragraph
sentence
word
2gram
3gram
4gram
5gram
Context Influence
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of removed Features

Results: Recommendation

- Stop using words in XAI while using shap, try to use grammatical constructs like sentences: Our data shows a decrease by 89,33% of the context-influence per word presented to the user.

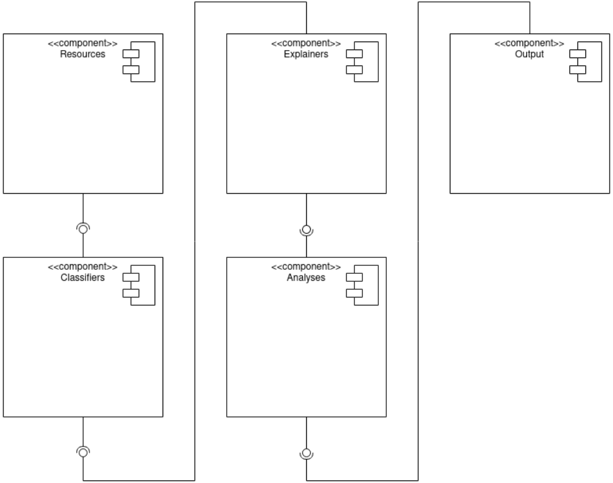


Figure: Framework overview

Figure: Outlook: Interactivity [1]

Thank you for listening

References

- Satyabrata Das. *Levels Of Interactivity In eLearning*. May 2021. URL: https://elearningindustry.com/levels-of-interactivity-elearning-modules.
- empty. *20 Newsgroups Dataset*. Ed. by empty. 2019. URL: http://people.csail.mit.edu/jrennie/20Newsgroups/.
- Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG].
- Scott M. Lundberg. *Welcome to the SHAP documentation*. URL: https://shap.readthedocs.io/en/latest/index.html.
- *Meaning of Prisoner's Dilemma With Real-life Examples*. Dec. 2014. URL: https://psychologenie.com/meaning-of-prisoners-dilemma-with-real-life-examples.