

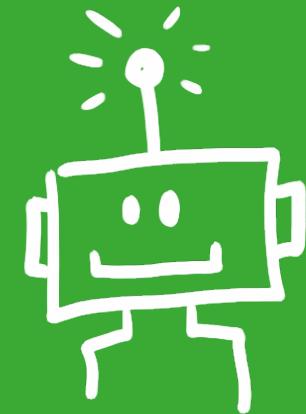


XAI meets Natural Language Processing

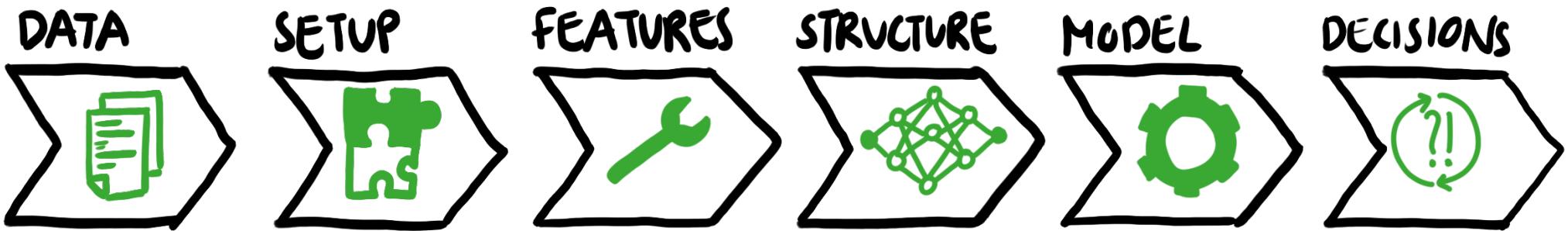
sovanta AG, 13.04.22

Larissa Haas

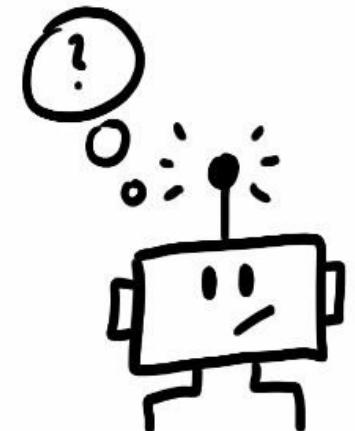
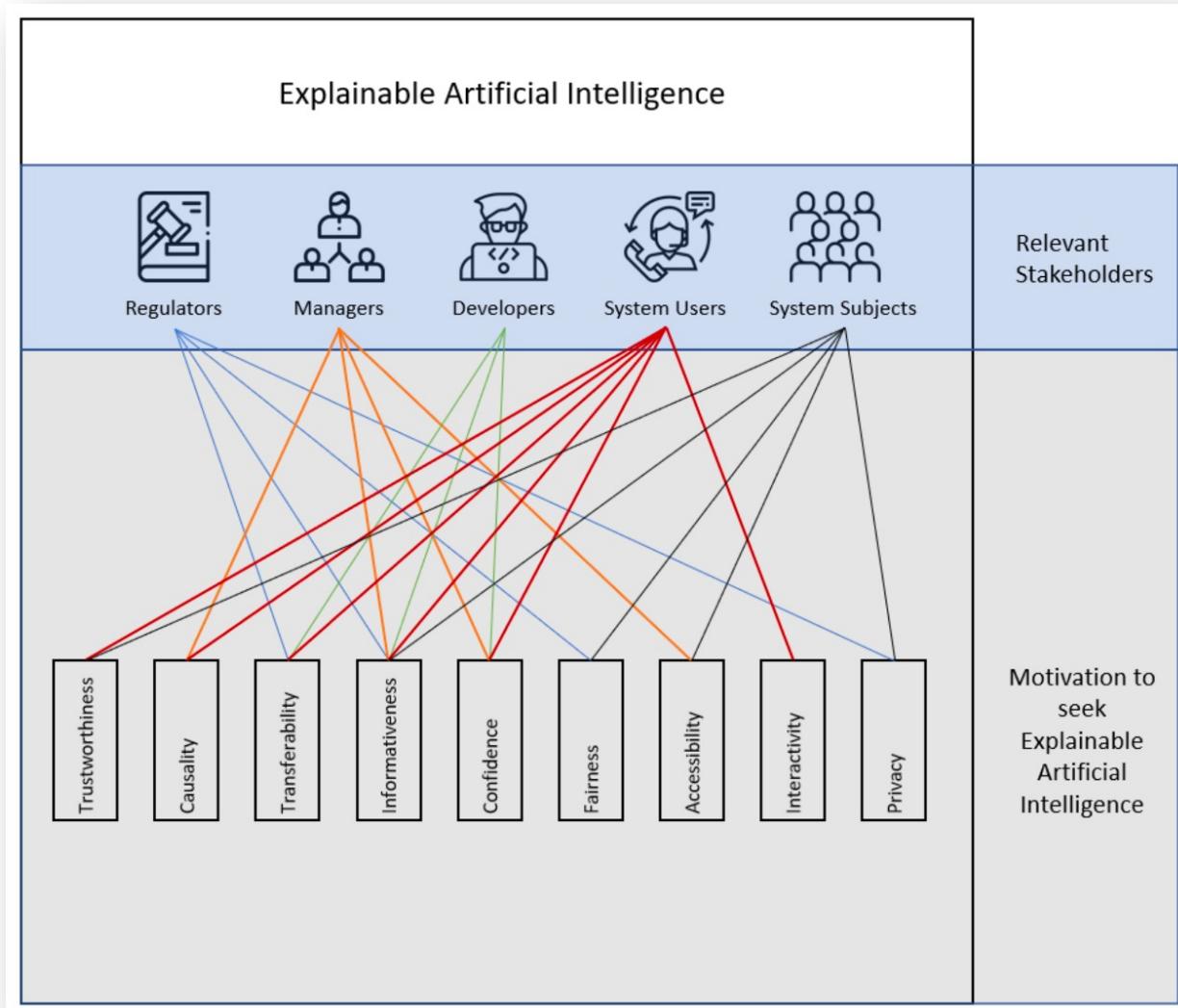
@l_r_ss_



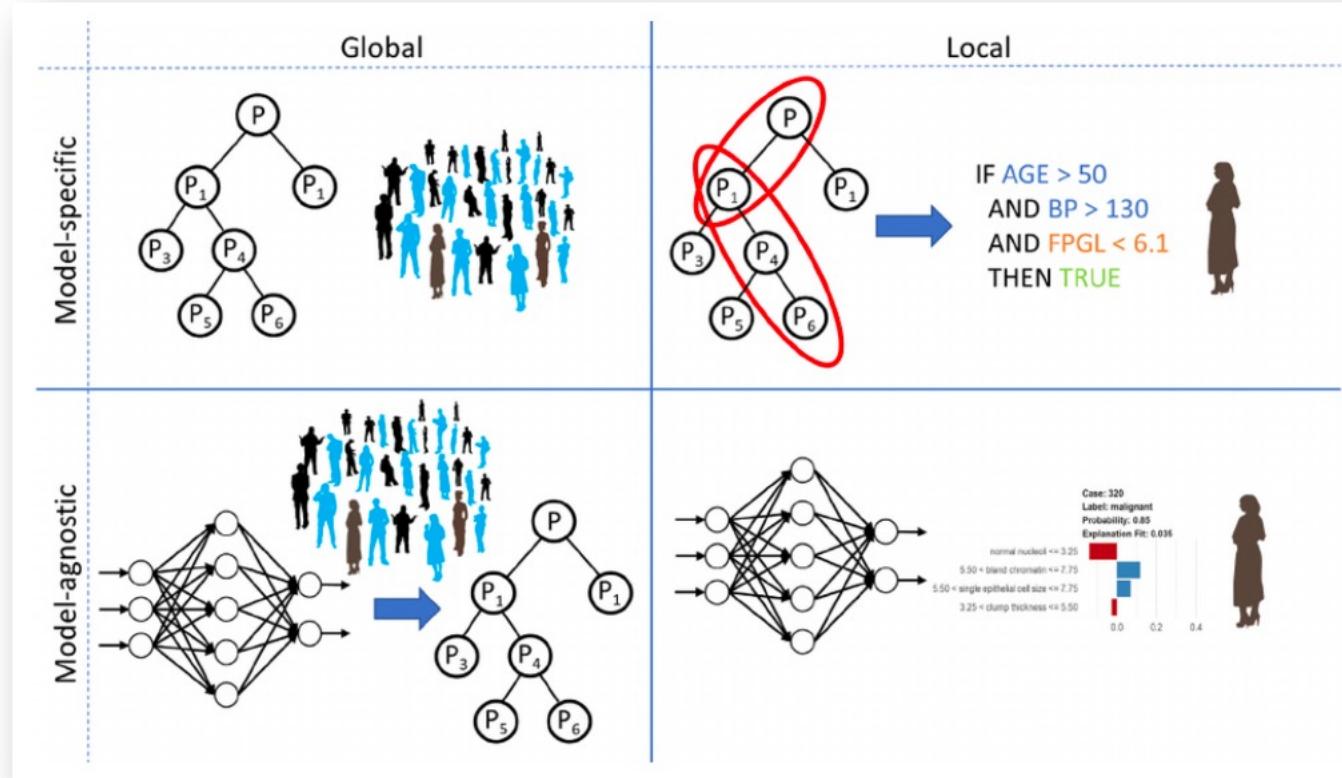
What is Explainable AI?



Why bother?



Common Approaches & Evaluations



Common Approaches & Evaluations

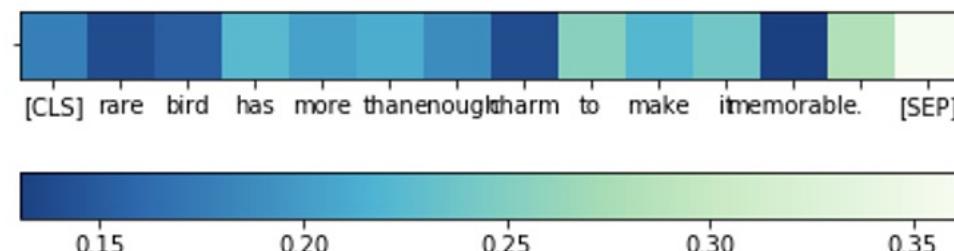
- Human annotations
- Feature importances and model structures
- Understandable models
- Correlations like LIME, SHAP, ELI5
- Graphical representations



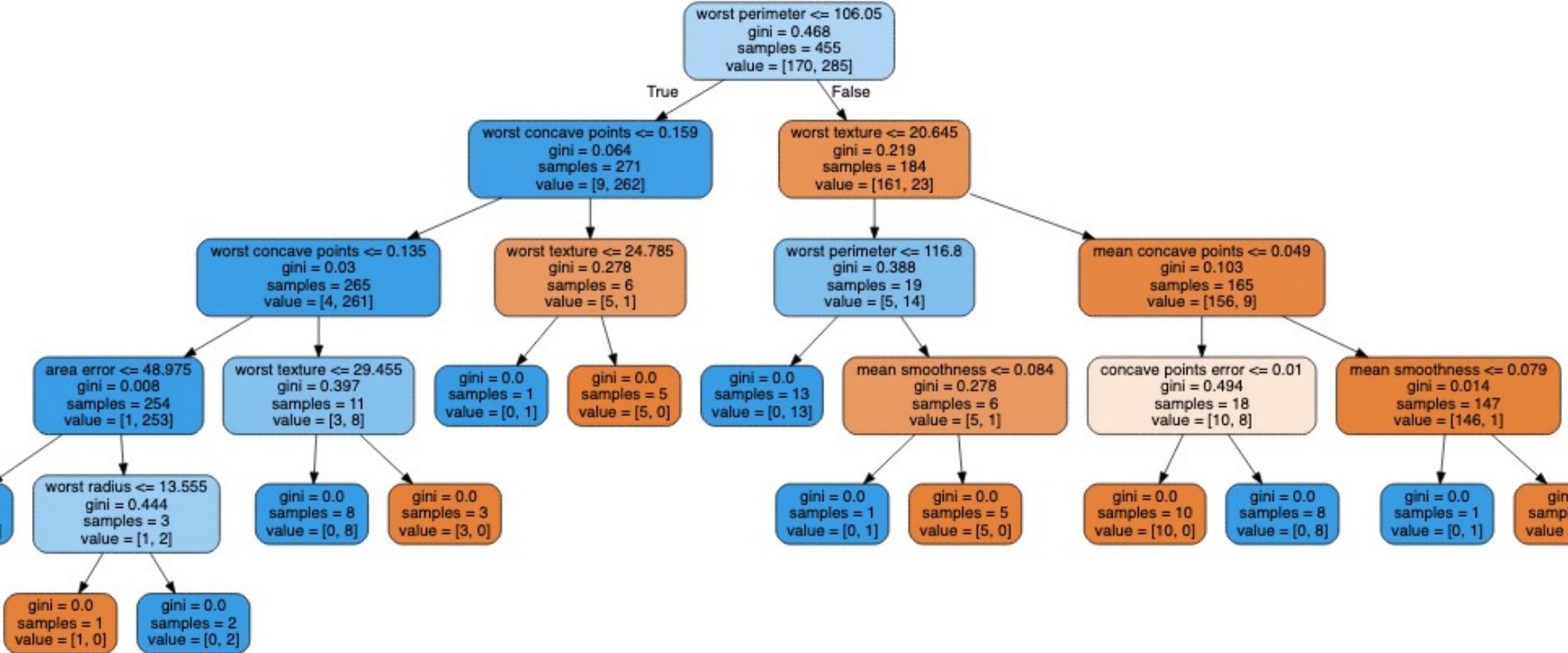
Feature Importance & Model Structures

- Use layers of BERT model to compute interpreter instances
- Sigma = the range that every word can change without changing **s** too much
- Sigma stands for the information loss of word \mathbf{x}_i when it reaches **s**

```
In [15]: interpreter_bert.visualize()
```



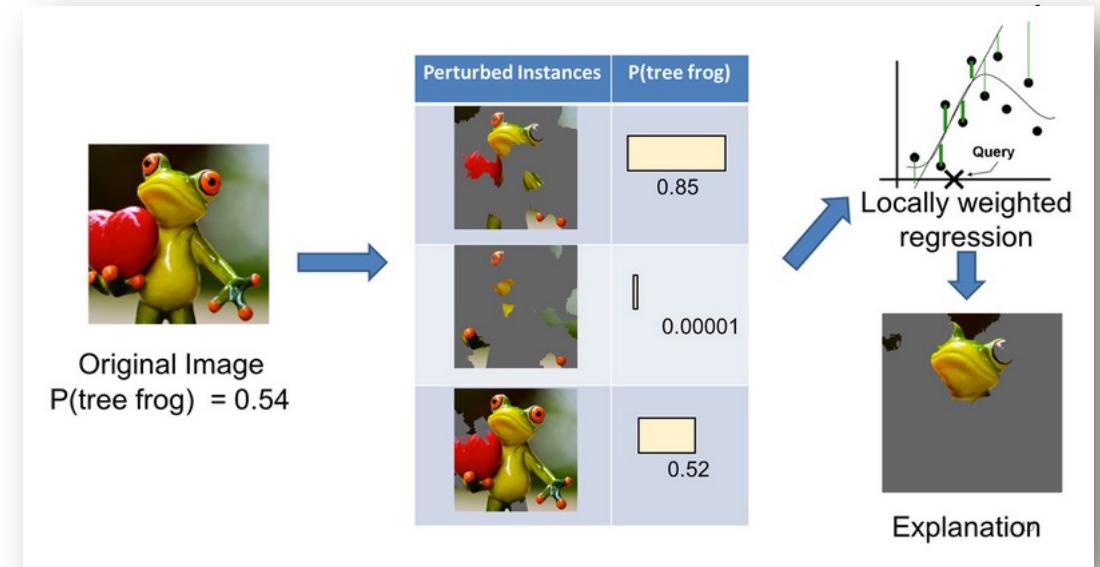
Understandable Models



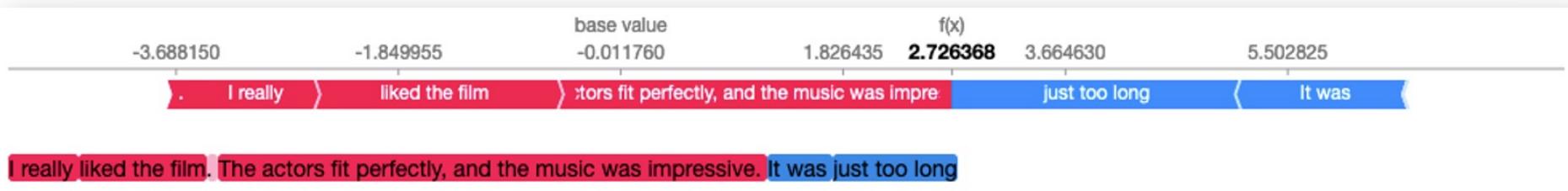
LIME & SHAP

- LIME (= Local Interpretable Model-agnostic Explanations)

- only locally



- SHAP values (= SHapley Additive exPlanations)



- locally and globally, but more costly

Difference between LIME & SHAP

- "For Christmas song visit my channel!" → SPAM

For	Christmas	Song	Visit	My	Channel	Prob	LIME	SHAP
1	0	1	1	0	0	0.17	0.5	0.42
0	1	1	1	1	0	0.17	0.67	0.68
1	0	0	1	1	1	0.99	0.67	0.68
1	0	1	1	1	1	0.99	0.83	0.91
0	1	1	1	0	0	0.17	0.5	0.42
0	0	0	0	0	1	0.83	0.17	0.91

- LIME: weighting on closeness to original
- SHAP: weighting on coalitions (high amount of 1s, high amount of 0s)

Example extended from <https://christophm.github.io/interpretable-ml-book/lime.html#lime>

```
In [31]: import eli5
```

```
eli5.explain_prediction(gnb, datarows["text_new"][2741], vec=tfid, target_names=['fake', 'true'])
```

```
Out[31]: y=fake (score -0.862) top features
```

Contribution?	Feature
+0.632	Highlighted in text (sum)
+0.230	<BIAS>

reuters news agency declares war on trump in the most perfect way, trump humiliated donald trump s treatment of journalists is so bad that the reuters news agency is now advising reporters to treat trump like they would treat a foreign dictator.on tuesday, trump and his team announced that they will no longer send surrogates to appear on cnn, an escalation of the war trump has been waging against the news network, which he calls fake news because they won t promote his agenda.trump has even praised fox news biased reporting as something cnn should copy.furthermore, trump and his team have repeatedly threatened and tried to intimidate journalists for doing their jobs just because they aren t writing puff pieces that kiss trump s ass.well, reuters is doing something about it.for 165 years, reuters has been bringing us news from around the world. they ve been in the most peaceful and democratic places, but they ve also been in war zones and reported on the most dangerous and tyrannical regimes in world history.in a message to staff on tuesday, reuters editor-in-chief steve adler advised his reporters to start dealing with trump the way they have dealt with brutal dictators in the past. it s not every day that a u.s. president calls journalists among the most dishonest human beings on earth or that his chief strategist dubs the media the opposition party, adler wrote. it s hardly surprising that the air is thick with questions and theories about how to cover the new administration. adler then revealed his solution for how reporters should handle trump.so what is the reuters answer? to oppose the administration? to appease it? to boycott its briefings? to use our platform to rally support for the media? all these ideas are out there, and they may be right for some news operations, but they don t make sense for

In [31]: `import eli5`

```
eli5.explain_prediction(gnb, datarows["text_new"][2741], vec=tfid, target_names=['fake', 'true'])
```

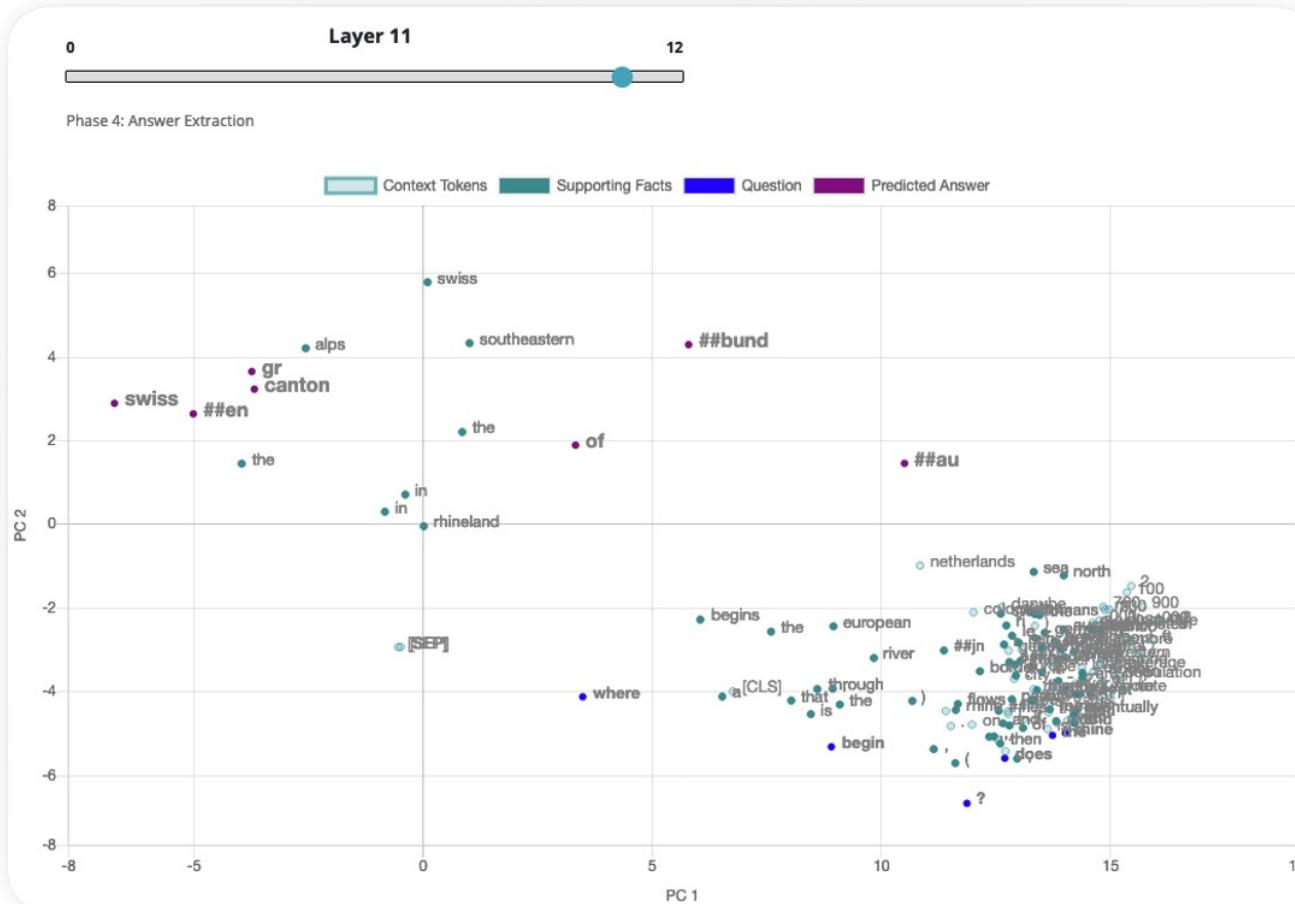
Out[31]: `y=fake (score -0.862) top features`

Contribution?	Feature
+0.632	Highlighted in text (sum)
+0.230	<BIAS>

reuters news agency declares war on trump in the most perfect way, trump humiliated donald trump s treatment of journalists is so bad that the reuters news agency is now advising reporters to treat trump like they would treat a foreign dictator.on tuesday, trump and his team announced that they will no longer send surrogates to appear on cnn, an escalation of the war trump has been waging against the news network, which he calls fake news because they won t promote his agenda.trump has even praised fox news s biased reporting repeatedly threatened and tried to intimidate journalists for doing their jobs just because they were doing something about it.for 165 years, reuters has been bringing us news from around the world, but they ve also been in war zones and reported on the most dangerous areas. reuters editor-in-chief steve adler advised his reporters to start dealing with trump today that a u.s. president calls journalists among the most dishonest human beings adler wrote. it s hardly surprising that the air is thick with questions and theories about how reporters should handle trump.so what is the reuters answer? to oppose the platform to rally support for the media? all these ideas are out there, and they may



Graphical Representations

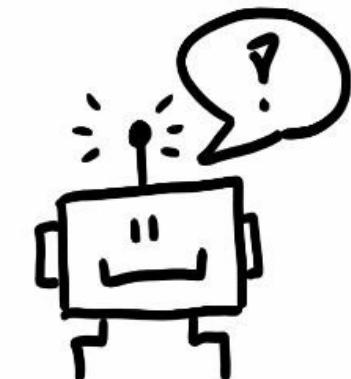


- good for word embeddings
- but only with dimensionality reduction

<https://visbert.demo.datexis.com/>

Common Approaches & Evaluations

- Human annotations
- Feature importances and model structures
- Understandable models
- Correlations like LIME, SHAP, ELI5
- Graphical representations



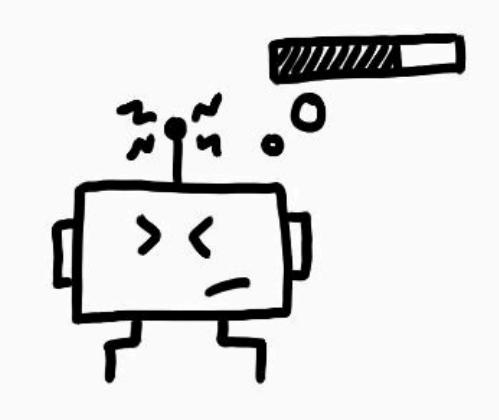
Common Approaches & Evaluations

approach	evaluation for NLP
human annotations	for all kinds of applications difficult, except if you have / create the right data
Feature importances & model structures	depends on model
understandable models & feature importances	depending on preprocessing: but mostly not applicable
correlations (LIME, SHAP, ELI5)	depending on preprocessing & model: can work quite well downside: slow can be manipulated by researcher
graphical representations	good e.g. for evaluating embeddings / word vectors, but also only via SVD / PCA / t-SNE

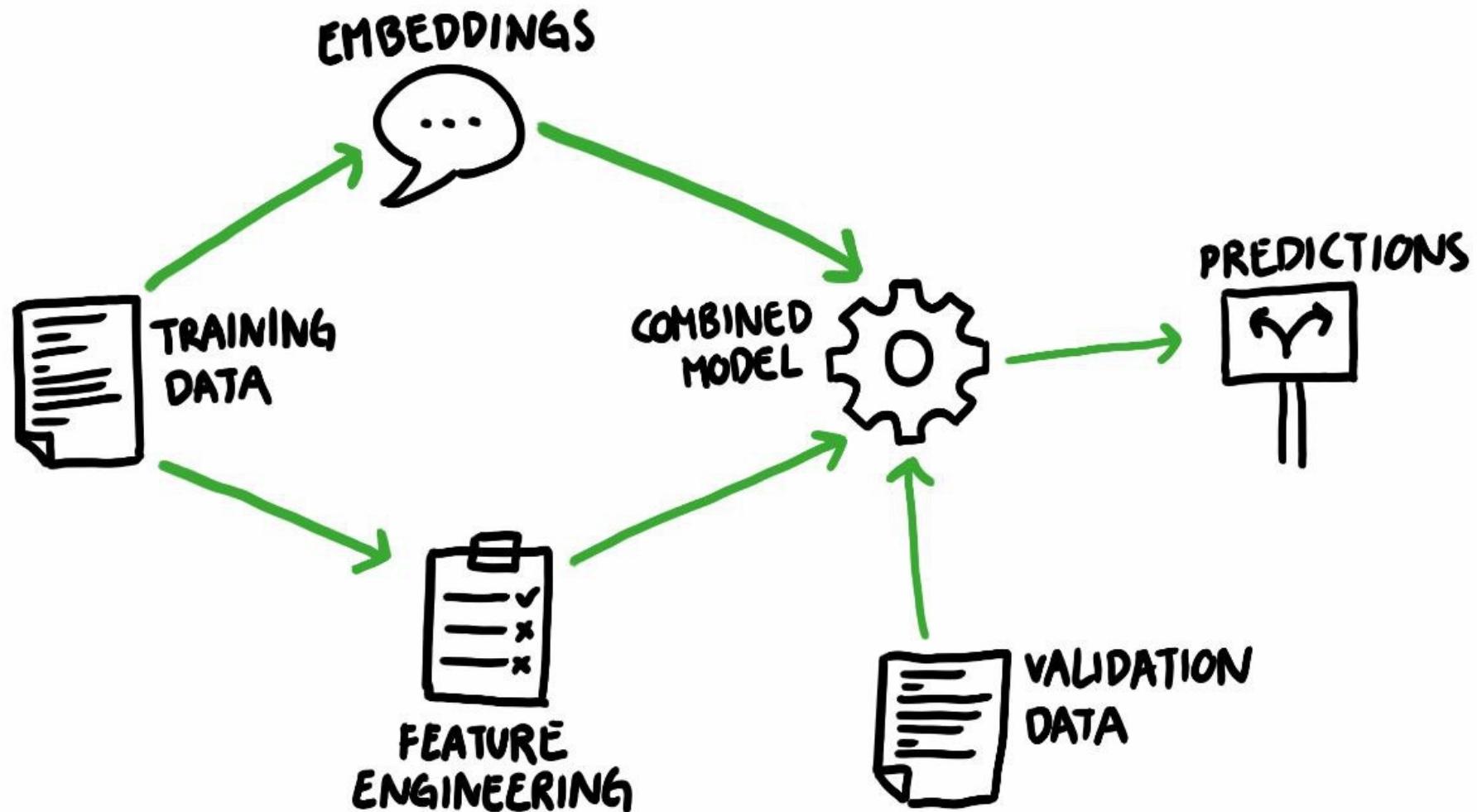
- See also <https://www.arxiv-vanity.com/papers/2106.07410/>

Difficulties for NLP Models

- Dimensions
- Variability in models and representations
- Set of tokens != variables
- Local → global
- Difficult to show graphically
- Joint effects and effects of absent words

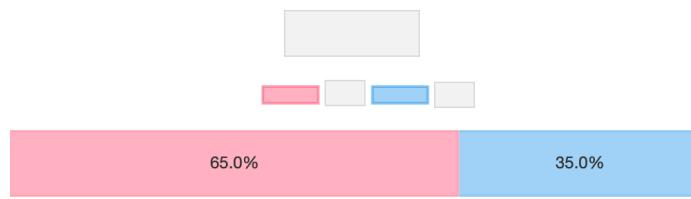


Real World Example- Setup



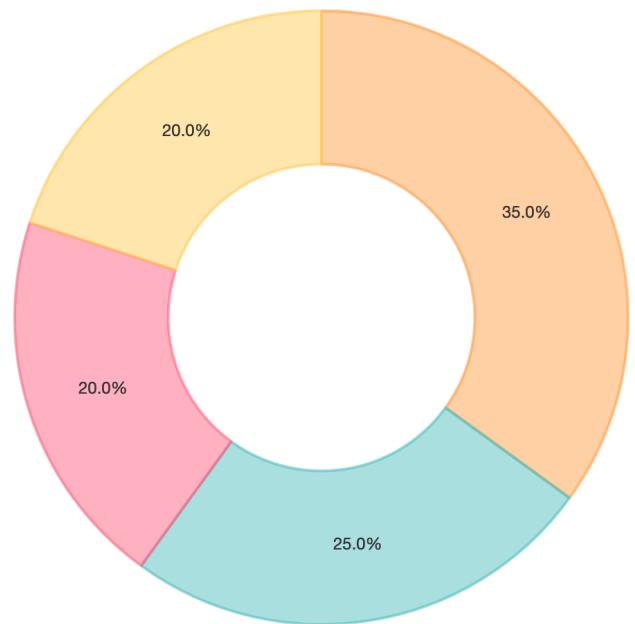
Ergebnisse:

Vorhersage und Verteilung:

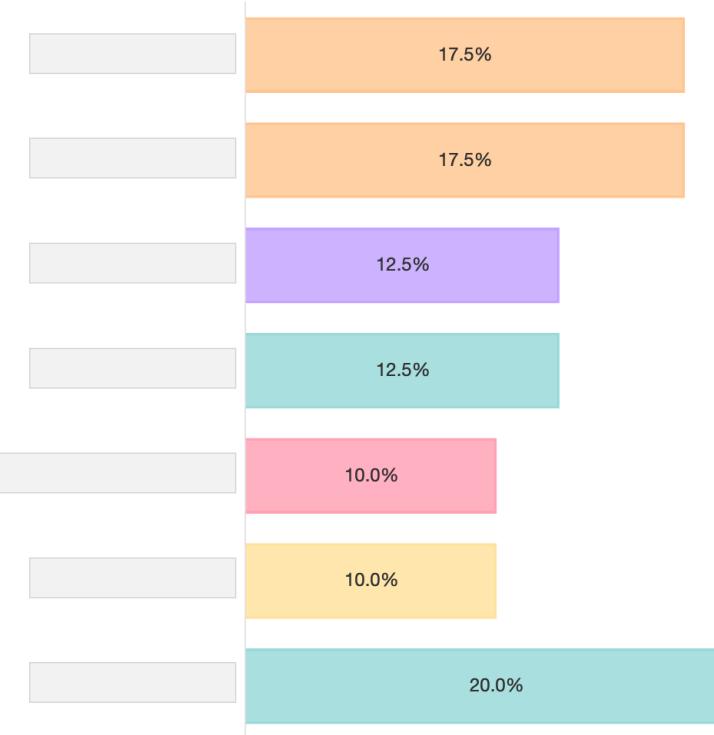


Text-Indikatoren (Zusammenfassung)

Text-Feature: [] Text-Feature: []
Text-Feature: [] Andere []



Feature-Indikatoren

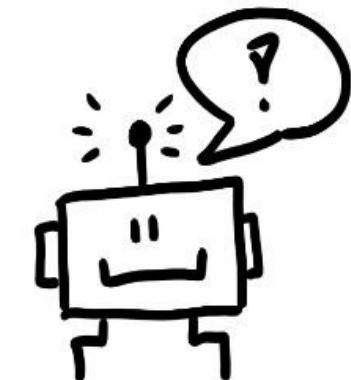


Kommentar



Lessons Learned

- Explain the explainability
- Absense of words
- Local impact VS global impact
- Find a fitting graphical representation
- Reduce complexity
- Include help texts



Links & Further Reading

- <https://volpato.io/articles/1907-nlp-xai.html>
- <https://www.arxiv-vanity.com/papers/2106.07410/>
- <https://analyticsindiamag.com/hands-on-guide-to-interpret-machine-learning-with-shap/>
- https://github.com/RajkumarGalaxy/StructuredData/blob/master/Interpret_ML_with_SHAP.ipynb
- <https://christophm.github.io/interpretable-ml-book/shap.html>
- <https://shap.readthedocs.io/en/latest/index.html>
- <http://lineardigressions.com/episodes/2019/8/18/data-shapley>
- <https://eli5.readthedocs.io/en/latest/overview.html>
- <https://mapmeld.medium.com/deciphering-explainable-ai-with-eli5-22c90a06a32a>
- <https://towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5>
- https://microsoft.github.io/nlp-recipes/examples/model_explainability/
- <https://visbert.demo.datexis.com/>
- 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. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
- Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

Thanks!

- GitHub
<https://github.com/LarissaHa>
- Twitter [@l_r_ss](https://twitter.com/l_r_ss)
- LinkedIn <https://www.linkedin.com/in/larissa-ha/>



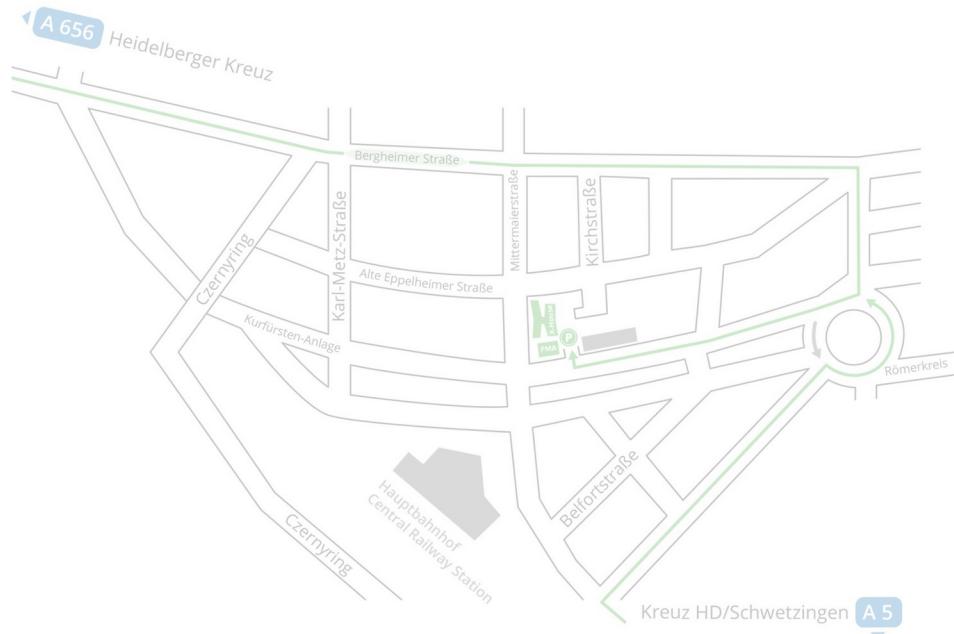
Kontakt



sovanta AG
X-House

Tel. +49 (0)6221 18733-0
Fax +49 (0)6221 18733-44
info@sovanta.com
www.sovanta.com

sovanta AG
X-House
Mittermaierstraße 31
69115 Heidelberg



Sitz der Gesellschaft: Heidelberg · Registergericht: Mannheim HRB 708906 · Ust-IdentNr.: DE269864971 · Vorstand: Prof. Dr. Claus E. Heinrich, Michael Kern · Vorsitzender des Aufsichtsrates: Dr. Georg Konrad · Haftungsausschluss Alle Informationen und Erklärungen dieser Website sind unverbindlich. Die sovanta AG übernimmt für die Richtigkeit und Vollständigkeit der Inhalte keine Gewähr. Es wird keine Garantie übernommen und keine Zusicherung von Produkteigenschaften gemacht. Aus den Inhalten der Internetseite ergeben sich keine Rechtsansprüche. Layout und Gestaltung der Website sowie die einzelnen Elemente sind urheberrechtlich geschützt. sovanta® ist ein eingetragenes Warenzeichen (Wortmarke). Alle anderen Produkte, die im Inhalt dieser Website erscheinen, sind registrierte oder nicht registrierte Warenzeichen der jeweiligen Firmen. Alle Rechte vorbehalten. Die sovanta AG übernimmt keine Garantie für die auf dieser Website vorhandenen Verweise („Links“) auf andere Websites. Für den Inhalt der verlinkten Seiten sind ausschließlich deren Betreiber verantwortlich. Die sovanta AG ist für den Inhalt einer Seite, die mit einem solchen Link erreicht wird, nicht verantwortlich. Alle Links sind lediglich als wertfreier Hinweis auf das bestehende, von Dritten erstellte Angebot anzusehen. Die sovanta AG übernimmt keine Gewähr für die Fehlerfreiheit von Daten und Software, die von der Website heruntergeladen werden können. Diese werden auf Virenbefall überprüft. Dennoch empfehlen wir Daten und Software nach dem Herunterladen auf Virenbefall mit der jeweils neuesten Virensuchsoftware zu prüfen. Im Falle von Schäden insbesondere für unmittelbare und mittelbare Folgeschäden, Datenverluste, entgangenen Gewinn, System- und Produktionsausfällen, die durch die Nutzung dieser Website oder das Herunterladen von Daten entstehen haftet die sovanta AG nicht.