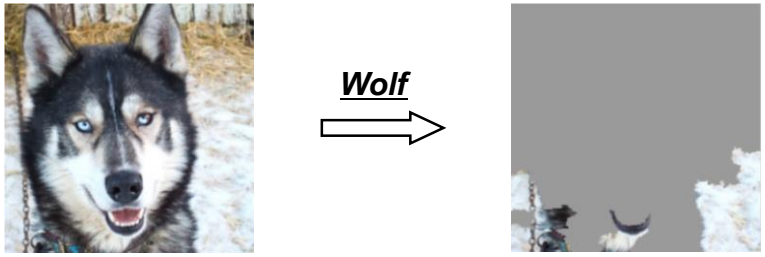
Interpretability of Neural Network Models Used in Data Analysis

Author: Branislav Pecher
Supervisor: Ing. Jakub Ševcech, PhD.

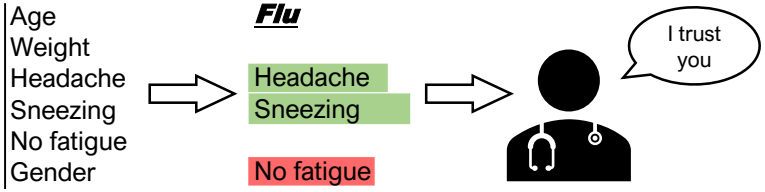
STU FIIT
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Why is interpretability important?

- Helps audit and find problems decision process of models



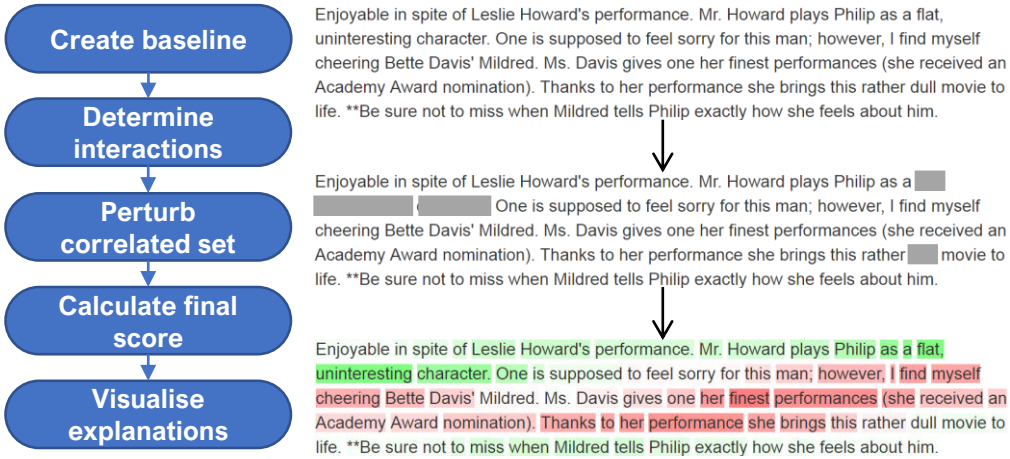
- Enhances trust in the model and therefore its use in practice



- Attribution methods most popular for neural networks
- Shortcomings of current methods:
  - Problems when encountering interactions in data
  - Mostly focused on problems involving image data
  - Explanations sometimes hard to interpret

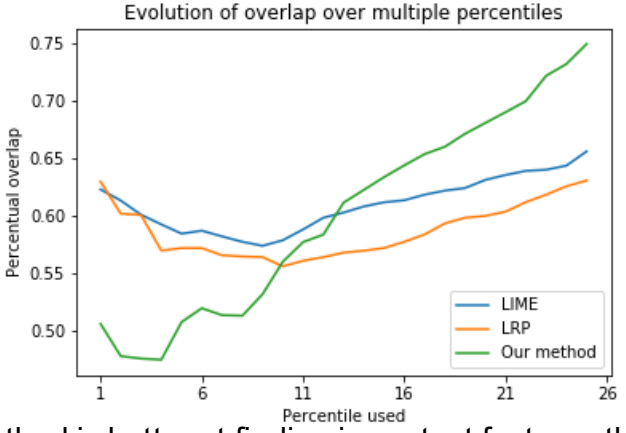
Proposed method

- Used for explanation of decisions in **text** classification with the use of word embeddings
- Take into consideration interactions when generating explanations – more precise explanations
- Generate insightful explanations – visualization of explanations
- 5 step process



Experiments and results

- Comparison with other attribution methods – LIME[1] and LRP[2]
- Solution: User experiment to determine ground truth



- Our method is better at finding important features that are otherwise overlooked due to the interactions

[1] Ribeiro, M.T, et al..: 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, ACM, 2016, pp. 1135–1144.
[2] Bach, S., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 2015, vol. 10, no. 7