Master Thesis

Explorative study of explainable artificial intelligence techniques for sentiment analysis applied for English language

by Rohan Kumar Rathore

Advisor: Dr. Anton Kolonin

Scientific workshop "Big Data Analytics"
Novosibirsk State University, Russia

March 16, 2021

Introduction

- Artificial intelligence : Artificial agents achieving goals smartly
- Machine learning : Algorithmic models responsible for smartness
- Explainable artificial intelligence : Techniques to explain the models

Outline

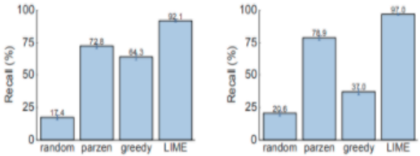
Explainable artificial intelligence (XAI) techniques for sentiment analysis

- Model development
  - Sentiment analysis model on IMDB movie reviews dataset
- Technique I
  - Local interpretable model-agnostic explanations (LIME): Explaining with surrogate models
- Technique II
  - Layer-wise relevance propagation (LRP): Explaining with propagated weights relevance scores of the network
- Technique III
  - Artificial neural network decision tree algorithm (Ruleex ANN-DT): Explaining by extraction of decision trees from artificial neural networks
- Performance Analysis
  - Simulatability test: A model is simulatable when a person can predict its behavior on new inputs

Exploration of LIME for sentiment analysis - I

Local interpretable model-agnostic explanations

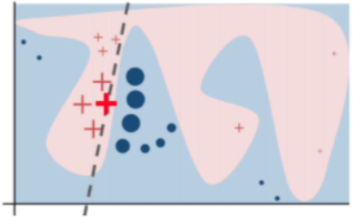
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



(a) Sparse LR

(b) Decision Tree

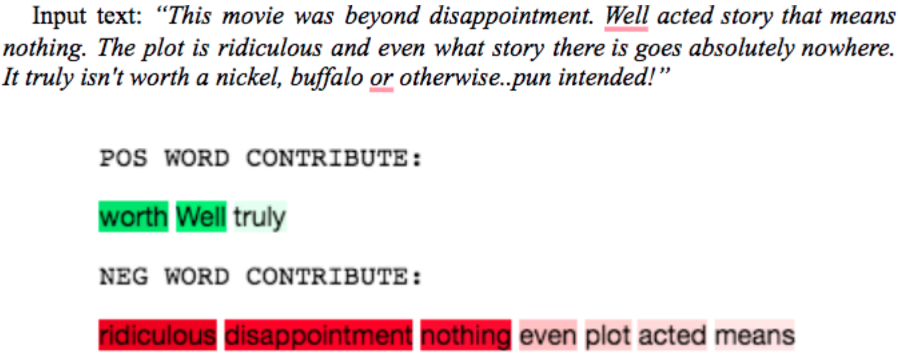
Recall on truly important features



Toy example to present intuition for LIME.

Exploration of LIME for sentiment analysis - II

Local interpretable model-agnostic explanations



Input text: *"This movie was beyond disappointment. Well acted story that means nothing. The plot is ridiculous and even what story there is goes absolutely nowhere. It truly isn't worth a nickel, buffalo or otherwise..pun intended!"*

POS WORD CONTRIBUTE:

worth Well truly

NEG WORD CONTRIBUTE:

ridiculous disappointment nothing even plot acted means

Exploration of LRP for sentiment analysis - I

Layer-wise relevance propagation

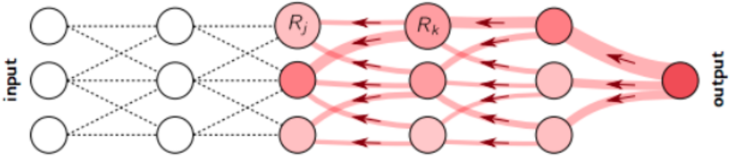


Fig. 10.2. Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

Basic Rule

$$R_j = \sum_k \frac{a_j \cdot \rho(w_{jk})}{\epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})} R_k,$$

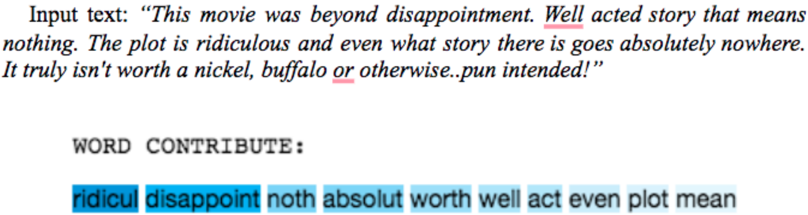
Epsilon Rule

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^{+})}{\sum_{0,j} a_j \cdot (w_{jk} + \gamma w_{jk}^{+})} R_k$$

Gamma Rule

Exploration of LRP for sentiment analysis - II

Layer-wise relevance propagation

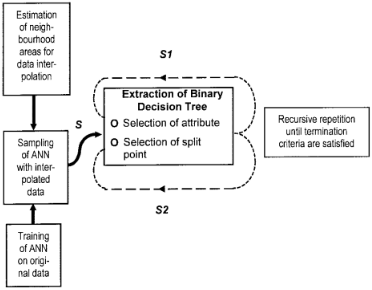


Input text: *"This movie was beyond disappointment. Well acted story that means nothing. The plot is ridiculous and even what story there is goes absolutely nowhere. It truly isn't worth a nickel, buffalo or otherwise..pun intended!"*

WORD CONTRIBUTE:

ridicul disappoint noth absolut worth well act even plot mean

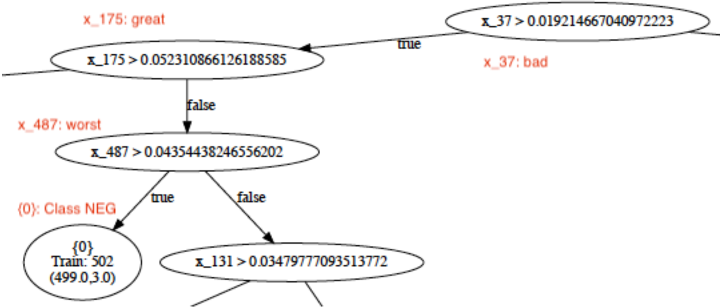
Exploration of Ruleex ANN-DT for sentiment analysis - I


Estimation of neighbourhood areas for data interpolation
Sampling of ANN with interpolated data
Training of ANN on original data
S
S1
S2
Extraction of Binary Decision Tree
O Selection of attribute
O Selection of split point
Recursive repetition until termination criteria are satisfied

- Selection of Attribute: Similar to CART algorithm of reducing the entropy $$V_w = \sum_{k=1}^{2} \frac{n_k}{n} \mathrm{Var}(O_k)$$
- Stopping criteria: Standard deviation or the variance is zero
- Statistical pruning technique: chi squared

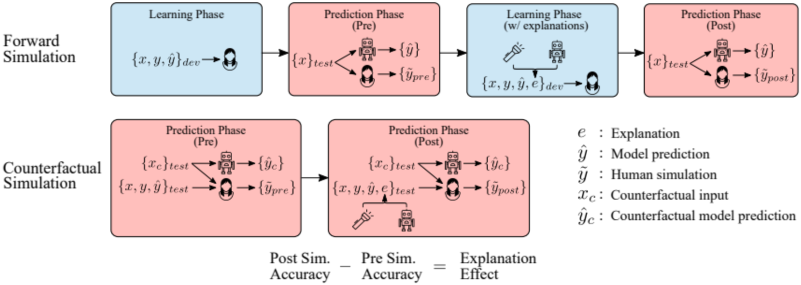
Exploration of Ruleex ANN-DT for sentiment analysis - II

- Decision Tree representation (showing sub-section here)


x_175: great
x_37 > 0.019214667040972223
true
x_175 > 0.052310866126188585
x_37: bad
false
x_487: worst
x_487 > 0.04354438246556202
true
false
{0}: Class NEG
{0}
Train: 502
(499.0,3.0)
x_131 > 0.03479777093513772

Performance Analysis Method

Simulatability test: Model is simulatable if person can predict its behavior on new inputs



Forward Simulation

Learning Phase

$\{x, y, \hat{y}\}_{dev}$

Prediction Phase (Pre)

$\{x\}_{test}$ $\{\hat{y}\}$ $\{\tilde{y}_{pre}\}$

Learning Phase (w/ explanations)

$\{x, y, \hat{y}, e\}_{dev}$

Prediction Phase (Post)

$\{x\}_{test}$ $\{\hat{y}\}$ $\{\tilde{y}_{post}\}$

Counterfactual Simulation

Prediction Phase (Pre)

$\{x_c\}_{test}$ $\{\hat{y}_c\}$

$\{x, y, \hat{y}\}_{test}$ $\{\tilde{y}_{pre}\}$

Prediction Phase (Post)

$\{x_c\}_{test}$ $\{\hat{y}_c\}$

$\{x, y, \hat{y}, e\}_{test}$ $\{\tilde{y}_{post}\}$

e : Explanation

$\hat{y}$: Model prediction

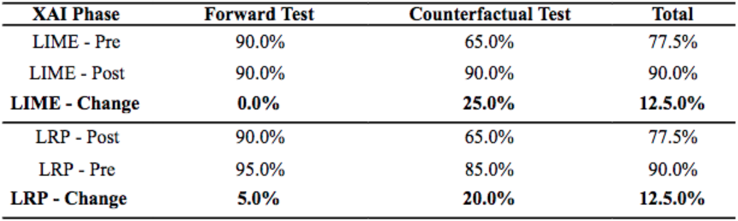
$\tilde{y}$: Human simulation

x_c : Counterfactual input

$\hat{y}_c$: Counterfactual model prediction

Post Sim. Accuracy − Pre Sim. Accuracy = Explanation Effect

Results



XAI Phase	Forward Test	Counterfactual Test	Total
LIME - Pre	90.0%	65.0%	77.5%
LIME - Post	90.0%	90.0%	90.0%
LIME - Change	**0.0%**	**25.0%**	**12.5.0%**
LRP - Post	90.0%	65.0%	77.5%
LRP - Pre	95.0%	85.0%	90.0%
LRP - Change	**5.0%**	**20.0%**	**12.5.0%**

- A total of 120 data points were collected
- Improvement the accuracy of model prediction capability of the human subject by 12.5%

Paper Readiness

- International Conference on Data Science and Applications, ICDSA 2021 (Accepted for presentation in conference)
- Improve results (Performance analysis on ANN-DT technique) and apply in other conferences

References

- Gregor P.J.S., Chris A., Francois S.G., ANN-DT: An Algorithm forExtraction of Decision Trees from Artificial Neural Networks, IEEETransactions on Neural Networks, Vol.10 No.6, 1999
- Montavon G., Binder A., Lapuschkin S., Samek W., Müller K.R.: Layer-Wise Relevance Propagation: An Overview. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller K.R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS, vol. 11700, pp. 193–209. Springer, Cham (2019)
- Ribeiro, M.T., Singh, S., Guestrin, C.: "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 (KDD '16), pp.1135–1144, Association for Computing Machinery, New York (2014)
- GHase, P., Bansal, M.: Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. arXiv:2005.01831 (2020)