A comparative study of instance-level explanations for big, sparse data

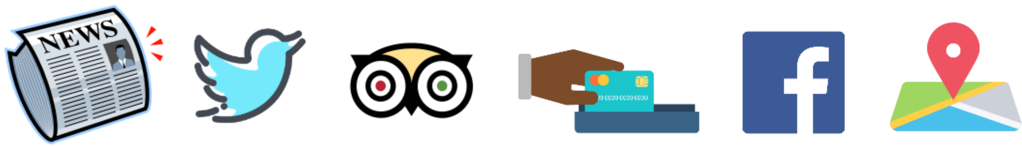
Yanou Ramon, David Martens
Applied Data Mining

Universiteit Antwerpen

19th of March, 2019
ECDA (University of Bayreuth, Germany)

1. Introduction

- Applications using **high-dimensional**, **sparse** data are ample



Behavioral data
payment data, visited websited or physical locations, FB likes…

Textual data
emails, news articles, Twitter posts…

1. Introduction

High-dimensional & sparse → Gender prediction using movie viewing data

ACTIVE FEATURE = "EVIDENCE"

	Star wars	Pearl Harbor	Django	…	Home Alone	*Target $\hat{y}$ Gender*
User 1	$\boxed{1}$	0	0		1	***M***
User 2	1	1	0		1	***F***
…						
User n	1	1	1		0	***M***

6,040 users

1. Introduction

- High predictive **performance** ⇔ **complex** models
- **Interpretability issues:** how are predictions made?

1. Introduction

- High predictive **performance** ⇔ **complex** models
- **Interpretability issues:** how are predictions made?

⬇

- Ethical objectives: privacy, fairness, safety
- Model improvement: debugging, data problems
- Trust/acceptance
- …

1. Introduction

- High predictive **performance** ⇔ **complex** models
- **Interpretability issues:** how are predictions made?

- Ethical objectives: privacy, fairness, safety
- Model improvement: debugging, data problems
- Trust/acceptance
- …

Instance-level explanations

1. Introduction

"Which **instance-level explanation method** is **most suitable** for explaining model predictions on **high-dimensional, sparse** data?**"**

- **Overview** of selected instance-level explanation methods
- Selection of **quantitative criteria**
- **Comparison** using **behavioral/textual** data

2. Explanation methods

Selection criteria

- **Model-agnostic** method: treats model as a black box
- **Computational ability** to cope with **high-dimensional** data

2. Explanation methods

Selection criteria

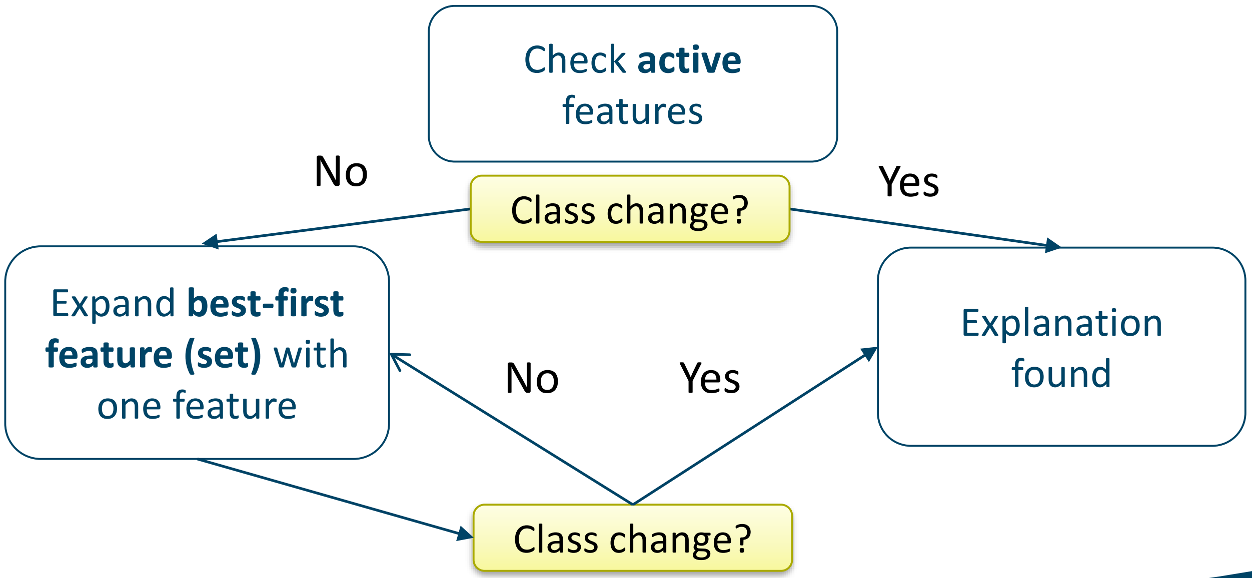
- **Model-agnostic** method: treats model as a black box
- **Computational ability** to cope with **high-dimensional** data

- Evidence Counterfactual (EDC) (Martens & Provost, 2013)
- Linear Interpretable Model-Agnostic Explainer (LIME) (Ribeiro et al., 2016)
- Shapley Additive Values (SHAP) (Lundberg & Lee, 2017)

2. Explanation methods

Evidence counterfactual

- **Minimal set of features** so that **removing** them results in a predicted class change
- **"Removing"** → set feature value to zero / remove evidence
- **Model-agnostic** algorithm based on **heuristic best-first** search



2. Explanation methods

Evidence counterfactual – example

Example: gender prediction using movie viewing data

User x_i: Sam

Sam watched 120 movies
Sam is predicted as male

2. Explanation methods

Evidence counterfactual – example

Example: gender prediction using movie viewing data

User x_i: Sam

WHY?

Sam watched 120 movies
Sam is predicted as male

2. Explanation methods

Evidence counterfactual – example

Example: gender prediction using movie viewing data

User x_i: Sam

IF Sam would not have watched ***{Taxi driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me, Interstellar}***, **THEN** his predicted class would change from male to female

2. Explanation methods

Evidence counterfactual – example

Example: gender prediction using movie viewing data

User x_i: Sam

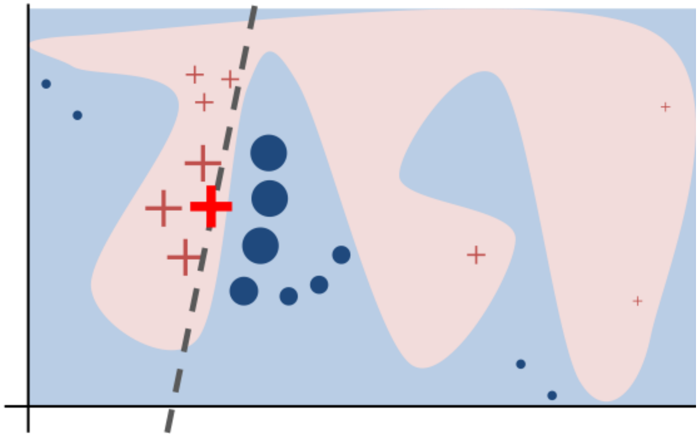
IF Sam would not have watched ***{Taxi driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me, Interstellar}***, **THEN** his predicted class would change from male to female

POSITIVE EVIDENCE = EVIDENCE *FOR* A PREDICTED CLASS

2. Explanation methods

LIME / SHAP

- Explanation model: **sparse, linear model**
- Explanation model **approximates** original model in the **neighborhood of the instance**
- **Perturbed instances**



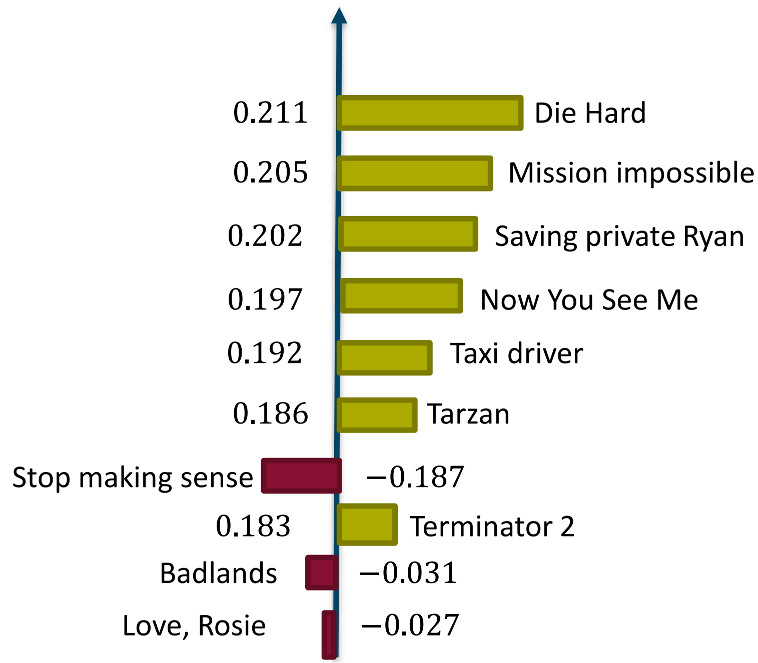
Source: Ribeiro et al., 2016

2. Explanation methods

LIME – example

Example: gender prediction using movie viewing data

User x_i: Sam
***k* = 10 features**
(feature selection)



2. Explanation methods

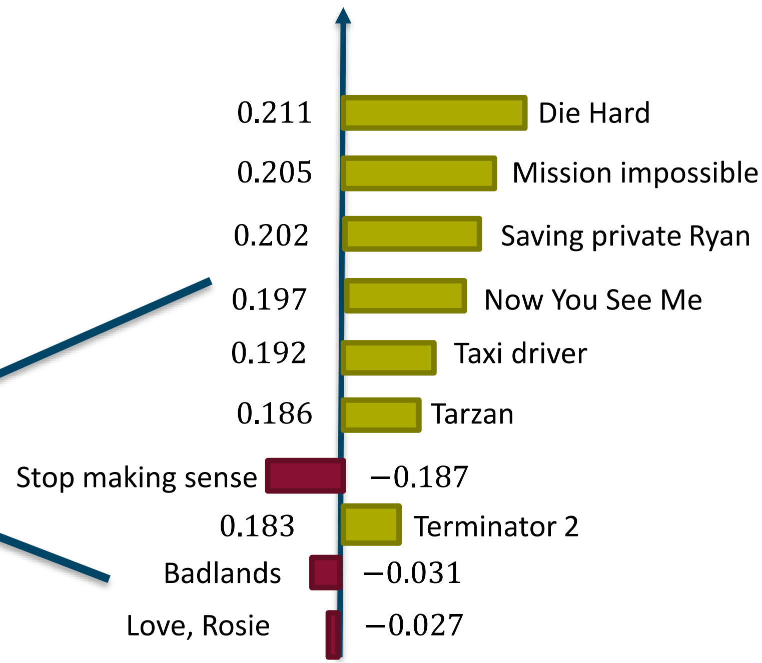
LIME – example

Example: gender prediction using movie viewing data

User x_i: Sam

***k* = 10 features**

(feature selection)



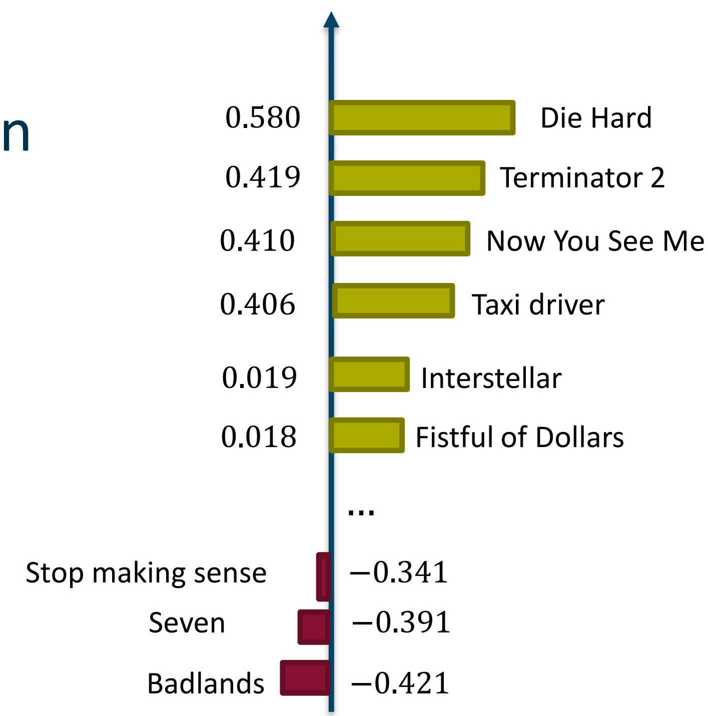
BOTH POSITIVE & NEGATIVE EVIDENCE

2. Explanation methods

SHAP – example

Example: gender prediction using movie viewing data

User x_i: Sam
Lasso regularization



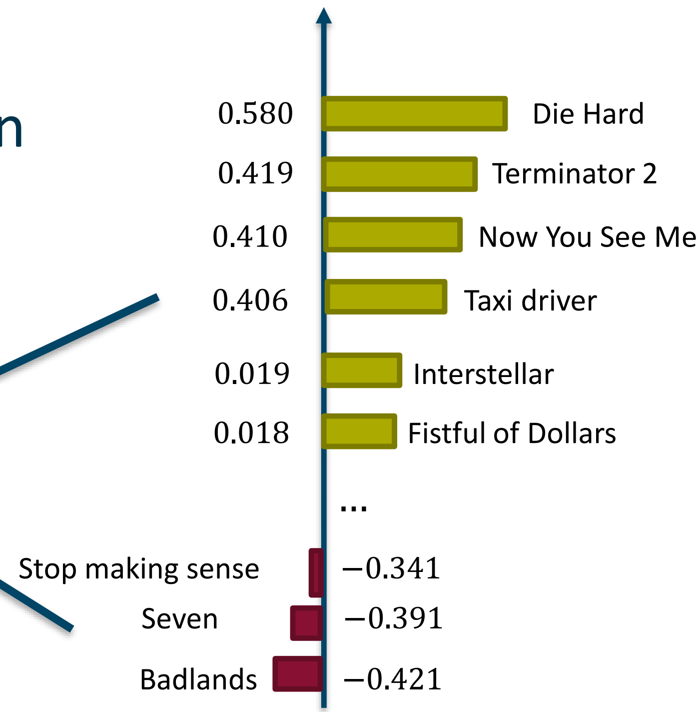
2. Explanation methods

SHAP – example

Example: gender prediction using movie viewing data

User x_i: Sam

Lasso regularization



BOTH POSITIVE & NEGATIVE EVIDENCE

3. Evaluation criteria

⇒ <u>**NOT**</u> a **qualitative** evaluation

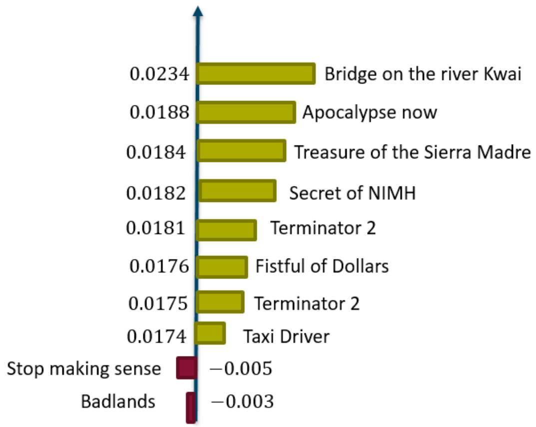
⇒ No evaluation of counterfactual vs linear model, negative evidence, output size, coefficients etc.

Counterfactual

IF Sam would not have rated ***{Taxi driver, North by Northwest, Bridge on the river Kwai, Terminator 2, Hunt for red October, Glengarry Glen Ross}***, **THEN** his predicted class would change from male to female

vs

Additive feature attribution

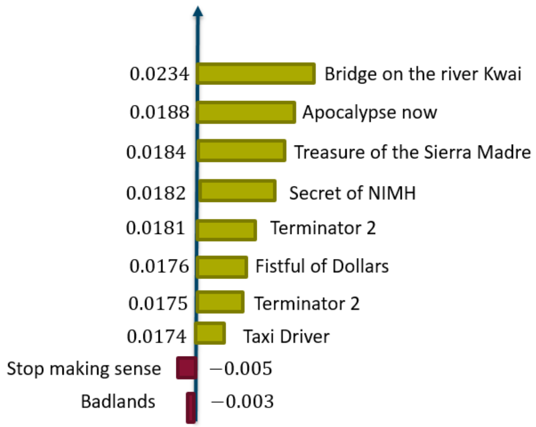


3. Evaluation criteria

⇒ **NOT** a **qualitative** evaluation
⇒ No evaluation of counterfactual vs linear model, negative evidence, output size, coefficients etc.

Counterfactual vs **Additive feature attribution**

IF Sam would not have rated ***{Taxi driver, North by Northwest, Bridge on the river Kwai, Terminator 2, Hunt for red October, Glengarry Glen Ross}***, **THEN** his predicted class would change from male to female



⇒ **Quantitative evaluation**

3. Evaluation criteria

For **a set of model predictions,** we want:

3. Evaluation criteria

For **a set of model predictions,** we want:

(1) to **generate** an explanation **output**

➔ Percentage of output generated

3. Evaluation criteria

For **a set of model predictions,** we want:

(1) to **generate** an explanation **output**

➔ Percentage of output generated

(2) that is **sparse** to be interpretable by humans

➔ Average output size

3. Evaluation criteria

For **a set of model predictions,** we want:

(1) to **generate** an explanation **output**

➔ Percentage of output generated

(2) that is **sparse** to be interpretable by humans

➔ Average output size

(3) that is **efficient** to compute

➔ Average computation time

3. Evaluation criteria

For **a set of model predictions,** we want:

(1) to **generate** an explanation **output**

➔ Percentage of output generated

(2) that is **sparse** to be interpretable by humans

➔ Average output size

(3) that is **efficient** to compute

➔ Average computation time

(4) that is able to **rank positive evidence from high to low relative importance**

➔ Average size of switching point
= number of features that need to be removed to change predicted class (**only** positive evidence)

4. Experimental setup

Collect data sets and build models

Generate explanations for test instances

Evaluation

Results & discussion

Text data: linear and rbf SVM

20NEWS

Behavioral data: LR and MLP

MOVIELENS

EDC

LIME

SHAP

% output generated

Avg output size

Avg computation time

Avg switching point

4. Experimental setup

Collect data sets and build models

- Text data: linear and rbf SVM — **20NEWS**
- Behavioral data: LR and MLP — **MOVIELENS**

Generate explanations for test instances

- EDC **≤ 10**
- LIME **= 10**
- SHAP

<u>Positively</u> predicted test instances

Small output sizes

Time limit: ≤10min

Evaluation

- % output generated
- Avg output size
- Avg computation time
- Avg switching point

Results & discussion

4. Experimental setup

Collect data sets and build models

Text data: linear and rbf SVM

20NEWS

Behavioral data: LR and MLP

MOVIELENS

Generate explanations for test instances

EDC **≤ 10**

LIME **= 10**

SHAP

<u>Positively</u> predicted test instances

Small output sizes

Time limit: ≤10min

Evaluation

% output generated

Avg output size

Avg computation time

Avg switching point

Results & discussion

Subset of instances for which output is generated

Subset of instances with SP
➔ Measured on <u>unrestricted</u> output sizes

5. Results

Table 1: Percentage generated for linear models (left) and nonlinear models (right)

Data set	Method	Percentage output generated
Movielens	**EDC ≤ 10**	75.5%
n = 302 ṁ = 327 Model: LR	**LIME=10**	100%
	SHAP	100%
20news	**EDC ≤ 10**	92.1%
n = 151 ṁ = 69 Model: lin-SVM	**LIME=10**	100%
	SHAP	100%

Data set	Method	Percentage output generated
Movielens	**EDC ≤ 10**	50.99%
n = 302 ṁ = 315 Model: MLP	**LIME=10**	100%
	SHAP	100%
20news	**EDC ≤ 10**	93.38%
n = 151 ṁ = 66 Model: rbf-SVM	**LIME=10**	100%
	SHAP	100%

5. Results

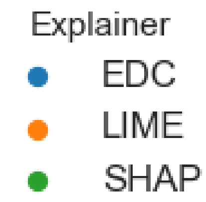


Figure 1 (a): Average absolute output size

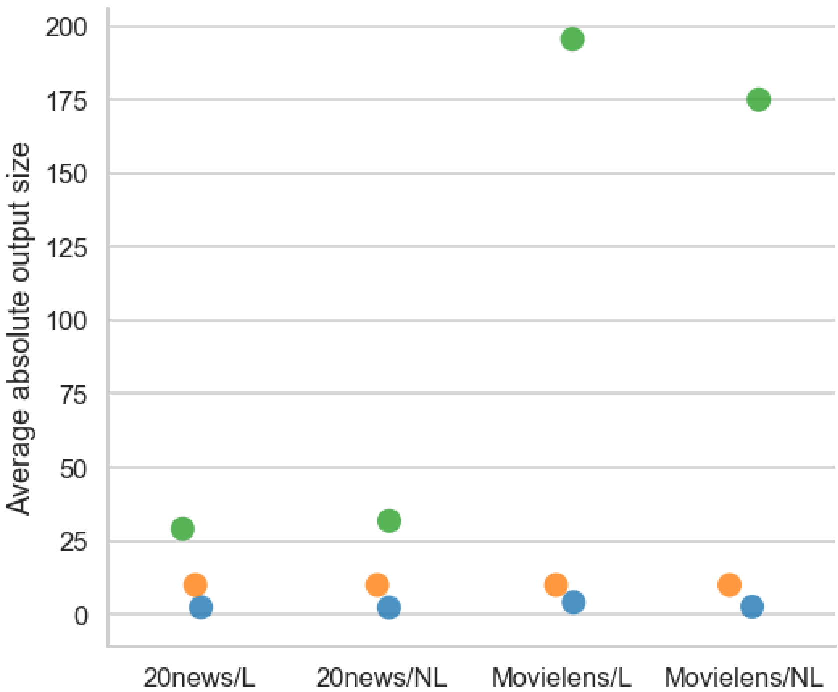
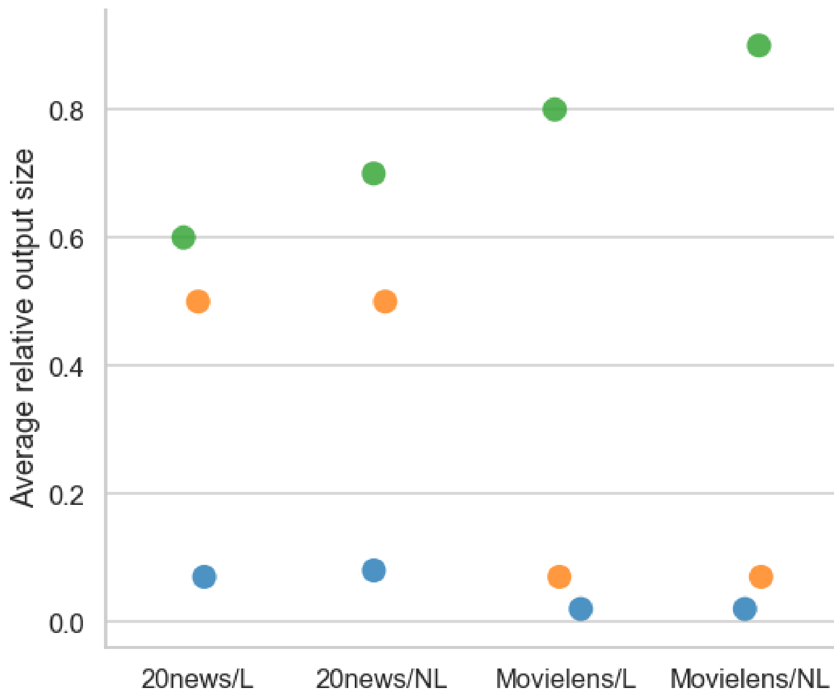
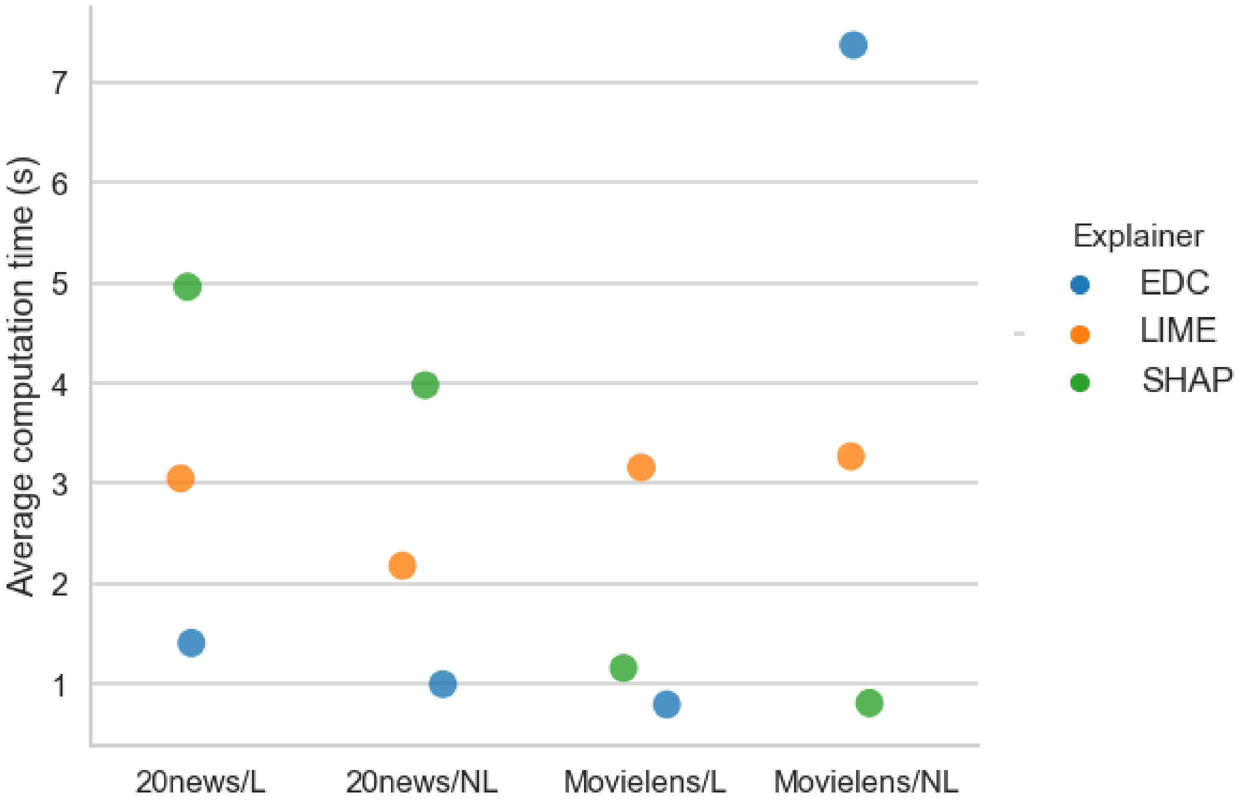


Figure 1 (b): Average relative output size



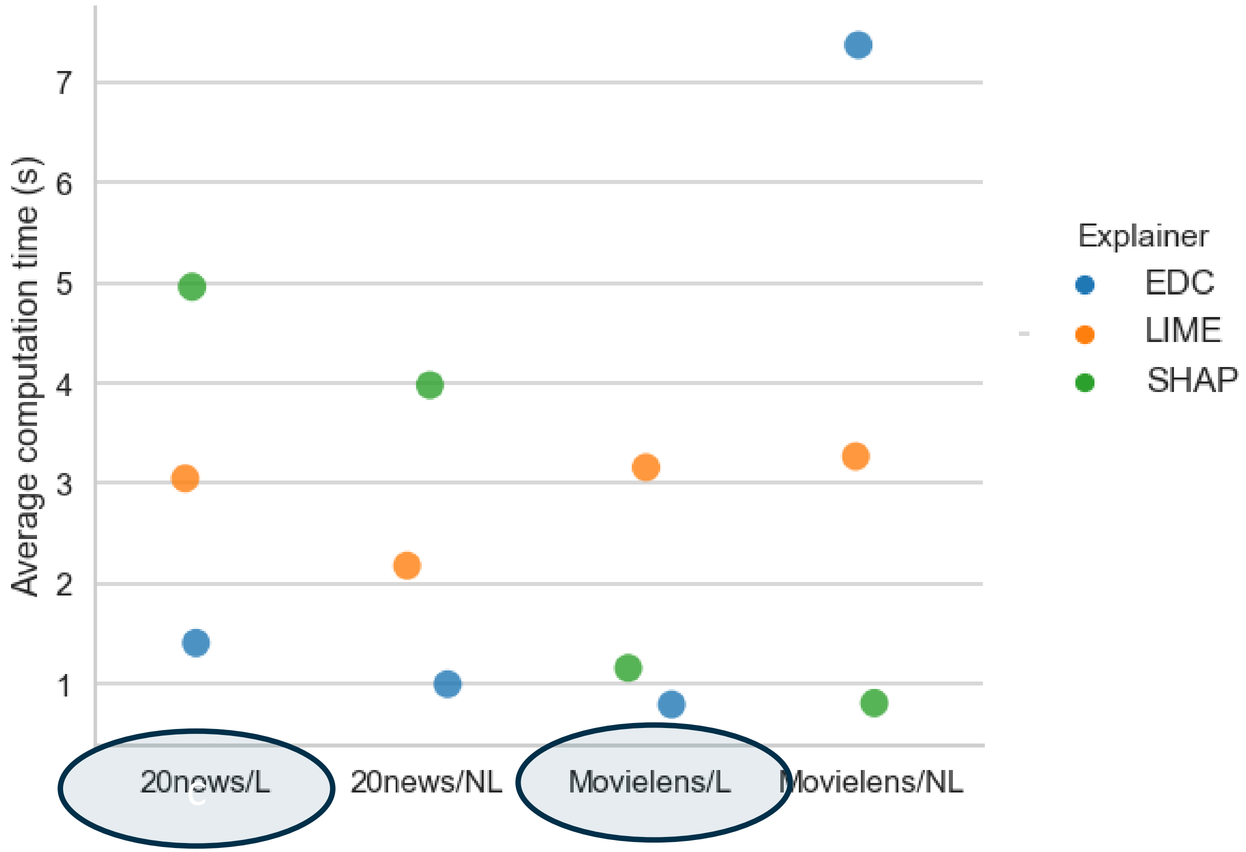
5. Results

Figure 2: Average computation time



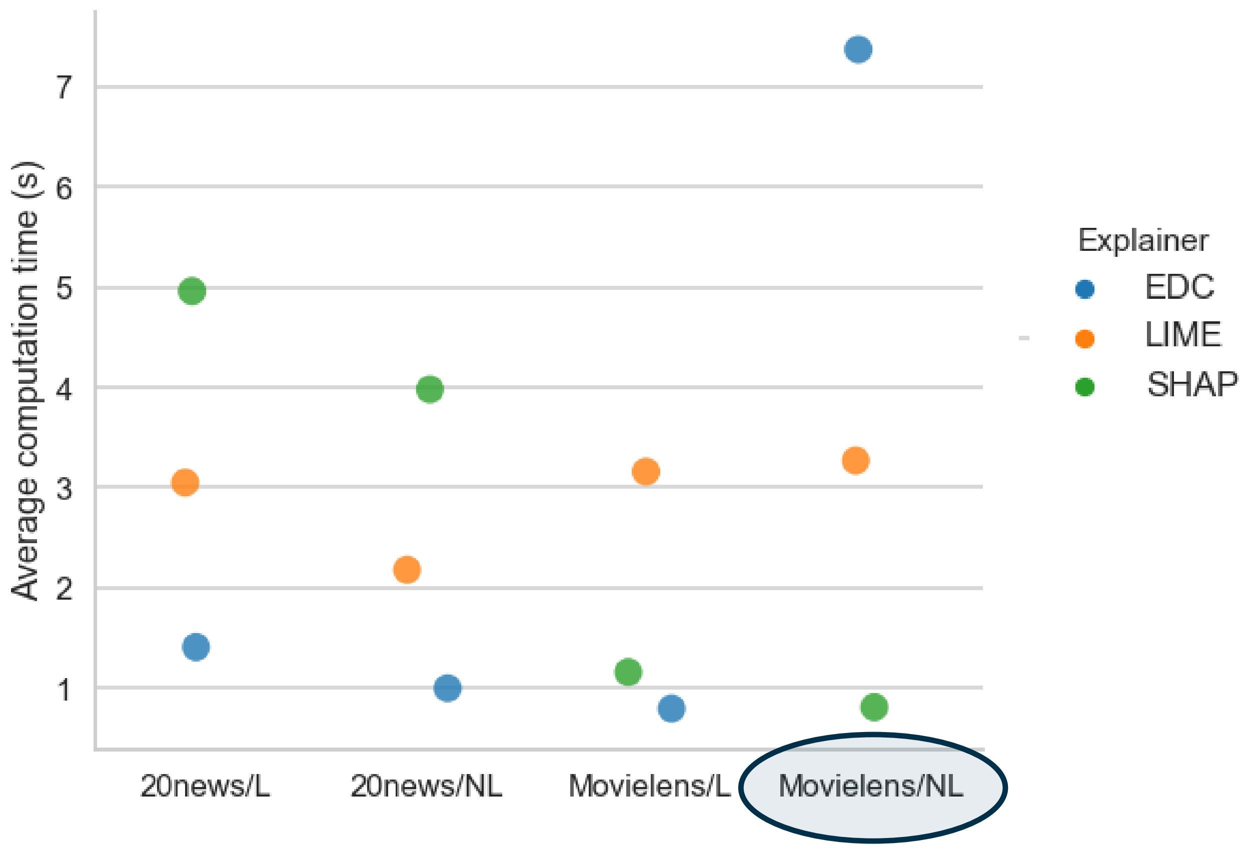
5. Results

Figure 2: Average computation time



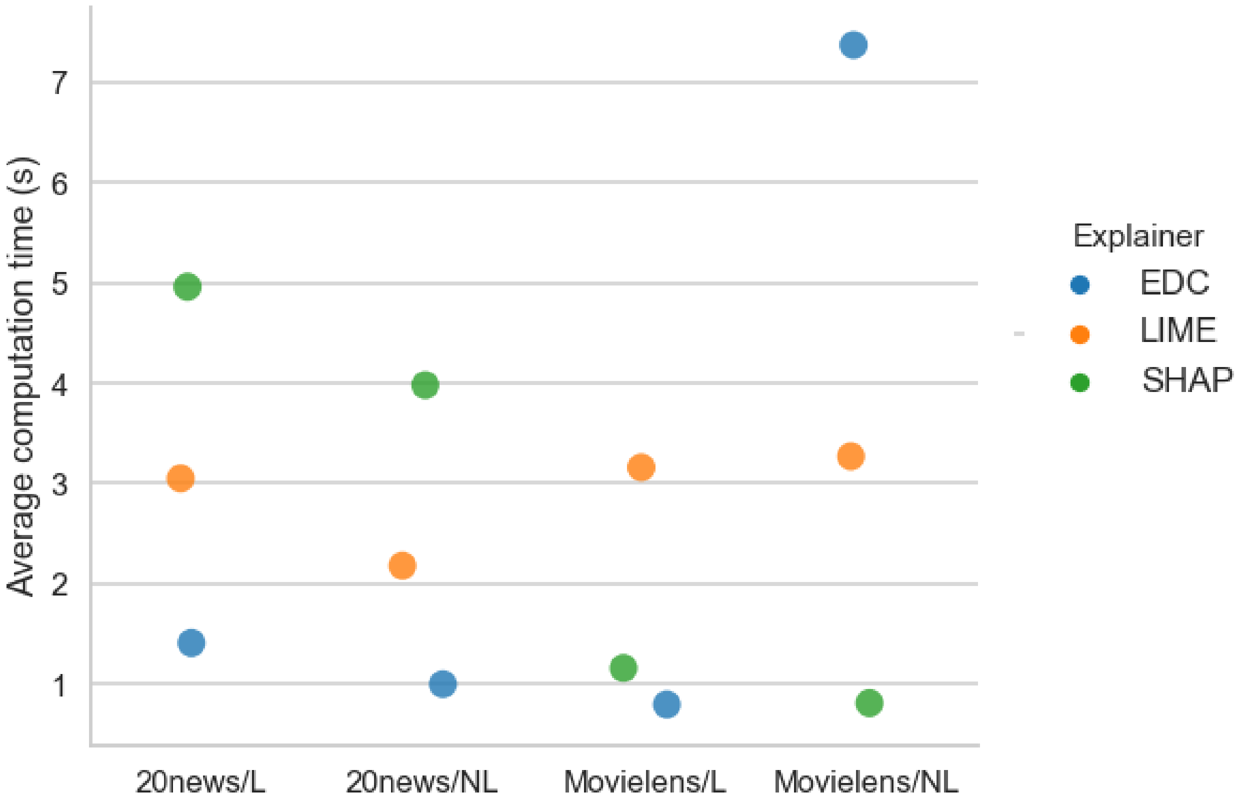
5. Results

Figure 2: Average computation time



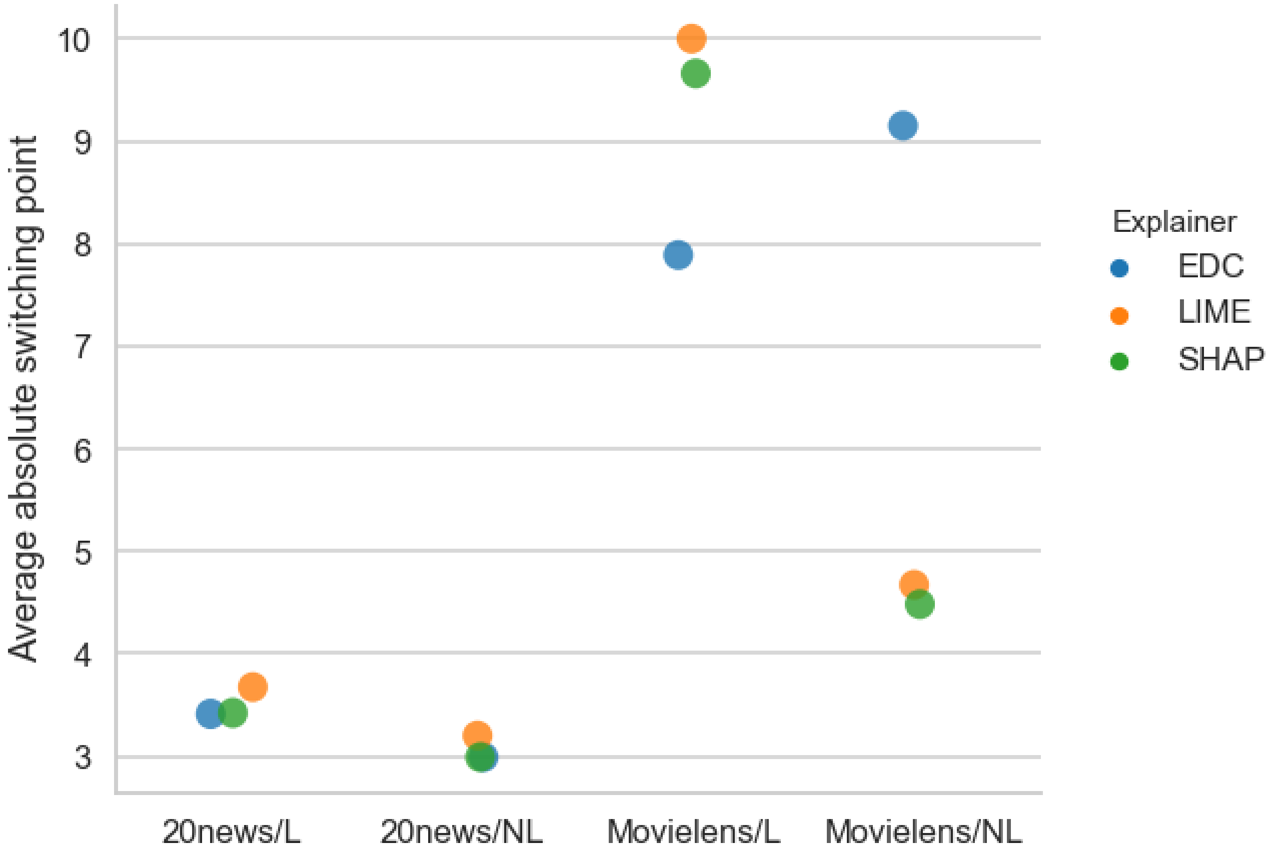
5. Results

Figure 2: Average computation time



5. Results

Figure 3: Average absolute switching point



5. Results

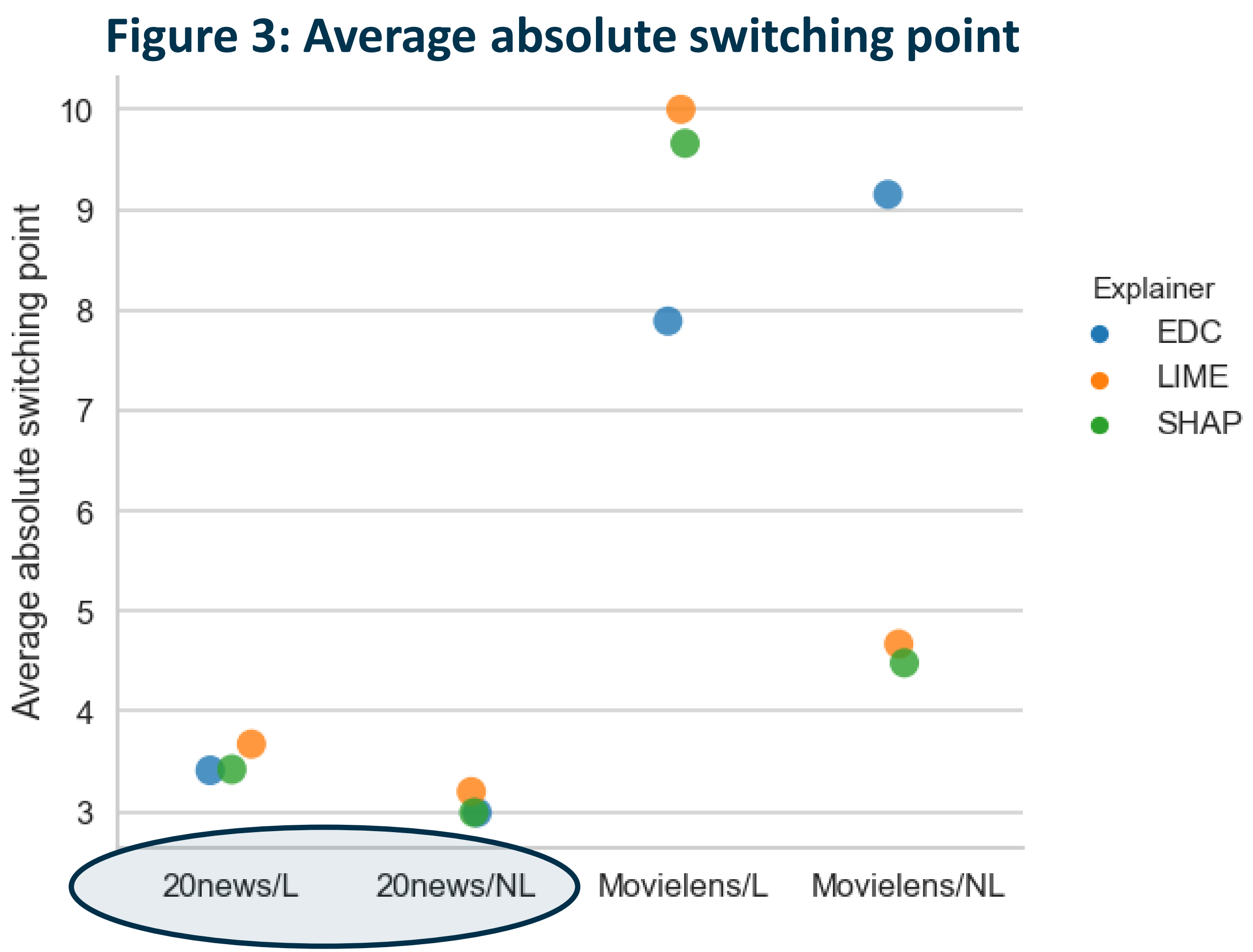
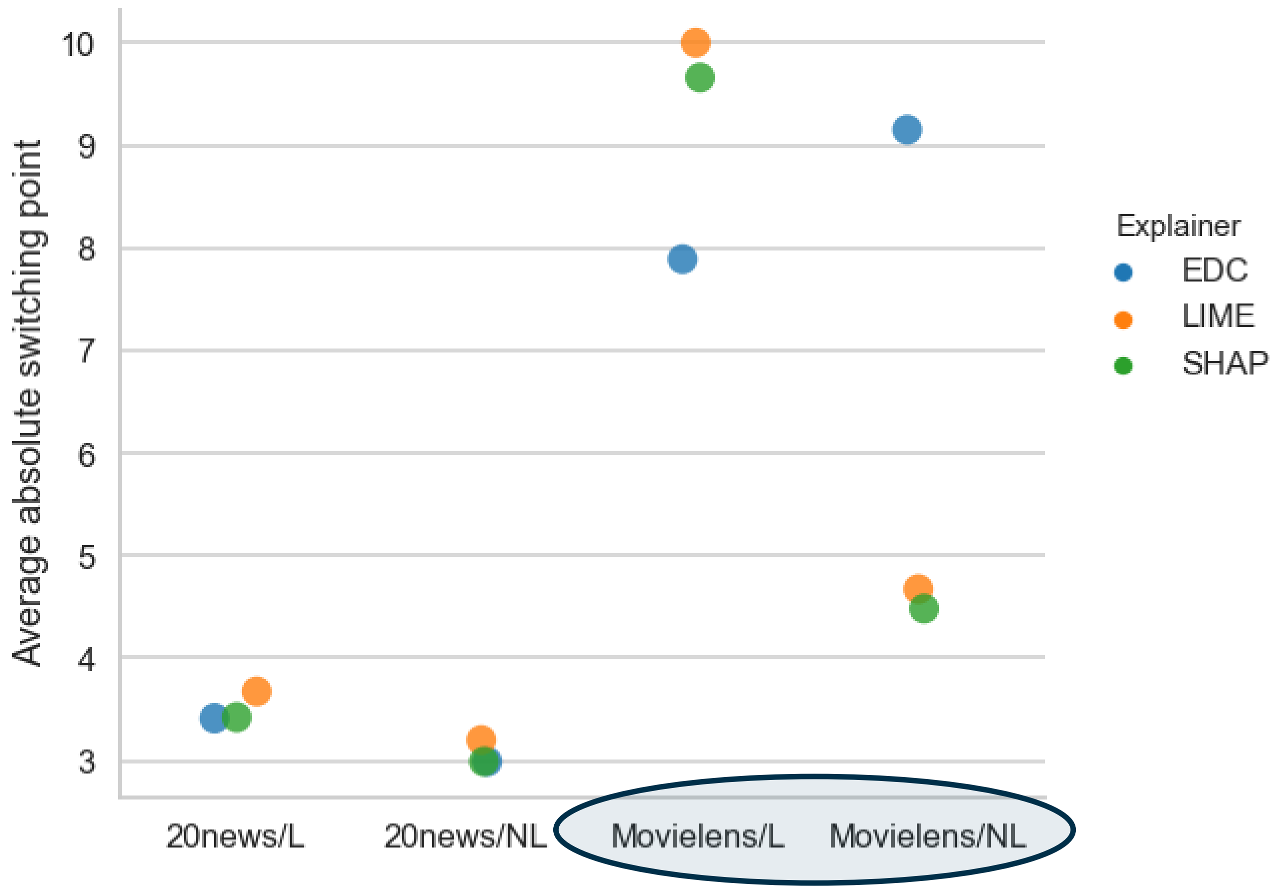


Figure 3: Average absolute switching point
Average absolute switching point
10
9
8
7
6
5
4
3
20news/L
20news/NL
Movielens/L
Movielens/NL
Explainer
EDC
LIME
SHAP

5. Results

Figure 3: Average absolute switching point

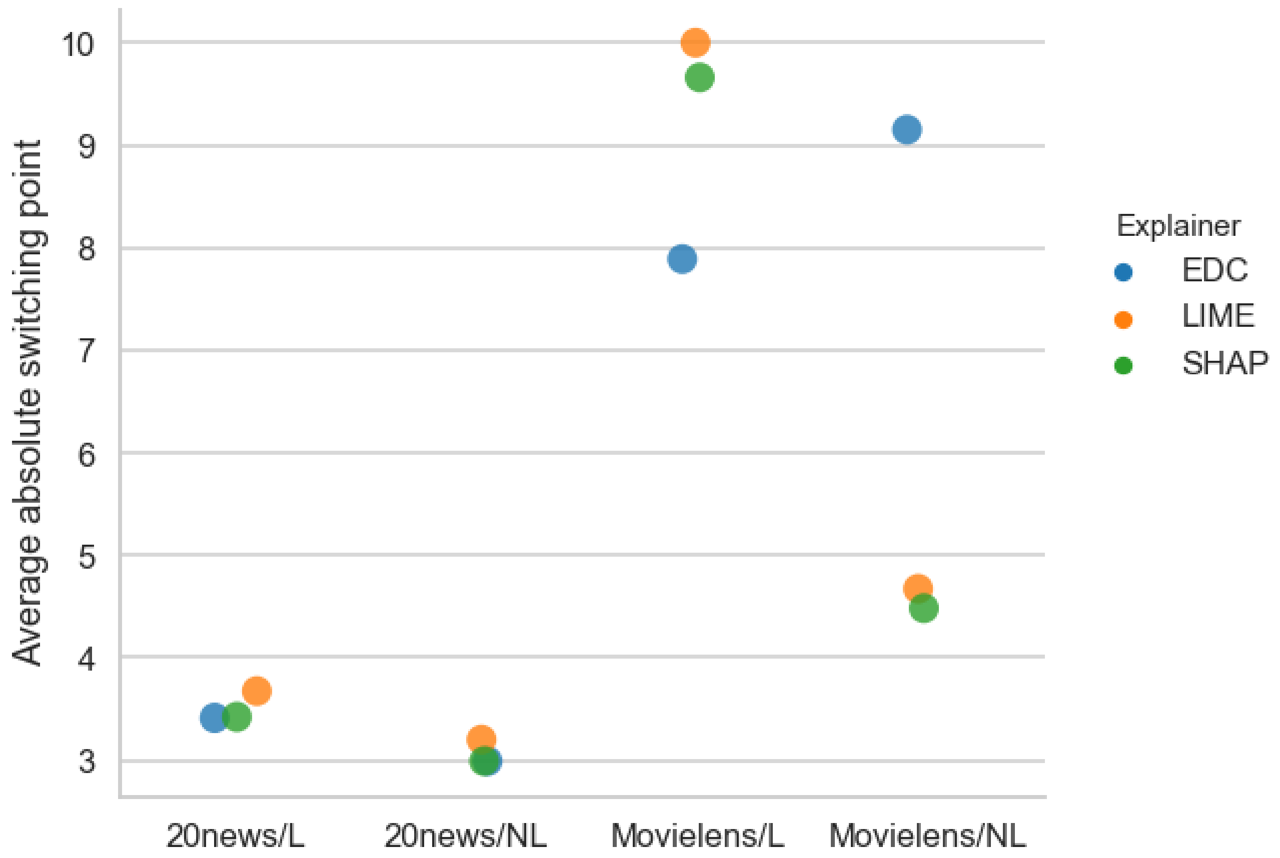


5. Results

Figure 3: Average absolute switching point

5. Results

Figure 3: Average absolute switching point



6. Discussion

Percentage output generated

- When restricting the output size (≤ 10), EDC does *not always* generate output

Explanation output size

- EDC provides *smallest* output sizes
- LIME can be *further reduced* if wanted
- SHAP cannot be *explicitly* restricted, ≥ 50% of active features included

Computational efficiency

- For *small* outputs and *linear* models, EDC is most efficient
- LIME and SHAP *relatively fast* for all scenarios

Ability to rank positive evidence ➔ switching point

- EDC provides smallest switching points for *linear* models
- Greedy approach EDC: worse results than LIME/SHAP for *some* non-linear models

7. Conclusion

A comparative study of instance-level explanations for big, sparse data

$\Rightarrow$ A **nuanced** conclusion:

- **EDC** seems best for smaller output sizes and linear models
- **SHAP**
  - Consistently relatively fast
  - Switching points close to the best
  - Very large outputs
- **LIME**: good trade-off
  - Consistently relatively fast ➔ most stable
  - Switching points close to the best
  - Ability to provide k

8. Further research

1. Adjustments of methods

- Adjust or combine methods → optimal approach

2. Extension of quantitative evaluation

- More data
- More models

3. Qualitative evaluation of explanation methods

- Relevance of negative evidence
- Counterfactual versus sparse, linear model

Thanks for your attention.
Questions?

https://www.linkedin.com/in/yanou-ramon

http://applieddatamining.com/cms/

yanou.ramon@uantwerp.be

Universiteit Antwerpen

References

Flach, P., & Sokol, K. (2018). *Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety.*
Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences.*
Martens, D. & Provost, F. (2013). *Explaining data-driven document classifications*. MIS Quarterly. 38(1), p.73-100.
Wachter, S., Mittelstadt, B., & Russell, C. (2017). *Counterfactual explanations without opening the black box: automated decisions and the GDPR.*
Ribeiro, M. T., Singh S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier.* Proceedings of KDD '16, p.1135-1144.
Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.*
Nguyen, D. (2017). *Comparing automatic and human evaluation of local explanations for text classification.*

References

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* arXiv preprint arXiv: 1702.08608,

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2013). *An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models.*

Moeyersoms J, d'Alessandro B, Provost F, & Martens D. (2016). *Explaining classification models built on high-dimensional sparse data*. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York, p. 36–40.

Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017). *Interpreting black-box classifiers using instance-level visual explanations.*

De Cnudde, S., Martens, D., Evgeniou, T., & Provost, F. (2017). *A benchmarking study of classification techniques for behavioral data.*

Arras, L., Horn, F., Montavon, G., Muller, K.-R., & Samek W. (2016). *Explaining predictions of non-linear classifiers in NLP.* Proceedings of the 1st Workshop on representation learning for NLP, p.1-7.

2. Explanation methods

LIME / SHAP

Perturbation: set feature value to zero / remove "evidence"

Original instance:

	Movie 1	Movie 2	Movie 3	Movie 4	…	Movie m-1	Movie m	Original predicted score
User x	1	1	0	1	…	1	1	0.96

Perturbed instances:

Weights		Movie 1	Movie 2	Movie 3	Movie 4	…	Movie m-1	Movie m	Predicted score (new label)
w1	z1	1	**0**	0	1	…	1	1	0.94
w2	z2	**0**	**0**	0	**0**	…	1	1	0.92
w3	z3	1	**0**	0	1	…	**0**	**0**	0.93
	…								

➔ **TRAIN SPARSE, LINEAR MODEL**

3. Evaluation criteria

- **Switching point** for **EDC**:

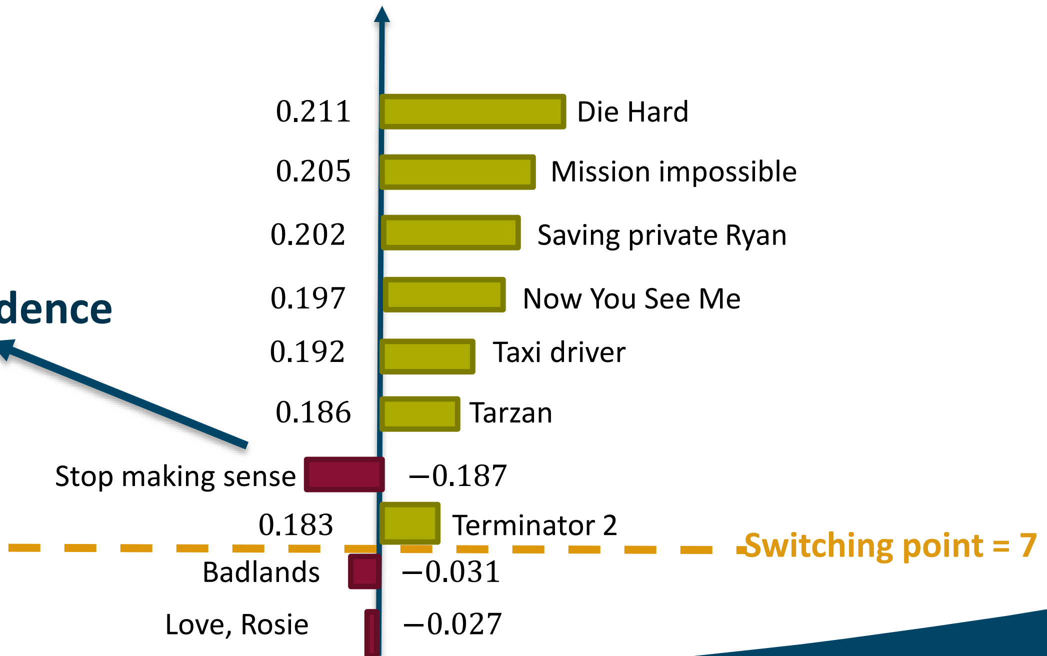
Relative importance

{Taxi Driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me}

Switching point = output size = 5

- **Switching point** for **LIME/SHAP**:

Ignore negative evidence



3. Evaluation criteria

To compare switching point, all methods should find one

EDC:

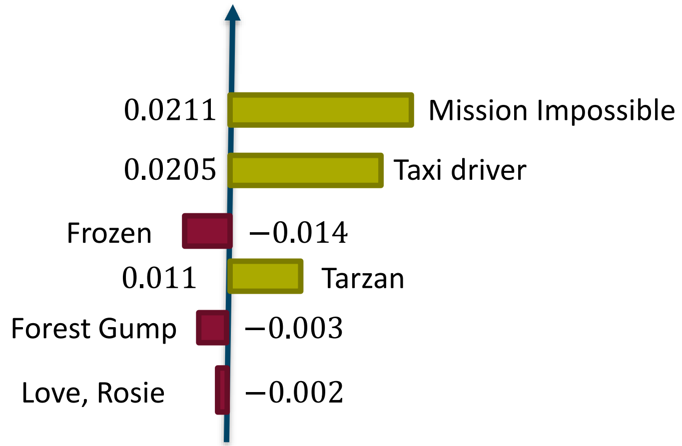
Relative importance

{Taxi Driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me}

➔ Switching point = 5

LIME for k=6:

➔ No switching point found



3. Evaluation criteria

To compare switching point, all methods should find one

EDC:

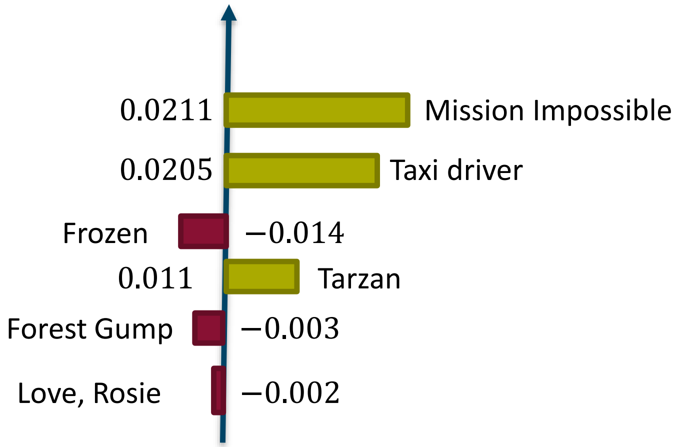
Relative importance

{Taxi Driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me}

➔ Switching point = 5

LIME for k=6:

➔ No switching point found



No comparison possible of ability to rank positive evidence from high to low relative importance

3. Evaluation criteria

To compare switching point, <u>all</u> methods should find one

EDC:

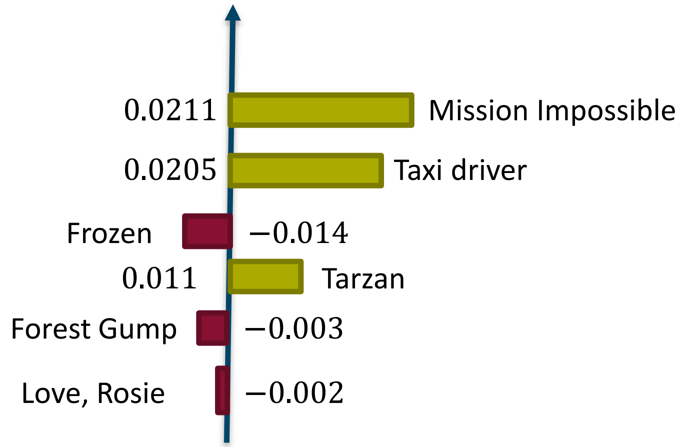
Relative importance

{Taxi Driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me}

➔ Switching point = 5

LIME for ~~k=6~~ k=#active features:

➔ No switching point found



No comparison possible of ability to rank positive evidence from high to low relative importance

➔ UNRESTRICT output size to measure switching point

3. Evaluation criteria

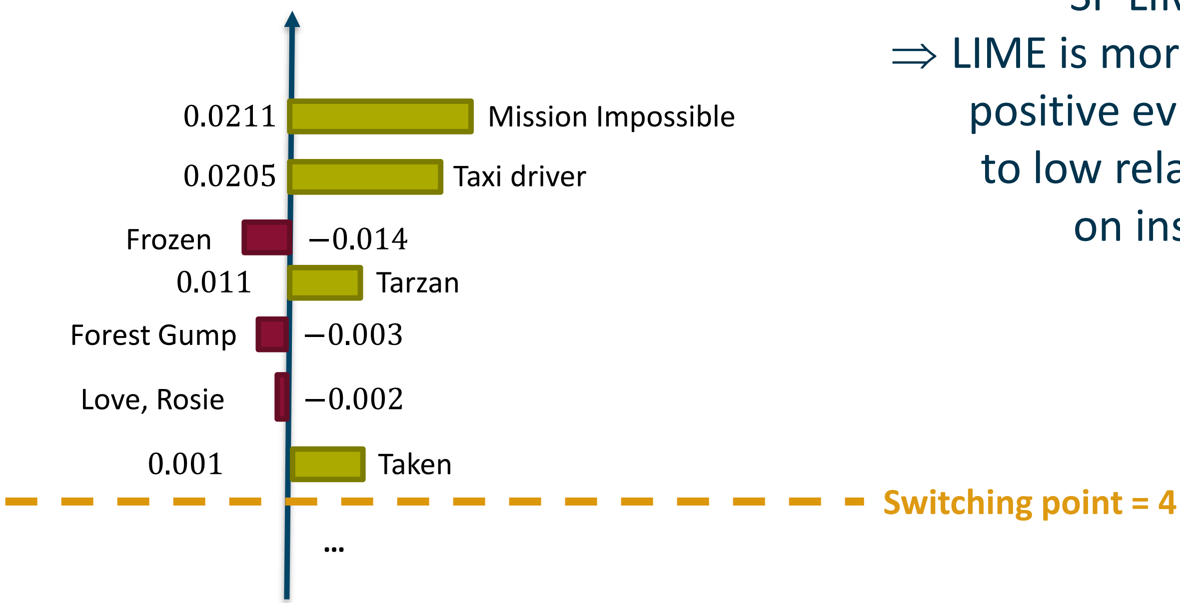
Example:

Relative importance

EDC output: *{Taxi Driver, Titanic, E.T., Taken, Gone girl}*

➔ Output size = switching point = 5

LIME output for ~~k=6~~ k=#active features:



Option 1:

SP LIME < SP EDC

$\Rightarrow$ LIME is more effective in ranking positive evidence from high to low relative importance on instance-level

3. Evaluation criteria

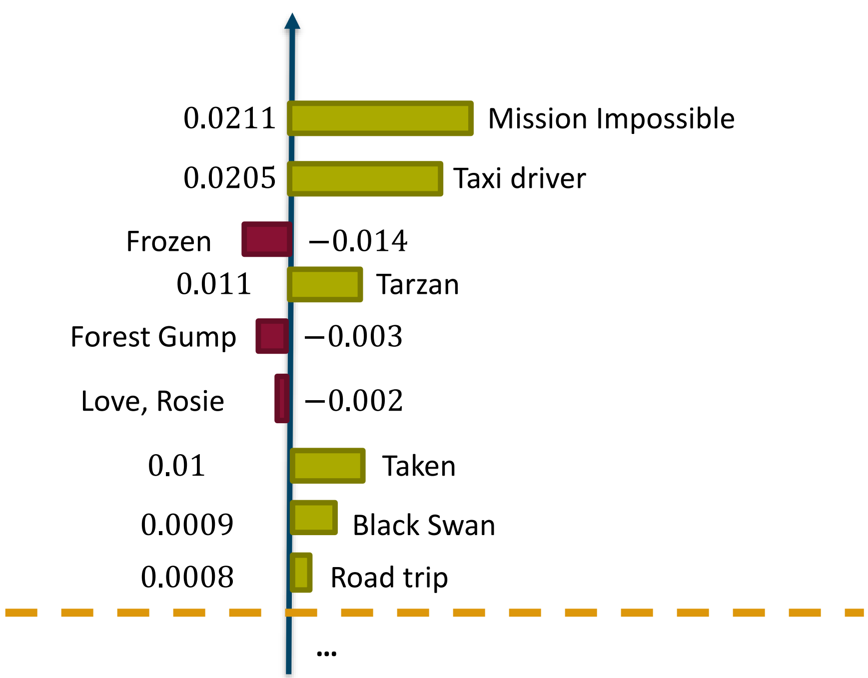
Example:

Relative importance →

EDC output: *{Taxi Driver, Titanic, E.T., Taken, Gone girl}*

➔ Output size = switching point = 5

LIME output for ~~k=6~~ k=#active features:



Option 2:

SP LIME > SP EDC

$\Rightarrow$ EDC is more effective in ranking positive evidence from high to low relative importance on instance-level

5. Results

Figure 4: Average relative switching point

5. Results

Table 1: Percentage generated & output size for LINEAR models

Data set	Textual/ behavioral	Explainer	Percentage output generated	Average output size	Average relative output size
Movielens	**Behavioral**	**EDC ≤ 10**	75.5%	**4.1 (2.7)**	**0.02 (0.02)**
n = 302 ṁ = 327 Model: LR		**LIME=10**	100%	10.0 (0)	0.07 (0.04)
		SHAP	100%	195.5 (112.6)	0.8 (0.1)
20news	**Textual**	**EDC ≤ 10**	92.1%	**2.4 (1.9)**	**0.07 (0.1)**
n = 151 ṁ = 69 Model: lin-SVM		**LIME=10**	100%	10 (0)	0.5 (1.2)
		SHAP	100%	29.1 (22.4)	0.6 (0.3)

(Standard deviations in parentheses)

5. Results

Table 2: Percentage generated & output size for NONLINEAR models

Data set	Textual/ behavioral	Explainer	Percentage output generated	Average output size	Average relative output size
Movielens	**Behavioral**	**EDC ≤ 10**	50.99%	**2.6 (2.2)**	**0.02 (0.03)**
n = 302 ṁ = 315 Model: MLP		**LIME=10**	100%	10 (0)	0.07 (0.1)
		SHAP	100%	174.95 (107.95)	0.9 (0.1)
20news	**Textual**	**EDC ≤ 10**	93.38%	**2.3 (1.98)**	**0.08 (0.1)**
n = 151 ṁ = 66 Model: rbf-SVM		**LIME=10**	100%	10 (0)	0.5 (1.2)
		SHAP	100%	31.8 (24.4)	0.7 (0.3)

(Standard deviations in parentheses)