Appendix: Human-grounded Evaluations of Explanation Methods for Text Classification

Piyawat Lertvittayakumjorn and **Francesca Toni**
Department of Computing
Imperial College London, UK
{pl1515, ft}@imperial.ac.uk

A CNN models

This section reports the performance of the trained CNN models on a test set of each dataset.

A.1 Amazon Dataset

1st CNN (better)	Prec.	Recall	F1	Support
Negative	0.92	0.89	0.90	50039
Positive	0.89	0.92	0.90	49961
micro avg	0.90	0.90	0.90	100000
macro avg	0.90	0.90	0.90	100000
2nd CNN (worse)	Prec.	Recall	F1	Support
Negative	0.82	0.81	0.81	50039
Positive	0.81	0.82	0.81	49961
micro avg	0.81	0.81	0.81	100000
macro avg	0.81	0.81	0.81	100000

Table 1: Precision, Recall, and F1 scores of both CNNs for the Amazon dataset

A.2 ArXiv Dataset

1st CNN (better)	Prec.	Recall	F1	Support
Computer science	0.94	0.93	0.93	10000
Mathematics	0.92	0.93	0.92	10000
Physics	0.96	0.94	0.95	10000
micro avg	0.94	0.94	0.94	30000
macro avg	0.94	0.94	0.94	30000
2nd CNN (worse)	Prec.	Recall	F1	Support
Computer science	0.96	0.74	0.84	10000
Mathematics	0.75	0.94	0.83	10000
Physics	0.89	0.88	0.89	10000
micro avg	0.85	0.85	0.85	30000
macro avg	0.87	0.85	0.85	30000

Table 2: Precision, Recall, and F1 scores of both CNNs for the ArXiv dataset

B Decision Trees

This section reports the decision trees performance in mimicking the CNNs' predictions (i.e., fidelity) on the test sets. All the DTs achieved over 80% macro-F1 in mimicking the CNNs predictions. As the F1 scores say, it's easier for the decision trees to mimic the behavior of the well-trained CNNs than the poor CNNs.

B.1 Amazon Dataset

1st CNN (better)	Prec.	Recall	F1	Support
Negative	0.84	0.84	0.84	48333
Positive	0.85	0.85	0.85	51667
micro avg	0.85	0.85	0.85	100000
macro avg	0.85	0.85	0.85	100000
2nd CNN (worse)	Prec.	Recall	F1	Support
Negative	0.81	0.82	0.82	49482
Positive	0.82	0.82	0.82	50518
micro avg	0.82	0.82	0.82	100000
macro avg	0.82	0.82	0.82	100000

Table 3: Performance of the decision trees in mimicking the CNNs' predictions for the Amazon dataset

B.2 ArXiv Dataset

1st CNN (better)	Prec.	Recall	F1	Support
Computer science	0.89	0.91	0.90	9971
Mathematics	0.89	0.87	0.88	10203
Physics	0.90	0.91	0.90	9826
micro avg	0.89	0.89	0.89	30000
macro avg	0.89	0.89	0.89	30000
2nd CNN (worse)	Prec.	Recall	F1	Support
Computer science	0.83	0.81	0.82	7653
Mathematics	0.82	0.88	0.85	12506
Physics	0.88	0.81	0.84	9841
micro avg	0.84	0.84	0.84	30000
macro avg	0.84	0.83	0.84	30000

Table 4: Performance of the decision trees in mimicking the CNNs' predictions for the ArXiv dataset

C Examples of the Explanations

Example 1: Amazon Dataset, (Actual: Positive, Predicted: Negative)

> "Source hip hop hits Volume 3: THe songs listed aren't even on the CD! I bought it for Bling Bling and it wasn't on the CD. the other songs are good, but not what I was looking for. Amazon needs to get the info right on this listing."

Top-5 evidence texts

- Random (W): . / get / hip / was / I
- Random (N): the CD ! I / the CD / needs to get the / info right on this / for .
- LIME: not / bought / 3 / info / Bling
- LRP (W): it / bought / . / listed / :
- LRP (N): ! I bought it / : THe songs listed / was looking for . / right on this listing / not what I
- DeepLIFT (W): it / bought / . / listed / :
- DeepLIFT (N): ! I bought it / : THe songs listed / was looking for . / right on this listing / not what I
- Grad-CAM-Text: n't even on the / not what I was / hits Volume 3 : / CD ! I / . Amazon needs to
- DTs: n't even on the / CD ! I

Example 2: ArXiv Dataset, (Actual: Physics (PH), Predicted: Computer Science (CS))

> "Multiple-valued Logic (MVL) circuits are one of the most attractive applications of the Monostable-to-Multistable transition Logic (MML), and they are on the basis of advanced circuits for communications. The operation of such quantizer has two steps : sampling and holding. Once the quantizer samples the signal, it must maintain the sampled value even if the input changes. However, holding property is not inherent to MML circuit topologies. This paper analyses the case of an MML ternary inverter used as a quantizer, and determines the relations that circuit representative parameters must verify to avoid this malfunction."

Top-5 evidence texts

- Random (W): not / This / one / basis / MML
- Random (N):) , and / , holding property is / are one of / sampled value even / circuit topologies
- LIME: paper / Logic / circuits / communications / applications
- LRP (W): paper / - / communications / topologies / the
- LRP (N): topologies . This paper / to - Multistable transition / valued Logic (MVL / circuits for communications . / the quantizer samples the
- DeepLIFT (W): paper / - / communications / Logic / the
- DeepLIFT (N): topologies . This paper / valued Logic (MVL / to - Multistable transition / circuits for communications . / the quantizer samples the
- Grad-CAM-Text: circuits for communications . / (MVL) circuits / MML ternary inverter used / topologies . This paper / - valued Logic
- DTs: MML ternary inverter / MVL) circuits are / advanced circuits / circuits for communications / to avoid this malfunction

Example 3: Amazon Dataset, (Actual: Positive, Predicted: Positive), Predicted scores: Positive (0.514), Negative (0.486)

"OK but not what I wanted: These would be ok but I didn't realize just how big they are. I wanted something I could actually cook with. They are a full 12" long. The handles didn't fit comfortably in my hand and the silicon tips are hard, not rubbery texture like I'd imagined. The tips open to about 6" between them.Hope this helps someone else know better if it's what they want."

Top-5 evidence texts

- Random (W): not / wanted / 'd / with / The
- Random (N): did n't / be ok / could actually cook / are hard / 12 " long .
- LIME: comfortably / wanted / helps / tips / fit
- LRP (W): are / not / 6 / hard / helps
- LRP (N): are hard , not / about 6 " between / not what I wanted / helps someone else know / wanted something I
- DeepLIFT (W): are / not / 6 / hard / helps
- DeepLIFT (N): are hard , not / about 6 " between / not what I wanted / helps someone else know / wanted something I
- Grad-CAM-Text: comfortably in my hand / I wanted : These / . The tips open / , not rubbery texture / Hope this helps someone
- DTs: imagined . The tips

Top-5 counter-evidence texts

- Random (W): texture / . / what / to / would
- Random (N): this helps someone else / , not / wanted something I / and the / I did n't
- LIME not / else / someone / ok / would
- LRP (W): : / tips / open / in / The
- LRP (N): . The tips open / : These would / in my hand and / could actually cook / I did n't realize
- DeepLIFT (W): : / tips / open / in / The
- DeepLIFT (N): . The tips open / : These would / in my hand and / could actually cook / I did n't realize
- Grad-CAM-Text: not what I wanted / not rubbery texture like / Hope this helps someone / would be ok / The handles did n't
- DTs: 'd imagined . / are . I wanted / would be ok

Example 4: ArXiv Dataset, (Actual: Computer Science (CS), Predicted: Mathematics (MA)), Predicted scores: Computer Science (0.108), Mathematics (0.552), Physics (0.340)

"The mnesor theory is the adaptation of vectors to artificial intelligence. The scalar field is replaced by a lattice. Addition becomes idempotent and multiplication is interpreted as a selection operation. We also show that mnesors can be the foundation for a linear calculus."

Top-5 evidence texts

- Random (W): intelligence / to / theory / is / by
- Random (N): replaced by a lattice / interpreted as a / linear calculus . / show that / The mnesor
- LIME: linear / a / idempotent / vectors / of
- LRP (W): lattice / theory / scalar / linear / of
- LRP (N): replaced by a lattice / . The scalar field / the adaptation of vectors / mnesor theory / a linear
- DeepLIFT (W): lattice / theory / scalar / linear / of
- DeepLIFT (N): replaced by a lattice / . The scalar field / the adaptation of vectors / mnesor theory / a linear
- Grad-CAM-Text: for a linear calculus / Addition becomes idempotent and / adaptation of vectors to / replaced by a lattice / mnesor theory is the
- DTs: Addition becomes idempotent and / becomes idempotent and multiplication

Top-5 counter-evidence texts

- Random (W): the / We / scalar / lattice / operation
- Random (N): lattice . Addition / The scalar / interpreted as a selection / for a linear calculus / .
- LIME: intelligence / scalar / field / The / lattice
- LRP (W): mnesors / interpreted / multiplication / can / foundation
- LRP (N): mnesors can be the / multiplication is interpreted as / to artificial intelligence / foundation for / field is
- DeepLIFT (W): interpreted / mnesors / multiplication / foundation / can
- DeepLIFT (N): mnesors can be the / multiplication is interpreted as / to artificial intelligence / foundation for / field is
- Grad-CAM-Text: . The scalar field / vectors to artificial intelligence / show that mnesors can / and multiplication is interpreted / The mnesor theory is
- DTs: vectors to artificial

D Score Distributions

This section presents the distributions of individual scores rated by human participants for each task and dataset. We do not include the random baselines in the plots to reduce the plot complexity.

D.1 Amazon Dataset

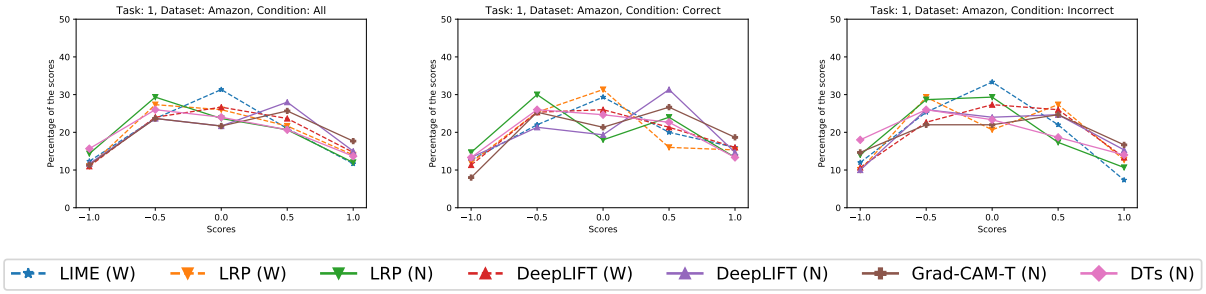

Figure 1: Distributions of individual scores from task 1 of the Amazon dataset ($\mathcal{A}$, ✔, ✘, respectively).

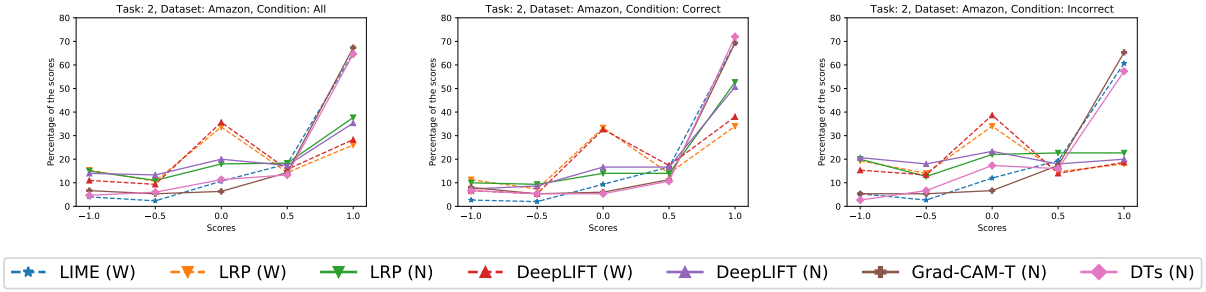

Figure 2: Distributions of individual scores from task 2 of the Amazon dataset ($\mathcal{A}$, ✔, ✘, respectively).

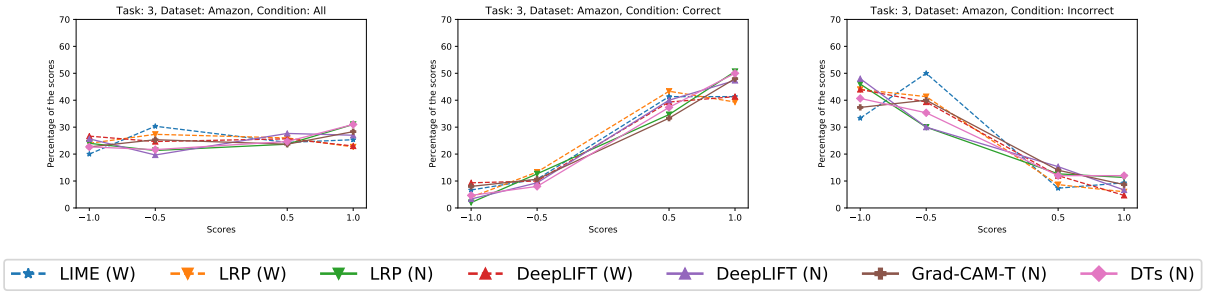

Figure 3: Distributions of individual scores from task 3 of the Amazon dataset ($\mathcal{A}$, ✔, ✘, respectively).

D.2 ArXiv Dataset

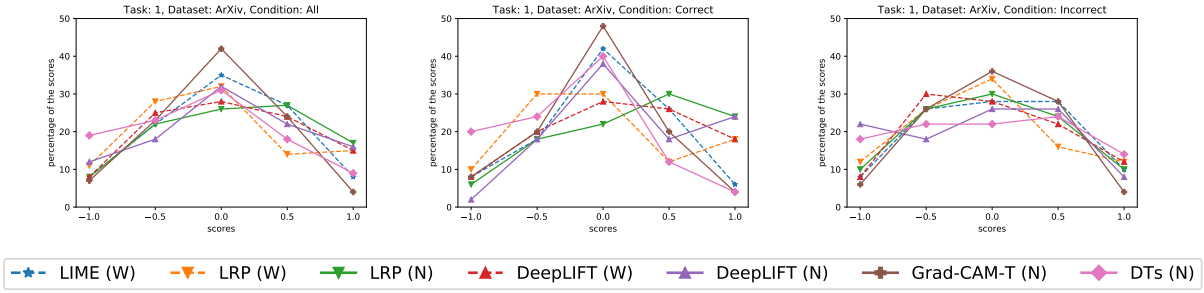

Figure 4: Distributions of individual scores from task 1 of the ArXiv dataset ($\mathcal{A}$, ✔, ✘, respectively).

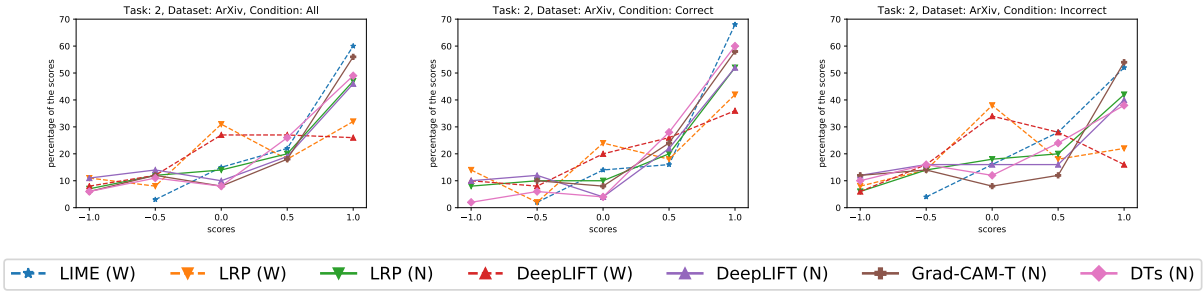

Figure 5: Distributions of individual scores from task 2 of the ArXiv dataset ($\mathcal{A}$, ✔, ✘, respectively).

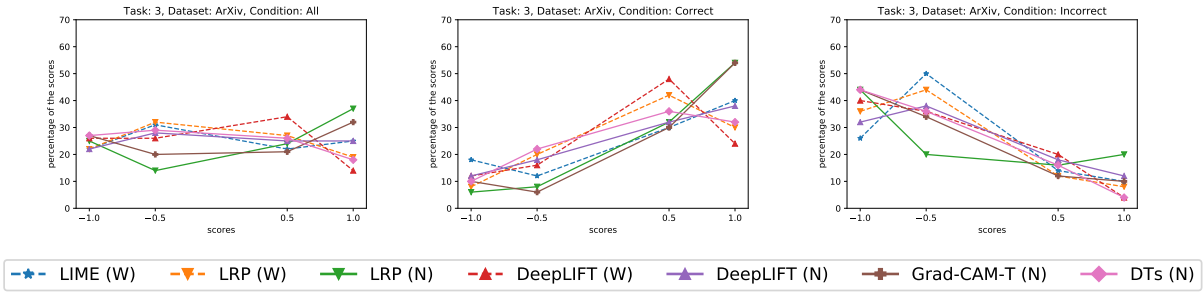

Figure 6: Distributions of individual scores from task 3 of the ArXiv dataset ($\mathcal{A}$, ✔, ✘, respectively).