Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Evaluating Deep Taylor Decomposition for Reliability Assessment in the Wild

Stephanie Brandl, Daniel Hershcovich, Anders Søgaard

University of Copenhagen
{brandl, dh, soegaard}@di.ku.dk

Abstract

We argue that we need to evaluate model interpretability methods 'in the wild', i.e., in situations where professionals make critical decisions, and models can potentially assist them. We present an in-the-wild evaluation of token attribution based on Deep Taylor Decomposition, with professional journalists performing reliability assessments. We find that using this method in conjunction with RoBERTa-Large, fine-tuned on the Gossip Corpus, led to faster and better human decision-making, as well as a more critical attitude toward news sources among the journalists. We present a comparison of human and model rationales, as well as a qualitative analysis of the journalists' experiences with machine-in-the-loop decision making.

Introduction

Deep neural NLP models increasingly assist humans in making documents easier to find and analyze. Generally, models are used for either *batch processing*, e.g., summarizing or indexing documents, or for *online decision support*, e.g., flagging probable fraud or toxic speech. Deep neural models are often thought of as black boxes, unable to provide rationales for their decisions, but recently, many methods have been developed for post-hoc interpretation of deep neural model predictions (Guidotti et al. 2018; Ancona et al. 2018; Poerner, Schütze, and Roth 2018; Atanasova et al. 2020). Most such methods provide rationales in the form of input token attributions. While alternatives exist (Søgaard 2021), we evaluate attributions below.

Rationales in the form of input token attributions have traditionally been evaluated automatically or through comparison with gold-standard, human rationales. Automatic evaluation methods look at the impact of removing (or keeping only) the tokens that are attributed most relevance or influence, while comparisons with human rationales typically evaluate predicted rationales in terms of matching metrics such as token-level F_1.

Related work Horne et al. (2019) evaluated if machine learning with and without feature attribution rationales can help *lay people* assess the reliability of claims. We follow their data collection set-up in most respects, but not their experimental protocol. We also rely on a different source of data. Horne et al. (2019) found a positive effect of model rationales on human decision making, but evaluated this only in terms of mean ratings for reliable and unrealiable articles, leaving it open whether human decision-making was more accurate at the article level. They found no positive effect of showing just model confidences. The most important difference between their study and ours, however, is that we evaluate the usefulness of reliability detection with feature attribution rationales on *professional experts* (journalists).

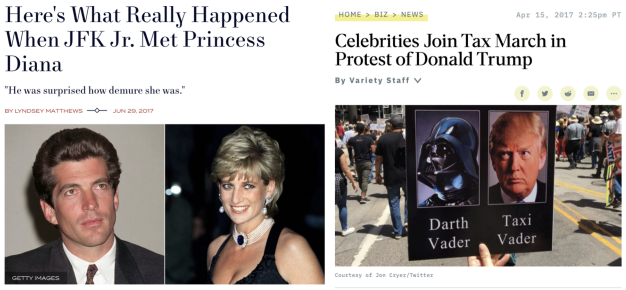

Figure 1: Celebrity news stories, the text of which (both longer than shown) are included in the GossipCop corpus. These examples are labeled *unreliable*. We ask professional journalists to relabel the texts, with or without the help of a RoBERTa-based document classifier and Deep Taylor Decomposition rationales.

More recently, Mohseni et al. (2021) evaluated the effect of model rationales (using self-explanatory models) across four fake news detection models, also with lay people. They found that using rationales is not significantly better than showing only model confidence, and generally found the effectiveness of explanations to vary across architectures. Our results align better with Horne et al. (2019) than with Mohseni et al. (2021), but provide a more nuanced picture and for a population of end users, who are trained to do reliability assessment, and who have a real need to perform this activity on a day-to-day basis.

In older work, Feng and Boyd-Graber (2019) evaluated the usefulness of model rationales in the context of human question answering. They recruited trivia experts and

novices to play QuizBowl with a computer as their teammate, with models returning both top-k predictions and rationales. Rationales was provided both in terms of feature attributions and training examples. Their work is also limited in several respects, relying on an inherently interpretable linear question answering model and therefore, in effect, not evaluating interpretability methods, but simply the usefulness of an interpretable model. Their task is also different than ours in having an open-ended output space. This means that simply listing top-10 candidate answers can contribute to human recollection of answers, an effect they exploit, but which we cannot for a binary classification task. Feng and Boyd-Graber (2019) find that only novices generally benefit from model rationales. Finally, QuizBowl is arguably not a task where interpretability is needed – in the sense that this is a necessary feature for decision support in reliability assessment to avoid the spread of misinformation. In sum, their work leaves open a) whether interpretability methods are useful when applied to deep neural models, b) whether interpretability methods can be useful to experts, and c) whether interpretability methods can be useful in situations where interpretability is critical. While Horne et al. (2019) and Mohseni et al. (2021) addressed a) and b), our work provides partial answers to all these three questions.

Other work that addresses human-in-the-loop evaluation of interpretability for deep neural models (a) includes Gonzalez and Søgaard (2020) and González et al. (2021), but both evaluate interpretability methods with lay people and on non-critical tasks, ignoring (b) and (c). Attempts to evaluate interpretability methods for experts performing critical tasks, have, to the best of our knowledge, been limited to automatic evaluation or evaluation against gold-standard human rationales.

Contributions In sum, our main contribution compared to previous work is that we evaluate feature attribution rationales for reliability assessment on professional journalists. While misinformation spread in social media is devastating, misinformation accelerates when authoritative platforms, including traditional media, pass on unreliable stories (Soares and Recuero 2021). As a way of evaluating interpretability methods, professional journalists present a much higher bar. Journalists are already trained to assess the reliability of news stories, and predictions and rationales have to be of very high quality to assist, rather than bias, such professional end users. In addition, we also evaluate a novel, state-of-the-art interpretability algorithm (Deep Taylor Decomposition for deep neural networks) on a new source of data, namely a corpus of celebrity gossip stories, a domain which is arguably a greater source of misinformation than any other domain and has been estimated to account for up to half of all online misinformation (Acerbi 2019). Finally, we consider the impact of the demographics and experience of the professional journalists on the usefulness of feature attribution.

Experimental Setup

Participants We recruited 25 professional journalists (6 female, 17 male, 2 without response) to participate in our experiment. All journalists were employed at the time of the experiment in the same media house, working for a tabloid newspaper, and had in this capacity daily contact with reliability assessment of celebrity-related news. The participants were compensated through salary.

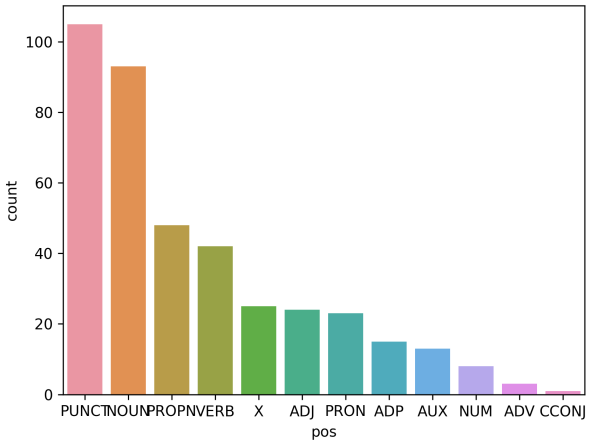

Figure 2: POS distribution of most highlighted words

	Time	Error
Text	**45.8**	0.31
+ Confidence	49.2	0.25
+ Confidence + Feat.attr.	48.6	**0.22**

Table 1: Deep Taylor Decomposition (last row) enables faster (Time) and better (Error) reliability assessment, even for professional journalists.

Data We sampled 80 articles for our experiments from the GossipCop section of the FakeNewsNet (Shu et al. 2019).[1] The articles were split in 2 batches of 40 articles, half of which our model (see below) had classified incorrectly. In each condition, each article was seen by a minimum of three journalists. Gossip as a genre has the advantage that reliability assessment here should be relatively unbiased by domain experience and political alignment, evaluating instead the plausibility of a reported state of affairs. We provide an example in Table 1.

Model We train a binary document classifier by fine-tuning RoBERTa-Large (Liu et al. 2019) on the rest of the GossipCop data. RoBERTa-Large – a pre-trained English language model based on Transformer blocks (Vaswani et al. 2017) – is augmented with a simple classification architecture to calculate cross-entropy. Our architecture is completely standard and applies drop-out to the Transformer's last layer hidden state. The regularized output is then fed into a single fully-connected layer and a softmax activation function to scale logits to probabilities. We use Deep Taylor De-

[1] github.com/KaiDMML/FakeNewsNet

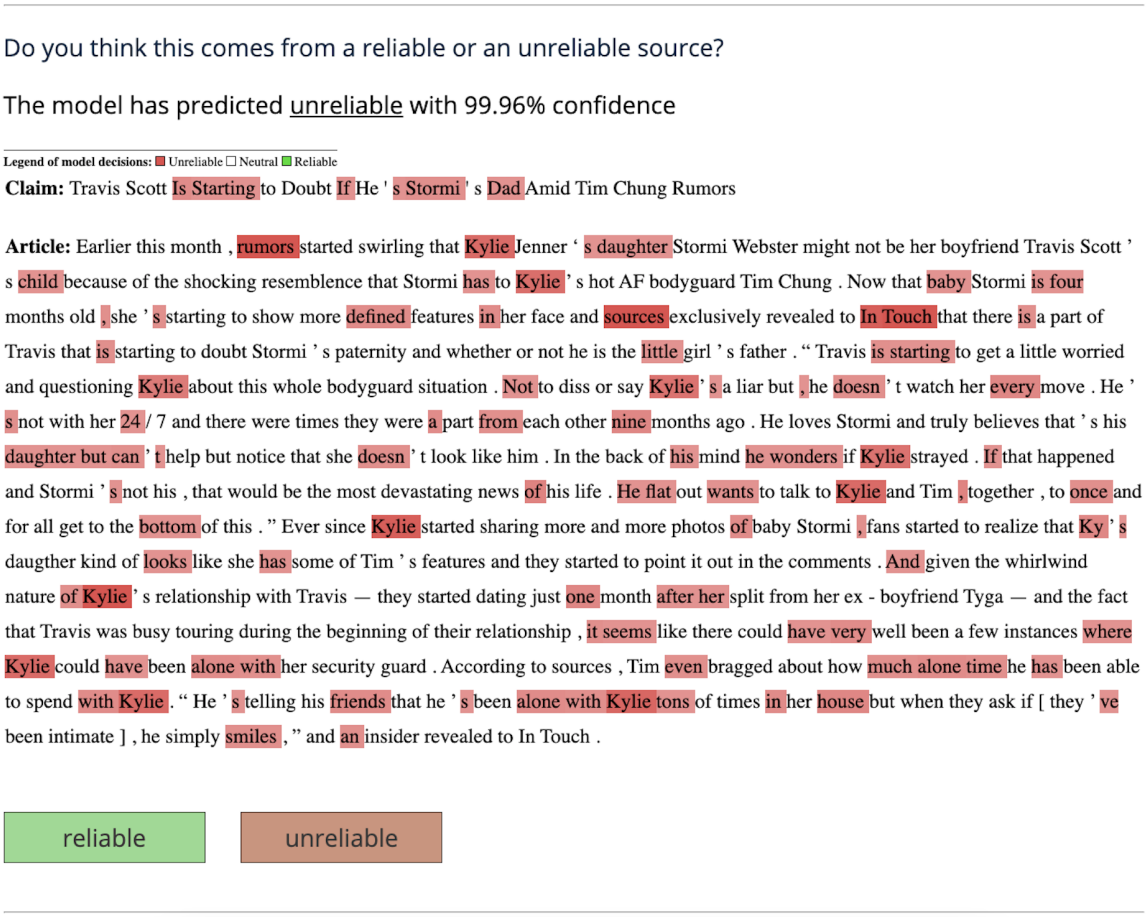

Figure 3: The interface used in our human reliability assessment with professional journalists receiving model prediction (with confidence) and rationales as feedback.

composition (Chefer, Gur, and Wolf 2021) to assign feature attribution scores to input words. See code for details. Deep Taylor Decomposition is non-trivial in Transformer models because propagation involves attention layers and skip connections, but Chefer, Gur, and Wolf (2021) present a novel, consistent method for computing attribution scores across such layers and connections.

Predictive words Deep Taylor Decomposition assigns feature attribution scores to input words. Reviewing aggregate statistics across the examples used in our experiments, we see the POS distribution for the most highlighted words in Figure 2. We see our model is very sensitive to punctuation, e.g., misplaced commas, and nouns, e.g., first names without last names. Several of the journalists, when asked for signs of unreliability, mentioned poor grammar as the main indicator outside of metadata.

Stimuli presentation The feature attribution scores were displayed to the journalists in one of three conditions; see Figure 3 to see the user interface. The interfaces for the other conditions were identical, except without color highlights and/or model predictions and confidence scores.

Metrics We evaluate the time used by professional journalists to assess the reliability of the news articles in terms of average wall-clock time in seconds. We evaluate their ability to assess reliability by evaluating their accuracy (1-0 loss) compared to the GossipCop gold-standard annotations.

Results

Our main results are listed in Table 1. We see that the time journalists take to assess reliability is not significantly impacted by the model feedback. We see a 3.5s increase on average when introducing model predictions and confidence scores, but on the other hand, the highlighting seems to increase reading speed a bit, leading to 0.6s faster reading times than in the setup without feature attributions.

More importantly, human error rate is substantially lower for professional journalists when they receive model feedback. In the baseline condition of *no* model feedback, they are only able to correctly estimate the reliability of two in three documents, whereas with model confidence scores, their accuracy is three in four, i.e., the error rate is 0.25. Model rationales in the form of feature attribution further decreases the error rate to 0.22.

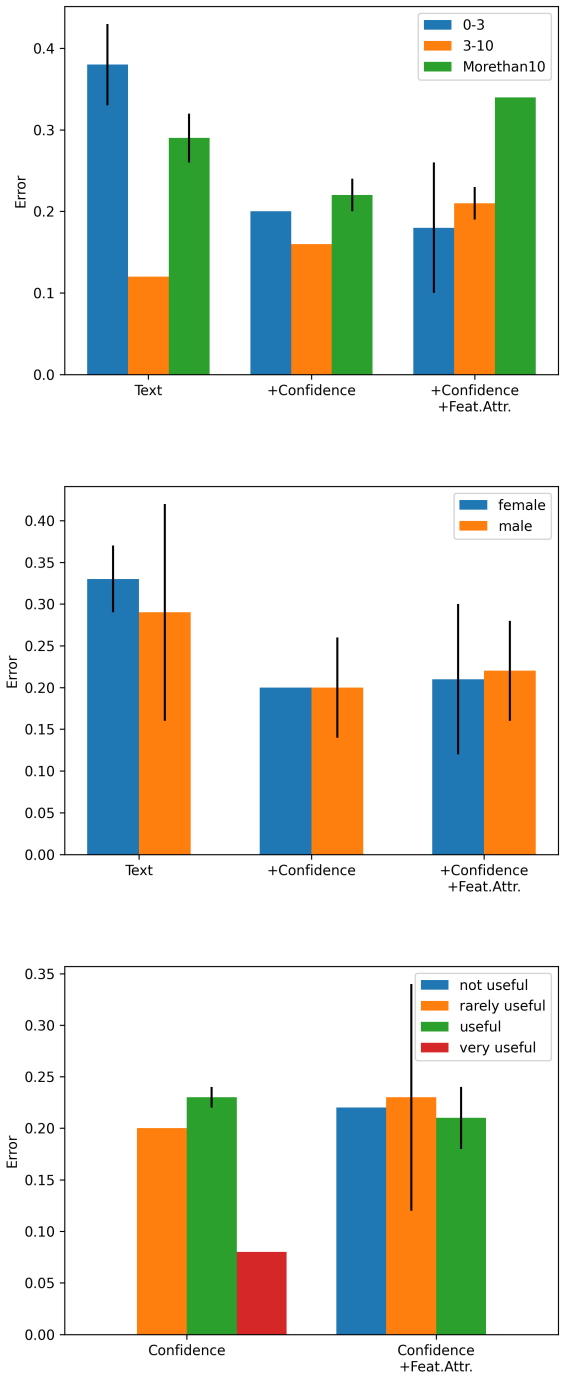

Figure 4: Error rate for all 3 conditions with journalists grouped by experience (top), gender (middle), and by how useful they rated the model feedback (right). Error bars are standard deviation across bootstrapped samples.

Analysis

As part of our experiment, we also asked the journalists to volunteer information about their seniority, gender, and how useful they found the model feedback. We use this additional survey data to perform a more fine-grained analysis of the usefulness of the feature attribution scores of Deep Taylor Decomposition for the task of reliability assessment.

Experience The 25 professional journalists in our experiment all work for the same media house, but some are more senior than others. The usefulness of feature attribution has previously been said to correlate with level of experience (Feng and Boyd-Graber 2019). Our data corroborates this claim (see Figure 4, left chart). In fact, it turns out the positive effect of interpretability is entirely with journalists with less than three years of experience. Model feedback, in contrast, seems to hurt the reliability assessments of more experienced journalists.

Demographics We also considered whether reported gender of the participants had an effect on the usefulness of model feedback, i.e., are men or women more prone to model suggestions? The positive effect of interpretability methods seems insenstive to the gender of the journalists. See Figure 4 (middle) for numbers.[2]

Usefulness Finally, we examined whether the perceived usefulness of the model feedback by the journalists correlated with their error. Mostly, it did not, but the small group of journalists who found the model feedback *very useful* had very low error rates. Most journalists rated model predictions as *useful* or *very useful*, whereas most journalists rated model predictions and rationales as *not useful* or *rarely useful*. This is remarkable in light of the better performance of journalists relying on both model predictions and rationales. It seems that **showing model predictions and rationales to professional journalists improves their reliability assessments, but hurts their perception of the usefulness of having a model in the loop**.[3]

Qualitative feedback The journalists were also asked to provide qualtiative feedback on their experience with model-supported reliability assessment. Example feedback ranged from *the model and the words marked worked fine for me* to *semed random at times*. Several journalists explicitly referred to the model as a source of bias, acknowledging its influence on their assessments.

Conclusions

We evaluated feature attributions of Deep Taylor Decomposition applied to a deep neural document classifier based on fine-tuned RoBERTa-Large representations, in a model-assisted human decision-making experiment with 25 professional journalists performing in-the-wild reliability assessments of English celebrity news stories. We found feature attribution to have a positive effect on the ability of journalists to assist the reliability of these stories. We also saw, however, that this effect was only with journalists with less than three years of experience. Interestingly, while the feature attribution scores had a positive net effect on reliability

[2] 23/25 reported they identified as either male or female.

[3] This seems paradoxical, but there are many similar paradoxes in human decision-making, e.g., Simon, Wang, and Keller (2011).

assessment, a setup where only model predictions and confidence scores were provided, was perceived of as more useful by the participants. We make the data and code involved in our experiments publicly available for replication and future work.[4]

Acknowledgements

This work was partially funded by the Platform Intelligence in News project, which is supported by Innovation Fund Denmark via the Grand Solutions program. Thanks to Ana Valeria Gonzalez for help with designing our experiments.

References

Acerbi, A. 2019. Cognitive attraction and online misinformation. *Palgrave Communications*, 5: 15.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–3274. Online: Association for Computational Linguistics.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer Interpretability Beyond Attention Visualization. In *CVPR*.

Feng, S.; and Boyd-Graber, J. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 229–239.

González, A. V.; Bansal, G.; Fan, A.; Mehdad, Y.; Jia, R.; and Iyer, S. 2021. Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1103–1116. Online: Association for Computational Linguistics.

Gonzalez, A. V.; and Søgaard, A. 2020. The Reverse Turing Test for Evaluating Interpretability Methods on Unknown Tasks. In *NeurIPS 2020 Workshop on Human And Model in the Loop Evaluation and Training Strategies*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5).

Horne, B. D.; Nevo, D.; O'Donovan, J.; Cho, J.-H.; and Adalı, S. 2019. Rating reliability and bias in news articles: Does AI assistance help everyone? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 247–256.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Cite arxiv:1907.11692.

[4]https://github.com/coastalcph/reliability-wild

Mohseni, S.; Yang, F.; Pentyala, S.; Du, M.; Liu, Y.; Lupfer, N.; Hu, X.; Ji, S.; and Ragan, E. 2021. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 421–431.

Poerner, N.; Schütze, H.; and Roth, B. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 340–350. Melbourne, Australia: Association for Computational Linguistics.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2019. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. arXiv:1809.01286.

Simon, J.; Wang, Y.; and Keller, L. R. 2011. *Paradoxes and Violations of Normative Decision Theory*. American Cancer Society. ISBN 9780470400531.

Soares, F.; and Recuero, R. 2021. How the mainstream media help to spread disinformation about COVID-19. *M/C Journal*, 24(1).

Søgaard, A. 2021. *Explainable Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. ISBN 9781636392141.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.