# Project Proposal: Rationalizing Neural Predictions Replication in PyTorch and Robustness Analysis

**Jason Wang** ⋆
jw7383@nyu.edu

**Chloe Zheng** ⋆
cz1300@nyu.edu

**Xu Han** ⋆
xh852@nyu.edu

## Paper Summary

Many recent advances in NLP use neural models that often come with a significant trade-off between accuracy and interpretability, which is a major downside in applications where transparency is necessary. The paper Rationalizing Neural Predictions addresses this concern by proposing a method to increase interpretability for rationale analysis. The paper's approach combines two modular components, a generator and an encoder, to extract short, continuous pieces of input text as justification for the prediction. The generator is a probability distribution over text fragments as candidate rationales that are then passed through the encoder for predictions. The loss function uses mean squared error and regularization that motivates shorter, continuous rationales. The paper implements its novel approach on two datasets, a BeerAdvocate dataset comprised of multi-aspect reviews and an AskUbuntu QA dataset. For rationale selection evaluation, precision was used in the BeerAdvocate dataset and MAP was used in the AskUbuntu dataset. In the implementation, RCNN's are used to represent both the generator and encoder.

## Core Experiments

Our core experiment will center around building the generator-encoder models. The model in the paper is built with Theano, but we will recreate the RNN model in PyTorch and perform both sentiment prediction and rationale selection on the BeerAdvocate review dataset.

**Timeline Draft (Deadline - Action Items)**
10/28 - EDA of BeerAdvocate dataset, Begin PyTorch implementation (Jason)
11/4 - Complete implementation, training (Jason)
11/11 - HP Tuning/Debug/Testing (Xu)
11/18 - EDA of RT reviews dataset, Begin extension (Chloe), Progress Update (All)
11/25 - Complete extension (Chloe), Finalize conclusions (All)
11/27 - Construct poster (All)
12/2 - Final Report (All)
12/9 - Finalize and practice presentation (All)

## Computational Feasibility

The authors provide a sample of the original BeerAdvocate dataset, about 520,000 multi-aspect reviews (out of the full 1.5 million reviews) split 50-50 between training and testing. The reviews contain multiple sentences that describe the overall impression, appearance, smell, palate, and taste of beers accompanied by ratings from 1 to 5 stars for each aspect. The manual annotations on the test data used for rationale analysis are also provided. Pre-trained word embeddings are also used to train the model, so we will use Stanford's GloVe embeddings. Google Colab and NYU HPC will be sufficient to run our core experiment as well as our extension.

## Extension

In the original paper, the authors were interested in future studies exploring the versatility of the novel generator-encoder approach and promoting interpretability in other domains. To contribute in this effort, we will test the robustness and interpretability of the model on a dataset similar to BeerAdvocate review dataset, Rotten Tomatoes Audience Reviews, comprised of 65,000 single-aspect reviews. To evaluate the robustness, we will compare results for the new dataset to results for the BeerAdvocate dataset evaluated on mean squared error for sentiment prediction and precision for rationale analysis. If time allows, a fine-tuned model will be provided.

## References


Tao Lei, Regina Barzilay, Tommi Jaakkola. 2016. Rationalizing Neural Predictions. arXiv:1606.04155

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078.

B Kim, JA Shah, and F Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In Advances in Neural Information Processing Systems

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015. Molding cnns for text: non-linear, non-consecutive convolutions. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?": Explaining the predictions of any classifier. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Omar Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pages 260–267.

Ye Zhang, Iain James Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. CoRR, abs/1605.04469.

Michiel Hermans and Benjamin Schrauwen. 2013. Training and analysing deep recurrent neural networks. In Advances in Neural Information Processing Systems, pages 190–198.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv:1707.07250

M.D. Devika, C. Sunitha, Amal Ganesh. 2016. Sentiment Analysis: A Comparative Study on Different Approaches. In Procedia Computer Science, Volume 87, Pages 44-49, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2016.05.124.

Shi, Tian and Rakesh, Vineeth and Wang, Suhang and Reddy, Chandan K. 2019. Document-Level Multi-Aspect Sentiment Classification for Online Reviews of Medical Experts. In Association for Computing Machinery.

L. Stappen, A. Baird, L. Schumann and S. Bjorn. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. In IEEE Transactions on Affective Computing. doi: 10.1109/TAFFC.2021.3097002.

Finale Doshi-Velez, Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608.