1.AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models:

The paper "AI Explainability 360: An Open-source Toolkit for Explainable AI" addresses the growing need for AI algorithms to explain their outputs in high-stakes societal applications. As AI algorithms become more prevalent, there is an increasing demand from multiple stakeholders for these algorithms to provide explanations. However, different personas of consumers of explanations have different requirements for explanations, making it challenging to address these needs effectively.

To tackle this challenge, the authors introduce the AI Explainability 360 toolkit. This toolkit is an open-source Python toolkit that offers ten diverse and state-of-the-art explainability methods and two evaluation metrics. It aims to provide a comprehensive set of tools and techniques for explaining AI algorithms' outputs. The toolkit includes not only the software itself but also guidance material, tutorials, and an interactive web demo. These resources are designed to introduce AI explainability to different audiences, making it accessible and understandable for users with varying levels of expertise.

One crucial aspect of the AI Explainability 360 toolkit is the taxonomy it provides. This taxonomy helps entities requiring explanations navigate the space of interpretation and explanation methods. It not only covers the methods included in the toolkit but also encompasses the broader literature on explainability. By organizing the methods according to their place in the AI modeling pipeline, the taxonomy allows users to identify gaps where more explainability methods are needed. This feature makes the toolkit a valuable platform for incorporating new explainability methods as they are developed.

In summary, the paper presents the AI Explainability 360 toolkit as a comprehensive solution for addressing the increasing demand for AI algorithms to explain their outputs. By providing a wide range of explainability methods, evaluation metrics, and a taxonomy, the toolkit offers a valuable resource for data scientists and other users. It not only helps them understand and interpret AI algorithms' outputs but also facilitates the development of new explainability methods to meet evolving needs in the field of AI explainability.

2. Interpretable Neural Network Approach to Predicting and Understanding Politeness in Natural Language Requests:

The paper presents an interpretable neural network approach to predicting and understanding politeness in natural language requests. The authors use simple convolutional neural networks (CNNs) directly on raw text to avoid the need for manual identification of complex sentiment or syntactic features. Their models perform better than previous feature-based models for politeness prediction.

The authors also focus on understanding what these successful neural networks are actually learning. They present several network visualizations based on activation clusters, first derivative saliency, and embedding space transformations. These visualizations help identify subtle linguistic markers of politeness theories. The analysis reveals multiple novel, high-scoring politeness strategies that, when added back as new features, reduce the accuracy gap between the original featurized system and the neural model. This provides a clear quantitative interpretation of the success of these neural networks.

The introduction section of the paper discusses politeness theories, including key components such as modality, indirection, deference, and impersonalization. Positive politeness strategies aim to make the hearer feel good through offers, promises, and jokes, while negative politeness examples include favor seeking, orders, and requests. Differentiating among politeness types is challenging due to factors such as context, relative power, and culture.

Previous work by Danescu-Niculescu-Mizil et al. (2013) proposed a computational framework for predicting politeness in natural language requests using various lexical and syntactic features based on key politeness theories. However, manually identifying these politeness features is difficult due to the complexity of the theories and the use of subtle markers and non-literal cues in natural language.

The paper proposes using CNNs to address politeness prediction directly on the raw text. This approach eliminates the need for complex, manually-defined linguistic features while still achieving better performance than featurized systems. The authors then present visualization strategies to interpret what the neural networks have learned. These strategies include activation clustering, first derivative saliency, and embedding space transformations, inspired by similar strategies in computer vision and recently adopted in natural language processing (NLP) for recurrent neural networks.

The activation clustering method rediscovers and extends several manually defined features from politeness theories and uncovers multiple novel strategies. The first derivative saliency technique allows the identification of the impact of each phrase on the final politeness prediction score through heatmaps, revealing useful politeness markers and cues. The authors also plot lexical embeddings before and after training to show how specific politeness markers move and cluster based on their polarity.

In conclusion, the paper presents an interpretable neural network approach to predicting and understanding politeness in natural language requests. The authors demonstrate the effectiveness of their models and provide insights into what these neural networks are learning through various visualization strategies. This work contributes to the field of politeness prediction and provides a quantitative interpretation of the success of neural networks in this task.

3. A causal framework for explaining the predictions of black-box sequence-to-sequence models:

The paper "Interpretable Sequence-to-Sequence Models for Natural Language Processing" by Ribeiro et al. introduces a method for interpreting the predictions of black-box structured input-structured output models. These models are often used in natural language processing tasks such as machine translation and summarization. However, they lack interpretability due to their complex architectures and large number of parameters.

The proposed method aims to provide explanations for the predictions made by these black-box models. It does so by identifying groups of input-output tokens that are causally related. To infer these dependencies, the method perturbs the inputs and queries the black-box model. The responses are used to generate a graph over tokens, and a partitioning problem is solved to select the most relevant components.

The paper focuses on sequence-to-sequence problems in natural language processing and adopts a variational autoencoder to generate meaningful input perturbations. The method is tested on several NLP sequence generation tasks.

The authors highlight the importance of interpretability in complex predictors. They argue that interpretability can lead to increased trust in model predictions, error analysis, model refinement, and detection of biases. While model interpretability, making the architecture itself interpretable, is challenging, prediction interpretability, explaining specific predictions, can be achieved for black-box systems.

In this work, the authors propose an approach to prediction interpretability that only requires oracle access to the black-box model. They focus on structured prediction, where both inputs and outputs are combinatorial objects. The explanation provided by their method consists of sets of input and output tokens that are causally related. These causal dependencies are identified by analyzing perturbed versions of the inputs.

The authors demonstrate the effectiveness of their method by recovering known dependencies. They show that a grapheme-to-phoneme dictionary can be largely recovered using their method. Additionally, they compare the explanations provided by their method with the attention scores used by a neural machine translation system and find close resemblance.

Overall, the paper presents a novel approach to interpreting the predictions of black-box structured input-structured output models in natural language processing. The method focuses on sequence-to-sequence problems and provides explanations consisting of causally related input and output tokens. The authors demonstrate the effectiveness of their method through quantitative evaluations and comparisons with attention scores.

4. Pathologies of Neural Models Make Interpretations Difficult:

The paper investigates the limitations of current interpretation methods for neural models in NLP and introduces a new approach called input reduction. The goal of interpretation methods is to identify the most important input features that contribute to the model's predictions. This is typically done through input perturbation or gradient-based methods.

Input perturbation measures the decrease in model confidence when a word is removed from the input. Similarly, gradient-based methods determine feature importance based on the gradient with respect to that word. These methods generate heatmaps that highlight the important words influencing the model's prediction.

To understand the limitations of these interpretation methods, the authors propose input reduction. This process iteratively removes the least important words from the input while maintaining the model's prediction. The expectation is that the remaining words after input reduction should be important for the prediction and align with the leave-one-out method's selections, which are known to closely match human perception.

However, the authors find that the reduced examples obtained through input reduction do not provide meaningful explanations of the original prediction. Instead, they resemble adversarial examples, where the reduced input is nonsensical to humans but still results in the same high-confidence prediction from the model.

The paper highlights the pathological behaviors of neural models trained with maximum likelihood and draws connections to adversarial examples and confidence calibration. These findings reveal difficulties in interpreting neural models and suggest the need for improvements.

To address these deficiencies, the authors propose fine-tuning the models by encouraging high entropy outputs on reduced examples. This approach aims to make the models more interpretable under input reduction without sacrificing accuracy on regular examples.

In conclusion, the paper explores the limitations of existing interpretation methods for neural models in NLP and introduces input reduction as a new approach. The findings reveal the pathological behaviors of neural models and propose a solution to improve interpretability.

5. MathQA: Towards Interpretable MathWord Problem Solving with Operation-Based Formalisms:

The paper "MathQA: A Large-Scale Dataset of Math Word Problems with Interpretable Solutions" introduces a new dataset called MathQA, which aims to address the challenges faced by current datasets in the domain of math word problem solving. These challenges include small scale and lack of precise operational annotations over diverse problem types.

The authors propose a new representation language to model precise operation programs corresponding to each math problem. This representation language improves both the performance and the interpretability of the learned models. The MathQA dataset significantly enhances the existing AQuA dataset by providing fully-specified operational programs.

To solve math word problems, the authors introduce a neural sequence-to-program model that is enhanced with automatic problem categorization. This model outperforms competitive baselines on both the MathQA and AQuA datasets, although the results are still lower than human performance.

The paper highlights the complexity of math word problem solving, which requires logical reasoning over implicit or explicit quantities expressed in text. The ability to extract salient information from natural language narratives and transform them into executable meaning representations is crucial for automatic solvers. Math word problems often involve narratives describing the progress of actions and relations over entities and quantities, requiring domain-specific knowledge and the ability to deduce implied constants.

The contributions of the paper include the introduction of a large-scale dataset of math word problems with densely annotated operation programs, a new representation language to improve the performance and interpretability of learned models, and a neural architecture that achieves competitive results on both the MathQA and AQuA datasets.

Overall, the paper presents MathQA as a valuable resource for future research in the field of math word problem solving and highlights the challenges that still need to be addressed to achieve human-level performance.

6.Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI):

The paper begins by highlighting the rapid adoption of artificial intelligence (AI) in our daily lives and the resulting shift towards a more algorithmic society. However, one major challenge with AI-based systems is their lack of transparency. While these systems can make powerful predictions, they often operate as black boxes, making it difficult to understand how they arrive at their decisions. This lack of explainability has sparked a new debate on Explainable AI (XAI).

XAI is a research field that aims to address the issue of transparency in AI systems. It is seen as essential for AI to continue progressing without disruption. The paper serves as an entry point for researchers and practitioners interested in XAI, providing an overview of the existing approaches, discussing trends in the field, and presenting major research directions.

The introduction section of the paper sets the context by highlighting the democratization of AI in everyday life. It mentions the significant investments being made in AI and its impact on society. While AI has become ubiquitous, there is a need for explainability in critical decisionmaking processes, such as disease diagnosis. The paper emphasizes the importance of understanding the reasons behind AI outcomes in such cases.

The paper acknowledges that AI algorithms, particularly machine learning algorithms, suffer from opacity, making it challenging to gain insight into their internal workings. This lack of transparency poses risks when important decisions are entrusted to AI systems. To address this issue, XAI aims to develop techniques that produce more explainable models while maintaining high performance levels.

The paper also discusses the landscape of XAI, highlighting the increasing attention and interest in the field. It mentions scientific events dedicated to XAI, such as the Fairness, Accountability, and Transparency (FAT-ML) workshop and the Workshop on Human Interpretability in Machine Learning (WHI).

In summary, the paper emphasizes the need for transparency and explainability in AI systems. It provides an overview of the existing approaches and trends in the field of XAI, serving as a valuable resource for researchers and practitioners interested in this topic.