NLP Explainability Techniques

Maria Ribalta i Albado - DL4NLP

Motivation

XAI is the discipline that explains why an AI model takes the decision it takes.
In 2020, Danilevsky et al provided a survey summarizing the most relevant XAI for NLP contributions were.
We have taken the explainability techniques on the paper and tried them out as well as proposed some additional techniques that have appeared since then.

Based on..

- A Survey of the State of Explainable AI for Natural Language Processing (Danilevsky et. al, 2020)

Models

- google-bert/bert-base-uncased
- meta-llama/Llama-2-7b-chat-hf
- sklearn's Random Forest

Data

- **Fine-tuning -** poem_sentiment
  - verses of poems classified from 0 to 3 according to the sentiment
  - 0: negative, 1: positive, 2: neutral, 3: mixed
- **Tests and experiments -** *All Too Well (10 Minute Version) (Taylor's Version) (From the Vault)* (manually labelled with the training's dataset criteria.

Feature Importance

Bertviz Attention Visualization System

The **Feature Importance** technique derives an explanation by investigating the importance scores of different features used to output the final prediction (e.g. latent features learned by NNs, attention scores, lexical features, etc). **BertViz** is the most famous library to plot attention scores and heads from any BERT model (or similars).
We have as well implemented the importance scores of each word of a sentence with **Representation Erasure** scores.

Word Representation Erasure Results

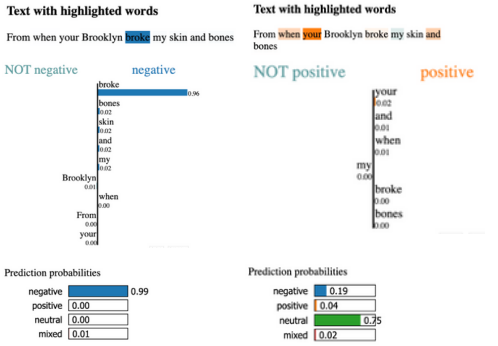
Example Driven

LRP results for the sentence: *Just between us, did the love affair maim you all too well?*

Example-Driven approaches explain the prediction of an input instance by identifying and presenting other instances, usually from available labeled data, that are semantically similar to the input instance.

LRP stands for layerwise relevance propagation and is an example-driven technique to showcase the relevance of certain features: positional embeddings, word embeddings, etc. The library **interpret_nlp** allows to extract the LRP scores from any Bert model, in our case, our fine-tuned Bert with the poem_sentiment data.

Surrogate Model



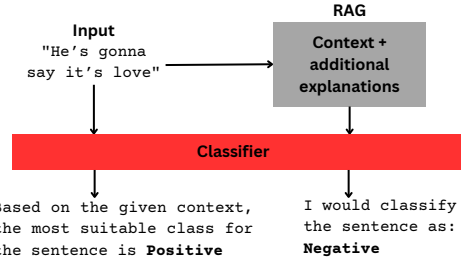
LIME's Explainability of a prediction of a Bert model

LIME's explainability of a prediction of an undertrained tree model

In **surrogate models**, predictions are explained by learning a second, usually more explainable model, as a proxy: this is explaining the decisions of the first model through a second. Surrogate model-based approaches are model-agnostic and can be used to achieve either local or global explanations.
LIME's algorithm consists on generating an input perturbation to evaluate how that affects the prediction and via the surrogate model assesses the explainability.

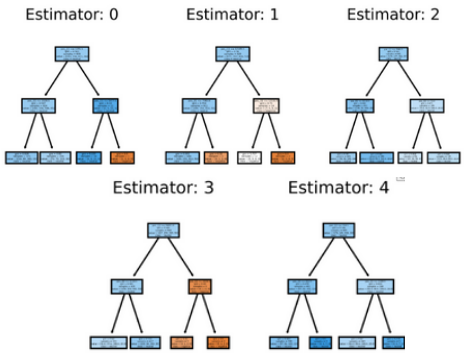
Provenance Based

The **provenance-based** technique is that in which explanations are provided by illustrating some or all of the prediction derivation process. This process is intuitive and effective and the final prediction is the result of a series of reasoning steps.
Techniques like **Chain of Thought** or **RAGs** could be considered provenance-based explainability techniques since the steps and additional information the model uses in order to make predictions are human understandable.
We generated a toy-RAG system that classifies verses within the context of the song.



Simulation of the process followed by our RAG-toy system to classify the sentence "he's gonna say it's love". In the context of the song, the sentence has a negative connotation.

Declarative Induction



Random Forest trained with our training dataset.

The **declarative Induction** technique is the one that works with human-readable representations, such as rules, trees, and programs are induced as explanations.

Decision trees generate a series of rules that step by step derive the data onto branches. **Random Forests** are the ensemble of different decision trees and are one of the most common and well-known self-explainable models.

Discussion

What is a good explanation? Who is the end user of that explanation? What task or objectives are we trying to evaluate? These questions don't permit the standardisation of simple processes for evaluating XAI methods. While there are different types of evaluation (informal examination, comparison to ground truth, human evaluation and others), how to evaluate an XAI approach is still an **open question**. In this project we have applied **human evaluation** to assess each of our approaches, but these cannot be directly compared:

- While **Feature Importance** is direct and easy to implement, recent studies show that, for instance, a feature such as attention is not always interpretable and that attention does not always lead to insight into model prediction.
- **Surrogate models** may have completely different mechanisms to make predictions than the original model. This has lead in many cases to concerns about the fidelity of this kind of approach.
- **Example-Driven** approaches are complex to code and each architecture will require a different implementation, which is costly in many levels.
- The **provenance-based** method is desirable but not always possible (we cannot always build interpretable architectures nor explainable datasets).
- **Declarative induction** techniques are simple and easy to understand, yet they are not always scalable (rule-based systems) or complex enough to generalize all tasks (random forests).

XAI is yet a huge field to investigate and we are sure many more techniques will arise with the creation of new models and new architectures,

Find the whole project in Github!