Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XXXVI cycle

Supervisor
Tania Cerquitelli
Elena Baralis

Explainable Artificial Intelligence techniques for Natural Language Processing tasks

PhD Candidate: *Salvatore Greco*

1. Context

Despite the high accuracy of deep learning models, their applicability in real-life settings is still limited. One reason is that they behave as black-boxes. Thus, model predictions should be explained to identify possible misbehavior or bias and increase the level of human trust and acceptance (XAI). Another reason is that they are trained and validated with static datasets. However, in real-life settings, data change and previously learned concepts may no longer be valid (i.e., concept drift). Therefore, models should be continuously monitored and explained.

2. Goal

My PhD goal is to design XAI solutions to explain the predictions of deep natural language models and concept drift detection techniques to explain models over time.

3. Method

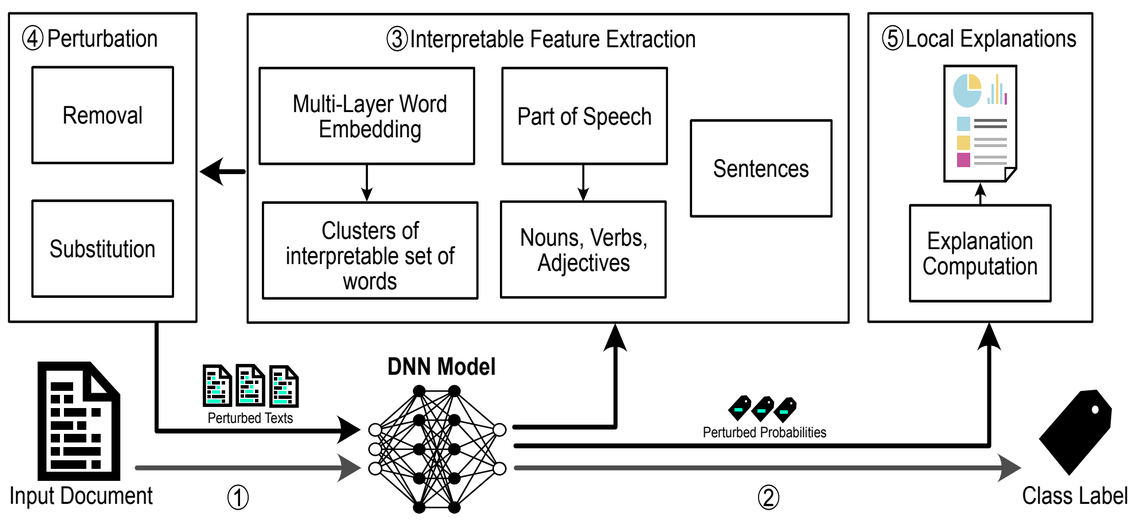

Figure 1. T-EBAnO: local explanation process

The XAI methodology, called **T-EBAnO** [1], produces **local explanations** (Figure 1) (i.e., to explain individual predictions of the model) by identifying the important words in the input text. It extracts a set of interpretable features (i.e., groups of words) with both **model-** and **domain-specific** techniques and quantifies their **influence** on the prediction with a **perturbation** process (i.e., removing words). An example is shown in Figure 2 (left) for a text predicted as *toxic* with the influent words in *cyan*.

T-EBAnO also produces **global explanations** to extract the most influent words for a class label. An example is shown in Figure 2 (right).

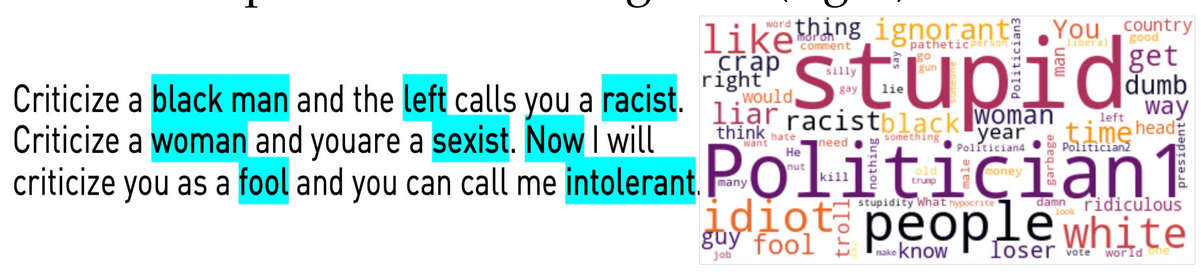

Figure 2. Local (left) and global (right) explanations

The methodology for monitoring and explaining the model over time is called **Drift Lens** [2]. It is an **unsupervised real-time** technique based on **embedding distribution distances** that performs **per-label drift detection** and **characterization**. Figure 3 shows an example of drift detection over time with an increasing drift.

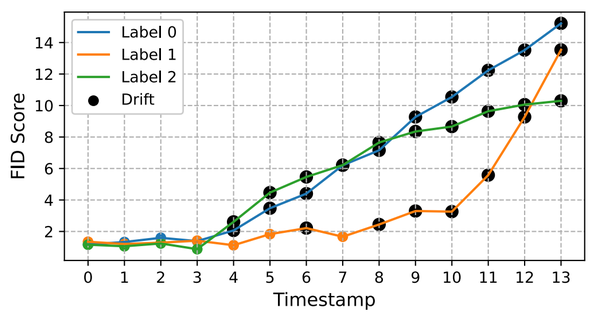

Figure 3. Incremental drift example

4. Results

The effectiveness and the quality of the explanations produced by **T-EBAnO** [1] have been evaluated on an extensive set of experiments with several deep language models and classification tasks. Moreover, the experimental comparison with other techniques proved that **T-EBAnO** [1] is more efficient and precise.

Drift Lens [2] was validated with a BERT model trained for topic detection, where the drift was simulated with text insertion of a new topic. Results show that **Drift Lens** is more accurate and faster than other state-of-the-art techniques.

5. References

1. Ventura, F., Greco, S., Apiletti, D., Cerquitelli, T. Trusting deep learning natural-language models via local and global explanations. Knowl Inf Syst 64, 1863–1907 (2022).
2. Greco S. and Cerquitelli T., "Drift Lens: Real-time unsupervised Concept Drift detection by evaluating per-label embedding distributions," 2021 International Conference on Data Mining Workshops (ICDMW), 2021, pp. 341-349.