Fake News Detection via Explainable Reinforcement Learning

Master Seminar Presentation

Szymon Socha

January 26, 2022

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

Table of Contents

1. **My plan for the Thesis**
2. **Topic**
   - Motivation
   - Literature Review
3. **Methodology**
   - Data
   - Reinforcement Learning
   - Extensions

Table of Contents

1 **My plan for the Thesis**

2 **Topic**
- Motivation
- Literature Review

3 **Methodology**
- Data
- Reinforcement Learning
- Extensions

Thesis

My plan for the master thesis in a nutshell:

- **Type:** scientific article
- **Title:** Fake News Detection via Explainable Reinforcement Learning
- **Keywords:** NLP, fake news, Reinforcement Learning, XAI, BERT

Table of Contents

1 My plan for the Thesis

2 **Topic**
- Motivation
- Literature Review

3 Methodology
- Data
- Reinforcement Learning
- Extensions

Fake News

I first became interested in the topic of fake news during the **2016 presidential election in the United States**. The term "fake news" has been widely used in the context of how false or misleading information can be disseminated through social media and other online platforms, and can have a significant impact on public opinion and political discourse.

Yet another moment when I realized the power of fake news was the **COVID-19 pandemic.** It was no longer just a political issue that influenced the outcome of the election. This time, human lives were at stake. False and manipulated information could lead to human death.

Another example of how quick detection of fake news can improve our well-being is the financial aspect - the **stock market**. Sometimes it happens that some false information (such as Tweets) causes rapid changes in the stock market. By quickly catching such false information, you can gain an advantage over others and adjust your investment strategy.

Detecting Fake News in Literature

Automatic detection of fake news is being frequently addressed in the literature:

- *Fake News Detection on Social Media: A Data Mining Perspective*
  A comprehensive review of detecting fake news on social media (fake news characterizations, basic concepts, algorithms, evaluation metrics and representative datasets)
- *Automatic Detection of Fake News*
  Sets of linguistic features extracted (Ngrams, Punctuations, Psycholinguistic features, Readability, Syntax), SVM classifier
- *Supervised Learning for Fake News Detection*
  Performance comparison of different classifiers (KNN, NB, RF, SVM, XGB)
- *Explainable Fake News Detection*
  Elements of XAI
- *New explainability method for BERT-based model in fake news detection*
  Fake news deteciton with LSTM + XAI
- *... and many more*

Reinforcement Learning for Text Classification

Detecting Fake News with Reinforcement Learning is much less popular than with other algorithms:

- *Domain Adaptive Fake News Detection via Reinforcement Learning*
  The only paper I've found that uses Reinforcement Learning and BERT for fake news detection. Cross Domain approach.
  This is the main paper that I'm going to base my thesis on.
- *Weak Supervision for Fake News Detection via Reinforcement Learning*
  RL used in labeling good quality data. CNN as a fake news detector.
  Performance change as the reward.

Explainable Reinforcement Learning

Both Explainable Reinforcement Learning and the use of RL itself in fake news detection are niche topics. This leaves the combination of these two topics **uncovered in the literature**.
Below are papers that relate to this topic to some extent:

- *Explainable Reinforcement Learning: A Survey*
- *A Survey on Explainable Reinforcement Learning: Concepts, Algorithms, and Challenges*

Table of Contents

1 My plan for the Thesis

2 Topic
- Motivation
- Literature Review

3 **Methodology**
- Data
- Reinforcement Learning
- Extensions

Data Collection

In order to enrich sources, topics, words, I plan to combine different datasets into one large dataset. The variety of topics can negatively affect the model's performance, but literature shows that RL handles this problem well.

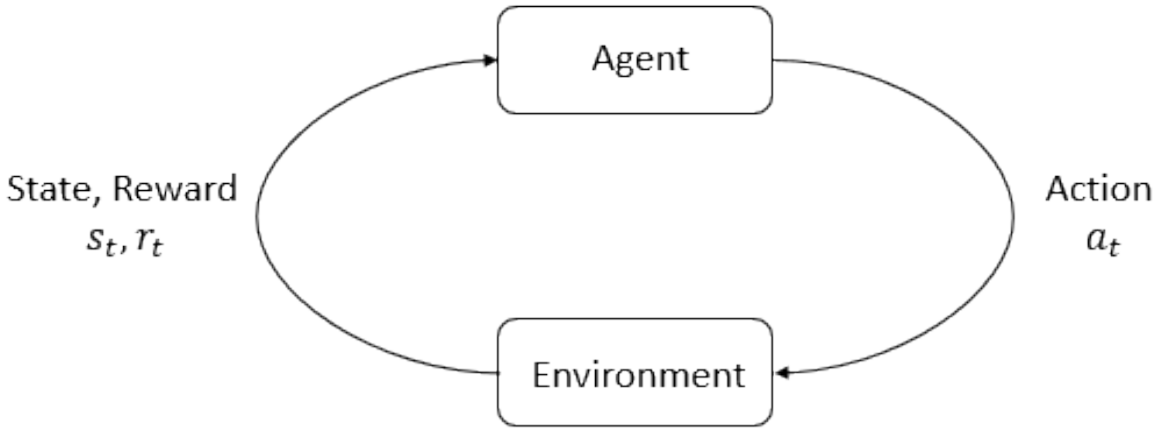
The subject of fake news is quite a sensitive topic and choosing reliable datasets is very important (well labeled news - into fake and real).

Examples:

- *FakeNewsNet*
- *Fake and real news dataset*
- *Fake News*

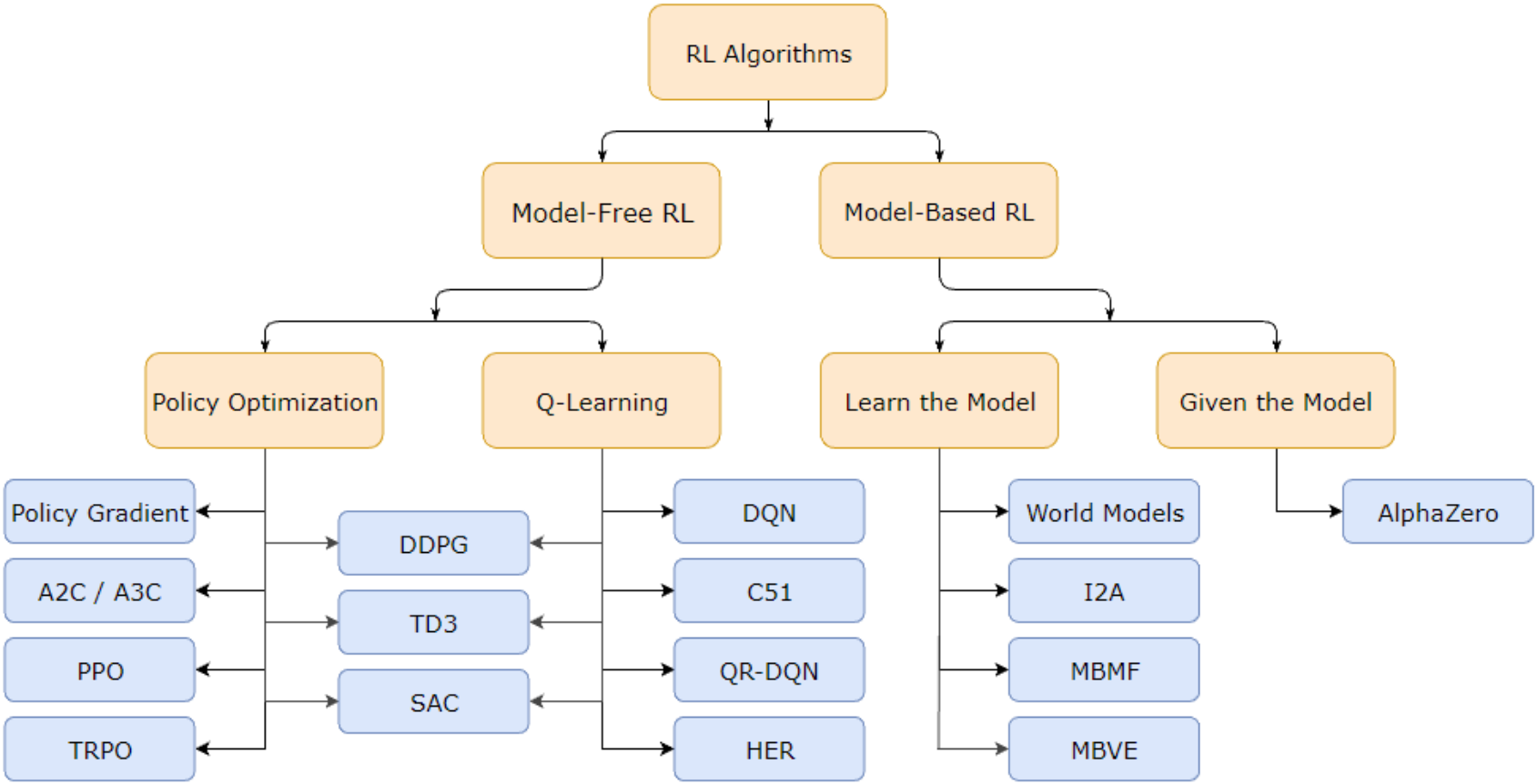
Key Concept

The main characters of RL are the **agent** and the **environment**. The environment is the world that the agent lives in and interacts with. At every step of interaction, the agent sees a (possibly partial) observation of the state of the world, and then decides on an action to take.



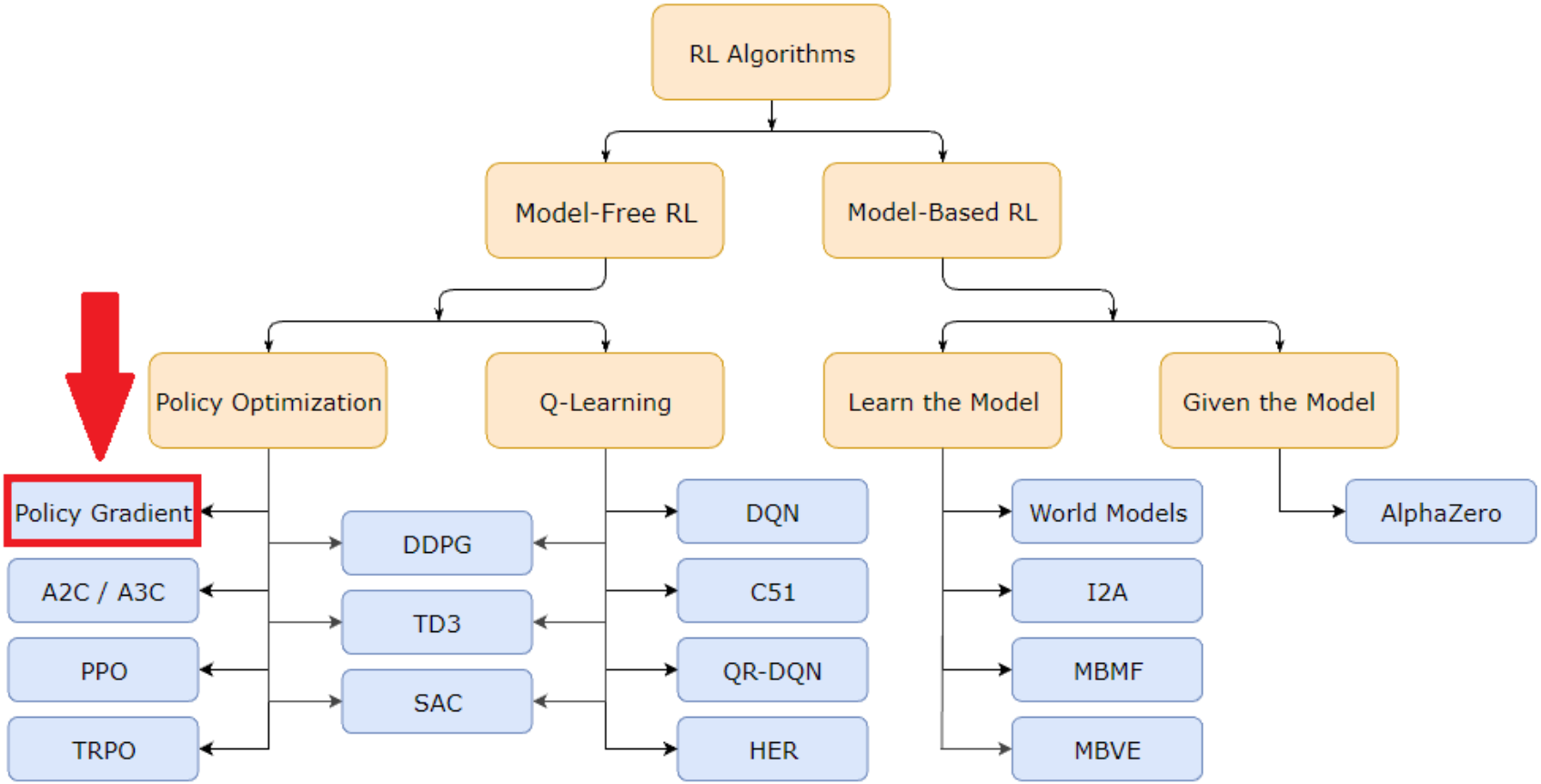
Source: OpenAI

A Taxonomy of RL Algorithms



Source: OpenAI

A Taxonomy of RL Algorithms



Source: OpenAI

Policy Gradient method

Using gradient ascent, we can move θ toward the direction suggested by the gradient $\nabla_\theta J(\theta)$ to find the best θ for π_θ that produces the highest return.

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\pi_\theta)|_{\theta_k} \tag{1}$$

Policy Gradient equation:

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi[Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a|s)] \tag{2}$$

This is an expectation, which means that we can estimate it with a sample mean. we collect a set of trajectories by letting the agent act in the environment using the policy π_θ. Then, after collecting the trajectory dataset, we can compute the policy gradient by averaging the results and take an update step (1).

BERT

Bidirectional Encoder Representations from Transformers (**BERT**) is a pre-trained deep learning model that is used for numerous different language understanding tasks.

It can be **cheaply** fine-tuned to achieve state-of-the-art performance on a staggering number of NLP tasks.
Many papers have confirmed that the use of BERT significantly improves the performance of models (including text classification).

XAI

I also intend to expand my thesis by applying elements of *eXplainable Artificial Intelligence* (**XAI**) to it.

My motivation is that using XAI will help convince more people that the model's results are reliable. For such a socially important topic as fake news, **transparency** of the tools used is essential.

Using XAI would also help with **future research** on this subject.

Thank you for your attention!