**Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis (Discussion Paper)**

[**http://ceur-ws.org/Vol-2646/18-paper.pdf**](http://ceur-ws.org/Vol-2646/18-paper.pdf)

**0. INTRODUCTION**

|  |
| --- |
| **[Summary]**  Sentiment Analysis (SA) 알고리즘을 XAI 기술에 적용하는 것에 대한 방법론으로, **NLP (Natural Language Processing)와 XAI를 결합한 사례**라고 할 수 있다.  **[Introduction and Related Works]**  여기서는 **state-of-the-art explainability method**를 **Sentiment Analysis**에 적용한 다음, **black-box prediction과 유사한 설명을 추출할 수 있는 attention 기반의 방법**을 탐색한다.   * Sentiment Analysis는 **binary sentiment classification** 방법을 자주 사용한다. 이때 negative sentence의 경우는 0, positive sentence는 1로 라벨링한다. 이때 XAI method는 각 word의 기여도를 평가하는, 다음과 같은 heatmap을 출력한다.      * Explanation을 생성하기 위한 방법으로 **LIME (Local Interpretable Agnostic-Model Explanations)**을 많이 사용한다. * 이 외에도 Integrated Gradient와 같은 기술을 사용한다. |

**1. METHODOLOGY**

|  |
| --- |
| 이 논문에서 다루는 모델은 **기존에 학습된 BERT 모델을 기반**으로 하고 있다. BERT 모델이 attention score를 이용하기 때문에, 설명을 하기 위해 이 모델의 점수를 이용한다. BERT 모델은 다음의 2개의 부분으로 구성된다.   * **Embedding creation part:** **텍스트를 벡터로 표현**한다. * **Classifier:** **Embedding creation part의 끝부분**에 위치하며, **전통적인 분류 방법**을 이용한다. * BERT가 attention 기반의 모델이므로, 모델의 결정에 대한 더 나은 insight를 위해 **이 2개의 부분 사이에 attention layer를 추가**한다.   attention layer를 통해 모델은 **prediction task에서 각 word의 중요도를 각 요소를 가중하여 평가**한다. |
| **NEXT PAGE** |

|  |
| --- |
| **PREVIOUS PAGE** |
| 여기서 각 값은 다음을 의미한다.   * : 각 word t에 대한 **embedding vector** * : **importance score**로, 각 word에 대하여 와 weight vector 를 곱해서 얻어진다. 학습 과정에서 **가 학습**된다. * : **attention importance score를 이용한 weighted sum**을 weight으로 하여 계산한 **output representation vector**로 다음 수식으로 나타낸다. * **(Explanation Score)** : 이때 는 learned attention score로 **항상 positive이므로 classifier로부터의 signal을 반영하지 못한다.** 따라서 signal을 반영하기 위해 다음과 같이 를 이용한다.   **학습 과정에서는 BERT weight을 original pre-trained model로 프리징**하고 attention 및 classification layer에서만 최적화한다. |

**2. Experiments**

|  |
| --- |
| **[Dataset]**  영화 리뷰에 대해 성능을 측정하기 위해 **Stanford Sentiment Treebank (SST)**의 데이터를 이용하였고, 영화 리뷰는 Rotten Tomatoes에서 오른쪽 그림과 같이 Fresh Tomato 또는 Rotten Tomato로 분류한 데이터를 이용했다.  **[Training]**  1. 처음에는 **BERT의 수정된 모델을 SST dataset에 대한 sentiment classifier로 학습**시켰다. 이때 하이퍼파라미터는 original paper [3] 에 있는 것을 이용하였다.  2. 그 다음으로 학습된 BERT를 black-box classifier로 하고, test set에 있는 모든 리뷰를 라벨링하여 **explanation score를 LIME, IntGrad, Attention을 이용하여 측정**하였다.  3. 마지막으로 **explanation score를 ground truth label 및 original black-box model의 test set에 대한 prediction과 비교**하였다.   * 이를 위해 explanation score를 통해 예측을 생성해야 하는데, 이를 위해 다음 수식을 이용한다. |
| **NEXT PAGE** |

|  |
| --- |
| **PREVIOUS PAGE** |
| **[VALIDATION]**  Explanation score를 비교하기 위하여 **Fidelity를 이용하여 XAI 기술이 얼마나 black-box 모델을 잘 모방하여 설명하는지 평가**하고, 또 **explanation score와 ground truth label을 비교**하여 그 유사성을 측정하였다.   * 이것들은 **ROC AUC score**를 이용하여 평가하며, 그 결과는 다음 그림과 같다.      * Attention layer에서는 **2분**, IntGrad에서는 black-box model을 모든 step에서 호출해야 하므로 **102분**, LIME은 **1440분**이 걸렸다.   XAI method의 유사성과 차이점을 보다 잘 알아보기 위하여 다음과 같이 **score assignment를 측정**하였다. 각 문장은 test set에 있으며, negative sentiment를 ground truth로 갖는다.    **XAI 방법은 다른 방법과 상충하는 경우**가 가끔 있는데, Figure 6은 correlation coefficient들의 분포에 대한 box-plot을 사용된 모든 test dataset에 대하여 보여 준다. 전체 score가 강하게 서로 연관되어 있지만, **서로 음의 상관관계를 보이는 일부 표본**이 있다. |
| **NEXT PAGE** |

|  |
| --- |
| **PREVIOUS PAGE** |
| Figure 7에서는 **2개의 서로 다른 uncorrelated behavior를 보이는 표본**을 볼 수 있다.   * 위쪽 차트에서는 **negative sentence “This is pretty dicey material”**을 확인할 수 있으며, LIME이나 IntGrad 중 어느 쪽도 dicey라는 단어의 실제 사용을 알아내지 못했다. * 아래쪽 차트에서는 **positive sentence “Not a bad journey at all”**을 확인할 수 있으며, 여기서 IntGrad는 bad라는 형용사의 사용을 알아내는 데 실패했다. * 상기의 예시에서 **explanation socre는 모델의 예측과 충돌**한다. * **Model explanation concordance는 일반적으로 보장되지 않으며**, XAI 기술을 개발하고 설명할 때 고려되어야 한다. |